

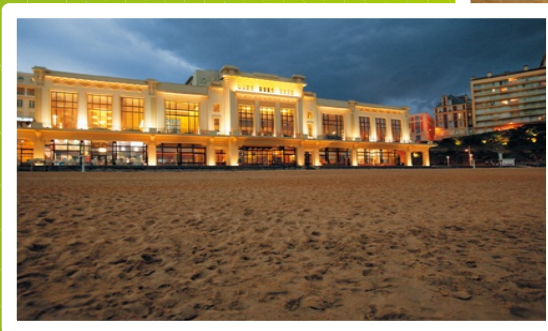
INFORSID

33^e édition - Biarritz

2015



Du 26 au 29 **Mai** 2015 ...



...au **Casino** de Biarritz

- ACTES -



Editeur : Inforsid, 44, Chemin de la Caille, 31750 Escalquens

Illustration : Ville de Biarritz

Impression : Service de reprographie de l'UPPA

ISBN : 2-906855-30-8

L'association INFORSID

Siège Social

44, Chemin de la Caille
31750 Escalquens
Web : <http://inforsid.irit.fr/>

INFORSID est une association régie par la loi de 1901 qui rassemble les chercheurs en informatique des organisations et systèmes d'information et qui a pour objectif de promouvoir les recherches effectuées dans ces domaines en faisant intervenir le plus largement possible les utilisateurs et les industriels. INFORSID centre son activité sur un ensemble de colloques et de séminaires périodiques au cours desquels le point est fait sur l'état des recherches en matière de système d'information et une orientation est donnée pour leur prolongement.

Composition du bureau

Présidente : Régine LALEAU, LACL, Université Paris-Est Créteil, IUT Sénart-Fontainebleau
Vice-présidente : Dominique RIEU, LIG, Université de Grenoble
Trésorier : Philippe ROOSE, LIUPPA, Université de Pau et des Pays de l'Adour, IUT de Bayonne
Secrétaire : Franck RAVAT, IRIT, Université Toulouse
Chargé de communication : Elöd EGYED-ZSIGMOND, LIRIS, Université de Lyon, INSA de Lyon

Présidents d'honneur

Jean-Bernard CRAMPES (Toulouse)
Gilles ZURFLUH (Toulouse)
André FLORY (Lyon)
Claude CHRISMENT (Toulouse)
Michel SCHNEIDER (Clermont-Ferrand)
Corine CAUVET (Aix-Marseille)
Chantal SOULE-DUPUY (Toulouse)

Table des matières

Préface	5
Comités	7
Conférenciers invités	9
Stéphane Frénot : <i>Plateformes d'intermédiations : Le Web, le temps et ... javascript</i> . . .	10
Cyril Labbé : <i>L'histoire de Ike Antkare et de ses amis : fouille de textes et SI scientifiques</i>	11
Oscar Diaz : <i>From Personalization to Individualization : the opportunities brought by browser-based transcoding</i>	12
Session Réseaux sociaux	13
Marie-Françoise Canut, Sirinya On-At, André Péninou and Florence Sedes <i>Enrichissement du profil utilisateur à partir de son réseau social dans un contexte dynamique : application d'une méthode de pondération temporelle</i>	15
Manel Mezghani, André Peninou, Corinne Zayani, Ikram Amous and Florence Sedes <i>Détection des intérêts d'un utilisateur par l'exploitation du comportement d'annotation de son réseau égocentrique</i>	31
Session Modèles et méthodes (1)	47
Rebecca Deneckere and Elena Kornyshova <i>Concept de Famille de Méthodes : qu'en pensent les utilisateurs ?</i>	49
Rachida Seghiri, Frédéric Boulanger, Claire Lecocq and Vincent Godefroy <i>Simulation des Systèmes d'Information des Smart Grids</i>	65
Session Web sémantique et ontologies	81
Firas Zekri, Emna Turki and Rafik Bouaziz <i>AlzFuzzyOnto : Une ontologie floue pour l'aide à la décision dans le domaine de la maladie d'Alzheimer</i>	83
Sebastien Laborie, Franck Ravat, Jiefu Song and Olivier Teste <i>Combining Business Intelligence with Semantic Web : Overview and Challenges</i>	99

Victor Charpenay, Elod Egyed-Zsigmond and Harald Kosch
Leveraging Knowledge from the Linked Open Data Cloud in the task of Reverse Geo-tagging 115

Session Modèles et méthodes (2) 131

Abdelkader Achi and Camille Salinesi
Proposition d'un modèle d'innovation par l'usage des SI 133

Salim Chehida, Akram Idani, Yves Ledru and Mustapha Kamel Rahmouni
Extensions du diagramme d'activité pour contrôler l'accès au SI 151

Amira Derradji
Un méta-modèle et un langage orientés patient pour la représentation de protocoles de soins auto-adaptatifs 167

Stephanie Chollet, Philippe Lalanda and Catherine Hamon
Maintien de la cohérence entre les architectures pour une gestion autonome de systèmes pervasifs 183

Session Données spatiales 199

André Fonteles, Sylvain Bouveret and Jérôme Gensel
Recommandation opportuniste de trajectoires pour l'accomplissement de tâches dans les systèmes crowdsourcing 201

Ziouel Tahar, Derbal Amieur Khalissa and Kamel Boukhalifa
Approche Guidée par les Structures Hiérarchiques Spatiales pour la Généralisation à la Volée dans SOLAP 217

Session Ingénierie des exigences 233

Christophe Ponsard, Robert Darimont and Arnaud Michot
Combining Models, Diagrams and Tables for Efficient Requirements Engineering : Lessons Learned from the Industry 235

Rémi Delmas and Thomas Polacsek
Vérification automatique d'exigences pour les politiques d'échange d'information 251

Juyeon Kang and Saint Dizier Patrick
Une expérience d'un déploiement industriel de LELIE : une relecture intelligente des exigences 267

Session Recommandation 283

Nesrine Ben Yahia, Wala Rebhi, Narjes Bellamine and Henda Ben Ghezela
Approche hybride de recommandation des nouvelles collaborations 285

Imen Akermi, Mohand Boughanem and Rim Faiz <i>Une Approche de recommandation proactive dans un environnement mobile</i>	301
Mariam Bambia, Mohand Boughanem and Rim Faiz <i>Un modèle de langue pour l'estimation de la fraîcheur et la pertinence des documents Web</i>	317
Ateliers	331
L'innovation par les SI dans l'écosystème	332
Autour du concept d'Observatoire en environnement	334
Modélisation des cartographies : principes, démarches, et techniques	335
SI, Connaissance, Homme numérique et décision	336
 Journée bilatérale AFIA / INFORSID SI&IA	 339

Préface

C'est avec un grand plaisir que je vous présente les actes de la 33ème édition du Congrès INFORSID qui se tient cette année à Biarritz, magnifique station balnéaire de la région Aquitaine avec son incomparable façade maritime sur l'Océan Atlantique.

Le congrès INFORSID (INFormatique des ORganisations et Systèmes d'Information et de Décision) est, chaque année, depuis 1982, un lieu d'échange privilégié entre chercheurs, jeunes chercheurs et industriels sur les recherches et développements liés à l'ingénierie, à la gouvernance et à l'évolution des systèmes d'information. Dans un contexte où les informations à traiter sont de plus en plus nombreuses et variées, où les services proposés inter-organisationnels permettent aux processus métier de devenir plus complexes et plus innovants, où le web et les réseaux sociaux sont omniprésents, les recherches en systèmes d'information doivent proposer de nouvelles approches et de nouvelles techniques permettant de développer et de faire évoluer des systèmes d'information de qualité, flexibles, adaptés aux nouveaux besoins, aux nouvelles technologies et aux acteurs des processus.

Cette année, le congrès INFORSID a reçu 43 soumissions d'articles issus de divers pays francophones (France, Algérie, Belgique, Maroc, Tunisie). Beaucoup de ces articles sont proposés en collaboration avec des entreprises, ce qui témoigne de l'intérêt du monde socio-économique pour la recherche en Systèmes d'Information. Aujourd'hui plus que jamais, les systèmes d'information tiennent une place stratégique dans les entreprises et les organisations qui n'ont de cesse de devoir faire preuve d'agilité et d'innovation afin de gagner en compétitivité et en performance.

Le processus de sélection des articles s'est déroulé en plusieurs phases. Dans un premier temps, chacun des articles soumis a été évalué par trois membres du Comité de Programme. Puis, les membres du Conseil du Comité de Programme ont, sur chaque article, dirigé des discussions entre les membres du Comité de Programme afin de résoudre d'éventuels conflits d'évaluation. Enfin, une réunion plénière entre les membres du Conseil du Comité de Programme a permis de sélectionner les articles acceptés pour présentation lors du congrès.

Au final, le programme du congrès INFORSID 2015 comporte 19 articles regroupés en 6 sessions thématiques :

- Réseaux sociaux,
- Modèles et méthodes,
- Web sémantique et ontologies,
- Données spatiales,
- Ingénierie des exigences,
- Recommandation.

La présentation de ces articles scientifiques est complétée par trois conférences invitées. *Stéphane Frénot*, (Laboratoire CITI, INSA, Lyon), *Cyril Labbé* (Laboratoire LIG, Université Joseph Fourier, Grenoble) et *Oscar Diaz* (University of the Basque Country, Donostia, Espagne) acceptent ainsi de nous livrer leurs travaux respectifs qui promettent de riches échanges scientifiques. Je tiens par

avance à leur exprimer toute ma gratitude pour la présentation de leurs travaux respectivement intitulés :

- Plateformes d’intermédiations : Le Web, le temps et ... javascript
- L’histoire de Ike Antkare et de ses amis : fouille de textes et SI scientifiques
- Personalization to Individualization : the opportunities brought by browser-based transcoding

En complément de l’appel à communications classique, le congrès INFORSID a souhaité cette année renouer avec le concept d’ateliers offrant la possibilité d’organiser, en parallèle du congrès, des ateliers de réflexion et de présentation de travaux autour d’une thématique spécifique et complémentaire à celles abordées au sein de la conférence. Quatre ateliers seront ainsi organisés permettant la présentation de travaux de recherche et l’organisation de débats entre chercheurs et professionnels dans des thématiques spécifiques ou novatrices liées aux systèmes d’information : *L’innovation par les SI dans l’écosystème, Le concept d’Observatoire en environnement, La modélisation des cartographies et Les Systèmes d’Information, Connaissance, Homme numérique et décision.*

Enfin, l’association INFORSID et l’association AFIA (Association Française pour l’Intelligence Artificielle) s’associent cette année pour organiser en marge du congrès, une journée commune intitulée *Les utilisateurs au centre du SI : des modèles aux raisonnements*. Cette journée originale regroupant des conférenciers invités et des présentations d’articles issus des deux communautés Systèmes d’Information et Intelligence Artificielle, promet également de riches échanges scientifiques autour du développement de nouvelles solutions issues de l’intelligence artificielle pour une meilleure prise en compte de l’utilisateur au centre du système d’information.

Pour conclure cette préface, je souhaite remercier vivement tous les acteurs de cette édition d’INFORSID 2015 :

- les auteurs des articles soumis au congrès, et les conférenciers venus nous présenter les articles sélectionnés,
- les conférenciers invités qui ont accepté de nous faire part de leurs travaux et de leurs réflexions,
- les porteurs des ateliers et de la journée bilatérale INFORSID / AFIA pour leur dynamisme permettant des échanges originaux autour de thématiques novatrices,
- les membres du Comité de Programme, du Conseil du Comité de Programme et les relecteurs additionnels pour l’important travail d’évaluation et de méta-évaluation qu’ils ont mené tout au long du processus de sélection,
- les membres du comité d’organisation qui, sous la houlette des co-présidents Philippe Roose et Philippe Lopistéguy, nous offrent un cadre de travail magnifique au sein du Casino de Biarritz,
- les participants au congrès qui j’en suis sûre, sauront être sources de discussions scientifiques et de conseils fructueux, en particulier à destination de nos jeunes chercheurs,
- et enfin les membres du bureau et du comité exécutif de l’Association INFORSID qui m’ont confié l’organisation scientifique du congrès, et qui n’ont de cesse de promouvoir cet événement national autour des systèmes d’information.

A tous, je souhaite un excellent congrès INFORSID 2015 !

Agnès FRONT
Présidente du Comité de Programme

Comités

Le comité de la 33e édition d'INFORSID est composé par les responsables de l'organisation ainsi que les membres du comité de programme et les membres du conseil du comité de programme. Les président(e)s sont mentionné(e)s par une étoile (*).

Comité d'organisation

Philippe Roose, LIUPPA, Université de Pau et des Pays de l'Adour *

Philippe Lopisteguy, LIUPPA, Université de Pau et des Pays de l'Adour *

Marie-Noëlle Bessagnet, LIUPPA, Université de Pau et des Pays de l'Adour

Patrick Etcheverry, LIUPPA, Université de Pau et des Pays de l'Adour

Olivier Le Goaër, LIUPPA, Université de Pau et des Pays de l'Adour

Alfredo Goñi, Faculté d'Informatique – UPV/EHU, San Sebastian (Espagne)

Sébastien Laborie, LIUPPA, Université de Pau et des Pays de l'Adour

Christophe Marquesuzaà, LIUPPA, Université de Pau et des Pays de l'Adour

Christian Sallaberry, LIUPPA, Université de Pau et des Pays de l'Adour

Nawel Takouachet, ESTIA-RECHERCHE, Bidart

Conseil du comité de programme

BENALI Khalid, LORIA – Université de Lorraine

CAUVET Corine, Université Aix-Marseille, LSIS

DENECKERE Rebecca, CRI, Université Paris 1 Panthéon – Sorbonne

FAVRE Cécile, Université de Lyon, ERIC Lyon 2

FRONT Agnès, Université de Grenoble, LIG

HUBERT Gilles, IRIT, Université Paul Sabatier

LOPISTEGUY Philippe, LIUPPA, Université de Pau et des Pays de l'Adour

RIEU Dominique, LIG, Université de Grenoble

SERVIGNE Sylvie, LIRIS, INSA de Lyon

TAMZALIT Dalila, LINA, Université de Nantes

Comité de programme

FRONT Agnès, Université de Grenoble, LIG *

BEIGBEDER Michel, École Nationale Supérieure des Mines de Saint-Étienne

BEN GHEZALA Henda, RIADI-ENSI, Université de la Manouba, Tunisie

BORNE Isabelle, Université de Bretagne Sud

BOUDJIDA Nacer, Université de Lorraine / LORIA

CALABRETTO Sylvie, LIRIS, INSA de Lyon
CHABROL Michelle, LIMOS Clermont-Ferrand
CRESCENZO Pierre, Université Nice Sophia Antipolis, CNRS, Laboratoire I3S
DALMAU Marc, LIUPPA, LIUPPA, IUT de Bayonne Pays basque
DELOT Thierry, LAMIH, Université de Valenciennes
DESCONNETS Jean-Christophe, UMR ESPACE-DEV, IRD
DUBOIS Hubert, CEA Tech Midi-Pyrénées
EGYED-ZSIGMOND Elöd, LIRIS, INSA de Lyon
FAUVET Marie Christine, LIG, Université de Grenoble
FOURNIER Sébastien, Université Aix-Marseille, LSIS
FREDJ Mounia, École nationale supérieure d'informatique et d'analyse des systèmes, Rabat
GERVAIS Marie-Pierre, Université Paris Ouest Nanterre La Défense, LIP 6
GRAIET Mohamed, Institut Supérieur d'Informatique et de Mathématique de Monastir
GZARA Lilia, G-SCOP, Grenoble-INP
KHAYATI Oualid, Institut supérieur des arts multimédia de la Manouba
KORNYSHOVA Elena, Conservatoire National des Arts et Métiers, Paris
LAFORST Frédérique, Laboratoire Hubert Curien, Université Jean Monnet, Saint Etienne
MASSEGLIA Florent, INRIA
MAZO Raul, Centre de Recherche en Informatique, Université Paris 1 Panthéon – Sorbonne
MOTHE Josiane, IRIT, ESPE de Toulouse Midi-Pyrénées
MURISASCO Elisabeth, LSIS (UMR CNRS 7296), Université de Toulon
OUSSALAH Mourad, LINA, Université de Nantes
PASTOR Oscar, Universitat Politècnica de Valencia, Espagne
PONCELET Pascal, Université Montpellier – LIRMM
RALYTE Jolita, Université de Genève
RAVAT Franck, IRIT, Université Paul Sabatier
ROCHE Mathieu, UMR TETIS, Cirad
ROUDIES Ounsa, Ecole Mohammadia d'Ingénieurs. Univ. Mohammed V. Rabat
ROUSSEY Catherine, Irstea, Clermont Ferrand
SALLABERRY Christian, LIUPPA, Université de Pau et des Pays de l'Adour
SANTORUM Marco Escuela, Politecnica Nacional, Quito, Ecuador
SAVONNET Marinette, Laboratoire LE2I, Université de Bourgogne
SHARP Bernadette, School of Computing, Staffordshire University
SIMONIN Jacques, Télécom Bretagne – Lab-STICC
SMAIL-TABBONE Malika, Université de Lorraine – Laboratoire LORIA
TESTE Olivier, IRIT, Université Toulouse Jean Jaurès
VILLANOVA Marlène, LIG, Université de Grenoble
WATTIAU Isabelle, Conservatoire National des Arts et Métiers, Paris

Relecteurs additionnels

Mikael Desertot, Noureddine Gasmallah, Kahina Kessi, Laila Kjiri, Léa Laporte, Annig Le Parc-Lacayrelle, Vincent Martin, Francois Pinet, Boukhedouma Saida, Yemna Sayeb, Mikal Ziane

Conférenciers invités

Plateformes d'intermédiations : Le Web, le temps et ... javascript

Auteur : Stéphane Frénot, Laboratoire CITI – INSA-Lyon

Résumé :

Les données numériques sont le grand sujet du XXIe siècle, la ressource d'une industrie émergente, comparable au pétrole pour l'industrie du siècle précédent. Les données sont porteuses de promesses les plus extraordinaires, comme des plus sombres menaces. En mettant en relation des millions d'utilisateurs, les plateformes d'intermédiation sont au coeur de cette révolution numérique. Cette mise en relation, qui était techniquement irréalisable autrefois, confère aux intermédiaires un pouvoir extraordinaire, qui bouleverse l'équilibre des rapports de force dans nos sociétés.

L'équipe INRIA Dice étudie l'impact des grands opérateurs numériques comme Google, Facebook, Netflix ou Uber sur nos sociétés. Je présenterai les grands principes des plateformes d'intermédiations, puis je reviendrai sur deux éléments spécifiques ayant stimulé cette mutation. J'insisterai sur l'influence des fondamentaux du Web dans les communications, le stockage de données et le développement logiciel, pour donner quelques pistes d'explications à notre sur-dépendance aux plateformes Nord Américaines.

L'histoire de Ike Antkare et de ses amis : fouille de textes et SI scientifiques

Auteur : Cyril Labbé, Laboratoire LIG – Université Joseph Fourier – Grenoble

Résumé :

Comment mesurer l'importance et l'impact des communications scientifiques? Que peut-on déduire de ces mesures? Sont-elles falsifiables et falsifiées? Des générateurs aléatoires d'articles scientifiques existent, notamment en informatique mais aussi en physique, en mathématiques ou encore en philosophie. Ces articles, dépourvus de sens, peuvent être utilisés de différentes manières. Ainsi, ils ont permis à "Ike Antkare" de devenir l'un des scientifiques les plus cités au monde, du moins si on en croit Google Scholar [1]. De tels articles apparaissent aussi dans des conférences réelles et des vraies-fausses publications sont parues chez de grandes maisons d'édition scientifique (IEEE et Springer). En conséquence, ces publications, sans aucun sens, ont été comptabilisées par les services de bibliométrie les plus réputés (Scopus, ISI-Web of Knowledge) [2, 3]. L'existence de ces documents peut être l'occasion de s'interroger d'une part sur les habitudes et les processus de validation des documents scientifiques mais aussi sur les modes d'évaluation de la recherche. On verra aussi que ces générateurs aléatoires d'articles scientifiques sont faciles à identifier de manière automatique et que des processus de vérification sont mis en oeuvre par les principaux acteurs du paysage de la diffusion scientifique (ArXiv, Springer, Hindawi, ...).

Bibliographie :

- [1] Cyril Labbé, "Ike Antkare, One of the Great Stars in the Scientific Firmament", 22th newsletter of the international society for scientometrics and informatics, June 2010.
- [2] Cyril Labbé, Dominique Labbé. "Duplicate and fake publications in the scientific literature : how many SCigen papers in computer science?" *Scientometrics* 94, no. 1 (2013) : 379-396
- [3] Richard Van Noorden. Publishers withdraw more than 120 gibberish papers. *Nature*, February 2014.

From Personalization to Individualization : the opportunities brought by browser-based transcoding

Auteur : Oscar Diaz, University of the Basque Country, Donostia, Spain

Résumé :

Surveys show that customers increasingly demand products or services that exactly meet their individual needs. Information Systems (ISs) are not exception. Traditional personalization techniques permit to adjust ISs to the user profile with none (a.k.a. adaptive techniques) or minimum (a.k.a. adaptable techniques) user intervention. However, user profiling might still imply a coarser-grained characterization that the one required for individualization. The keynote addresses the opportunities brought by browser extensions for individualization. Browser extensions permit “to augment” (transcode) rendering by third parties once websites are in operation. The notions of “augmentive” and “augmentable” are introduced to characterize those websites that provide or support the development of browser extensions for individualization purposes. Amazon is used as a case study.

Session Réseaux sociaux

Enrichissement du profil utilisateur à partir de son réseau social dans un contexte dynamique : application d'une méthode de pondération temporelle

Marie-Françoise Canut, Sirinya On-at, André Péninou, Florence Sèdes

*Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, 31062 TOULOUSE Cedex 9
{Marie-Francoise.Canut, Sirinya.On-at, Andre.Peninou, Florence.Sedes}@irit.fr*

RESUME. Le profil de l'utilisateur est un élément central dans les systèmes d'adaptation de l'information. Les réseaux sociaux numériques représentent une source d'informations très riche sur l'utilisateur. Nous nous intéressons au processus d'enrichissement du profil utilisateur à partir de son réseau social. Ce processus extrait les intérêts de l'utilisateur à partir des individus dans son réseau égocentrique afin de construire la dimension sociale du profil de l'utilisateur. Afin de prendre en compte le caractère dynamique des réseaux sociaux, nous proposons, dans ce travail, de construire cette dimension sociale en intégrant un critère temporel afin de pondérer les intérêts de l'utilisateur. Ce poids "temporel", qui reflète la pertinence d'un intérêt, est calculé, d'une part, à partir de la pertinence des individus du réseau égocentrique de l'utilisateur en prenant en compte la fraîcheur de leurs liens avec l'utilisateur et, d'autre part, à partir de la pertinence des informations qu'ils partagent en prenant en compte la fraîcheur de ces informations. Les expérimentations sur les réseaux de publications scientifiques DBLP et Mendeley ont permis de montrer que notre proposition fournit des résultats plus satisfaisants que ceux du processus existant.

ABSTRACT. The development of users' profiles is central for adaptive system to propose information that corresponds to user specific needs. Online social networks provide valuable information sources to collect and identify user's information and user's interests. This work focuses on extracting user's interests from his social egocentric network to build the social dimension of his user profile. To improve this social dimension, we take into account the dynamic of the social network by applying a time-weight strategy in order to drift the user interests. The time-weight of an interest is computed by combining the accuracy of individuals in his egocentric network (computed by taking into account the freshness of their ties) with the information accuracy (computed by taking into account its freshness). The experiments on scientific publications networks (DBLP/Mendeley) allowed us to demonstrate the effectiveness of our proposition comparing to the time-agnostic user profiling approach.

MOTS-CLES : Profil utilisateur, Réseau social, Réseau égocentrique, Pondération temporelle.

Keywords: User Profile, Social Network, Egocentric Network, Time-weight Method.

1. Introduction

Dans les systèmes d'adaptation de l'information, pour pouvoir offrir des informations qui correspondent aux besoins de l'utilisateur, les mécanismes de personnalisation/adaptation doivent disposer d'informations sur les utilisateurs telles que leurs caractéristiques personnelles, leurs préférences générales, leurs centres d'intérêt. De ce fait, le profil utilisateur construit à partir de ces informations devient central dans ce type de mécanisme. Dans nos travaux, c'est le processus de construction du profil utilisateur qui nous intéresse particulièrement. Dans la littérature, plusieurs modèles de profil utilisateur ont été proposés. Nous représentons le profil utilisateur sous forme de vecteur de termes pondérés qui décrivent les centres d'intérêt de l'utilisateur définis par un ensemble de mots clés (foot, tennis, danse,...).

Le contexte de nos travaux porte sur le processus de construction du profil utilisateur à partir de son réseau social représenté par un graphe de relations entre les individus. Nous travaillons en particulier dans le contexte des médias sociaux qui s'appuient généralement sur des réseaux sociaux numériques. Ce processus de construction permet d'une part de compléter le profil d'un nouvel utilisateur ou celui d'un utilisateur peu actif pour lequel le profil serait insuffisant pour les mécanismes d'adaptation, et d'autre part d'enrichir un profil existant. Nous utilisons dans ce papier, le terme profil social pour désigner un profil construit par ce type de processus.

Nous partons du constat que les intérêts de l'utilisateur évoluent au fil du temps, en particulier, dans le cas des intérêts extraits depuis les médias sociaux. En effet, dans les médias sociaux, l'information partagée évolue sans cesse du fait des interactions sociales en ligne (partage, échange d'informations) qui génèrent (plus facilement) un volume important d'informations volatiles. Pour un utilisateur, les intérêts qui sont extraits à une période donnée peuvent ne plus être significatifs ultérieurement. Par exemple, pour un utilisateur qui regarde le foot uniquement pendant la période de la coupe du monde (une fois tous les 4 ans) et qui partage des informations dans son réseau social, ses intérêts pour la coupe du monde ou le foot ne sont significatifs pour son profil que pendant cette période.

Un autre problème qui peut être rencontré est la pertinence des liens de l'utilisateur avec les membres de son réseau social ainsi que celle des informations qu'il partage. Les utilisateurs peuvent créer des contacts en ligne sans forcément connaître les personnes dans la vie réelle. Si on reprend l'exemple précédent, un utilisateur peut suivre les informations à partir des comptes de joueurs de foot qui sont alors enregistrés comme contacts sans pour autant les connaître personnellement. Après la compétition, les liens avec ces joueurs deviennent moins importants pour l'utilisateur. Cela nous montre bien que l'on ne peut pas prendre en compte (ou donner la même importance à) toutes les informations existant dans les réseaux sociaux pour refléter les intérêts d'un utilisateur à un moment donné. Nous nous trouvons donc dans la problématique suivante : dans le contexte des médias sociaux, comment construire un profil social de l'utilisateur qui soit à la fois pertinent et à jour ?

Pour répondre à ce problème, nous considérons que l'évolution des intérêts des utilisateurs pousse ces derniers à modifier leurs relations sociales et les informations

qu'ils partagent. L'évolution du réseau social reflète alors l'évolution des intérêts des utilisateurs. Nous proposons donc de prendre en compte la caractéristique dynamique (plus précisément des informations temporelles) du réseau social pour sélectionner efficacement les informations provenant de ce réseau afin de calculer un profil social pertinent et à jour. Nous envisageons donc dans ce travail, de construire la dimension sociale de l'utilisateur par l'intégration d'une mesure temporelle dans le calcul du poids (représentant la pertinence) des intérêts liés au profil social de l'utilisateur. Ce poids, appelé poids temporel, est calculé, d'une part, à partir du poids de pertinence temporelle des individus et du poids de pertinence temporelle des informations que ces individus partagent. Le poids de pertinence des individus est calculé en appliquant une méthode de prédiction de liens temporelle afin de sélectionner les individus ayant les liens les plus actifs avec l'utilisateur. Le poids de pertinence temporelle des informations partagées est calculé en prenant en compte la fraîcheur de ces informations afin de prendre en compte les informations les plus récentes et en extraire des intérêts pertinents et à jour.

Nous présentons tout d'abord un état de l'art sur la construction du profil utilisateur à partir de réseaux sociaux, puis nous insistons sur la caractéristique dynamique des intérêts de l'utilisateur et des réseaux sociaux. Nous présentons ensuite notre proposition permettant de prendre en compte cette double dynamique dans le processus de construction du profil social de l'utilisateur. Ensuite, nous décrivons nos expérimentations effectuées dans deux réseaux de publications scientifiques (DBLP et Mendeley). Nous terminons par une conclusion et les perspectives de ce travail.

2. Etat de l'art : dérivation du profil utilisateur à partir de réseaux sociaux

L'approche de construction du profil social de l'utilisateur où les intérêts de l'utilisateur sont enrichis à partir des informations partagées par les individus de son réseau social permet d'améliorer la représentation du profil utilisateur mais aussi de résoudre le problème de démarrage à froid ou le cas d'un utilisateur très peu actif dans le système (qui interagit moins avec le système et fournit donc moins d'informations pour déduire ses intérêts) (Massa et Avesani, 2007).

(Cabanac, 2011) s'appuie principalement sur des utilisateurs considérés individuellement pour proposer un système de recommandation sociale d'articles scientifiques aux chercheurs. (Carmel et al., 2009) utilise le même principe pour proposer un système de recherche d'information sociale. Ces méthodes peuvent être définies comme des méthodes de type « autoritaire » pour lesquelles les utilisateurs les plus actifs ou les plus influents dans le réseau social seront privilégiés dans la dérivation de leur profil social. Pourtant, dans le contexte des réseaux sociaux numériques, il est difficile de considérer que tous les centres d'intérêt d'un utilisateur influent dans le réseau social, peuvent représenter fidèlement cet utilisateur. Ce problème peut être mieux perçu si l'on considère des environnements de réseaux sociaux numériques comme Facebook dans lesquels un utilisateur peut être ami avec plus de 1000 individus. Parmi ces individus, très peu sont réellement ses amis proches qui permettraient de fournir des informations permettant de caractériser correctement cet utilisateur, même s'ils sont très actifs ou influents.

Les travaux de (Tchunte, 2013) portent sur l'étude de la dérivation du profil utilisateur à partir des communautés de son réseau égocentrique, un réseau largement utilisé en sociologie. Il s'agit d'un graphe composé des relations entre les individus situés à distance 1 (directement reliés) de l'utilisateur (appelé égo), l'égo étant bien entendu exclu de ce graphe. Ces travaux proposent donc une méthode plutôt de type « affinitaire », dans laquelle c'est la présence d'affinités, de liens, de relations entre les individus d'une communauté du réseau égocentrique de l'utilisateur qui permet de dériver des informations à associer à son profil. Par rapport aux méthodes autoritaires, cette approche éliminera, d'une part, les individus ayant des relations superflues avec l'utilisateur (amitiés acceptées au hasard ou pour démontrer un certain pouvoir par le nombre d'amis), et d'autre part, les éléments de profil non significatifs pour l'utilisateur. Cette étude propose tout d'abord un modèle générique du profil utilisateur suivant deux dimensions : une **dimension utilisateur** dont les centres d'intérêt sont calculés à partir des activités propres de l'utilisateur et **une dimension sociale** (profil social) dont les centres d'intérêt sont calculés à partir des activités de son réseau égocentrique. Ces deux dimensions étant complémentaires et indépendantes, elles peuvent être utilisées par les mécanismes d'adaptation, recommandation soit individuellement soit couplées. L'étude présente ensuite un processus de dérivation de la dimension sociale du profil utilisateur à partir des communautés extraites de son réseau égocentrique (détaillé dans la section 5).

Toutefois, la gestion dynamique du profil utilisateur est un problème soulevé et non traité dans le travail existant de (Tchunte, 2013). En effet, les centres d'intérêt d'une personne sont amenés à évoluer dans le temps (changement d'environnement ou de contexte du travail). Nos travaux s'appuient sur la proposition de (Tchunte, 2013), en tentant de prendre en compte la gestion dynamique du profil utilisateur dans le processus de dérivation de la dimension sociale de ce profil. Il nous paraît alors important d'étudier l'évolution de ce profil en fonction de l'évolution du réseau social de cet utilisateur. Nous nous sommes alors intéressés à ces deux types d'évolution.

3. Gestion de l'évolution des intérêts dans le profil utilisateur

L'étude de l'évolution des intérêts de l'utilisateur consiste à prendre en compte le changement de ses intérêts à travers le temps (Crabtree et al., 1998). A partir des travaux dans la littérature, nous pouvons distinguer deux approches de gestion de l'évolution des intérêts dans le profil utilisateur. La première approche consiste à gérer la dynamique des intérêts de l'utilisateur après la phase d'extraction des intérêts et correspond à un processus de mise à jour du profil utilisateur. La deuxième approche consiste à prendre en compte la dynamique des centres d'intérêt pendant l'étape d'extraction des intérêts.

Nous nous intéressons dans ce travail à la deuxième approche, dans le but de construire un profil social pertinent dès sa première utilisation. Dans ce type d'approche, les modèles de profil utilisateur utilisés sont, soit des modèles à court terme où le profil utilisateur se construit et s'utilise en exploitant les informations récentes de l'utilisateur, soit des modèles à long terme où les intérêts sont extraits et

enrichis à travers le temps. Dans la recherche d'information personnalisée, on utilise l'historique à court terme de l'utilisateur lié à une seule (la dernière) session de recherche pour extraire ses intérêts (Bennett et al., 2012). Avec le même principe, plusieurs travaux proposent une approche utilisant un critère temporel pour mieux cerner l'évolution et la dynamique des informations étudiées. La plupart de ces travaux se basent sur l'approche « time-forgotten » qui ignore les informations trop anciennes (Cheng et al., 2008; Maloof et Michalski, 2000). Dans ce type d'approche, on oublie complètement les informations dépassant une date limite. Pourtant, certaines de ces informations ignorées peuvent être utiles et ne pas les prendre en compte peut entraîner une perte d'informations intéressantes. En effet, (Tan et al., 2006) ont prouvé que l'historique de recherche à long terme est très important pour améliorer la tâche de recherche d'informations dans le cas de requêtes récurrentes.

Dans l'utilisation de modèles à long terme, toutes les informations de l'utilisateur sont conservées (et peuvent contenir éventuellement des biais), il est alors difficile de sélectionner les informations pertinentes pour représenter l'utilisateur à un instant donné. Il se peut que des intérêts anciens de l'utilisateur ne soient plus significatifs à ce jour. Cette remarque peut être retrouvée dans (Kacem et al., 2014; Li et al., 2013) qui proposent d'appliquer une fonction temporelle pour pondérer les intérêts de l'utilisateur selon leur fraîcheur. Cette idée peut être retrouvée également dans le contexte de la construction du profil utilisateur à partir d'un réseau d'annotations comme dans (Zheng et Li, 2011) qui utilise des fonctions temporelles pour pondérer des tags avant d'en extraire les intérêts de l'utilisateur. Dans ce type d'approche, toutes les informations existantes de l'utilisateur sont exploitées mais de manière plus restreinte. Nous nous intéressons dans ce travail à cette dernière approche.

Il est important de noter que dans nos travaux, les principales interactions que nous souhaitons étudier pour détecter un changement de centres d'intérêt ne sont pas ciblées sur l'utilisateur lui-même mais sur les éléments de son réseau social (liens entre les membres, informations qui circulent entre les membres). L'évolution de la dimension sociale du profil utilisateur est liée à l'évolution de son réseau social. Pour étudier l'évolution des intérêts de l'utilisateur dans notre contexte, il est donc important de prendre en compte des informations temporelles en s'appuyant sur les mécanismes d'évolution du réseau social de l'utilisateur.

4. Evolution du réseau social

Depuis plusieurs années, l'étude sur les propriétés et les caractéristiques des réseaux sociaux (densité, degré de distribution, classification, composants connexes, communautés, ...) a été considérée comme une piste importante de recherche. Cependant, la plupart des études ont été conduites avec une vision statique du réseau alors qu'un réseau social numérique est considéré comme un réseau dynamique (qui évolue au fil du temps). Pour répondre à ce problème, l'analyse de la dynamique du réseau social ne peut s'effectuer qu'en prenant en compte un critère temporel pour comprendre les évolutions qui se produisent dans le réseau (Spiliopoulou, 2011). Nous pouvons distinguer deux types de dynamique dans un réseau social : la dynamique de la structure du réseau et la dynamique des informations partagées.

4.1 Dynamique de la structure du réseau social (dynamique dans le réseau)

La dynamique de la structure du réseau social provient de la dynamique des interactions des individus dans le réseau, liée à la création et/ou à la suppression des liens mais aussi à la persistance des liens déjà créés. L'analyse de l'évolution de la structure du réseau social porte sur la visualisation de l'évolution du réseau, sur le modèle de l'évolution de la structure du réseau pour prédire comment le réseau va évoluer à travers le temps (Kumar et al., 2006; Leskovec et al., 2008) et sur la prédiction de liens (Liben-Nowell et Kleinberg, 2003; Hasan et Zaki, 2011) qui se focalise sur la formation de liens entre les nœuds dans le réseau. Etant donné un réseau social, on détermine, pour chaque paire de nœuds, s'il y a une possibilité de formation de lien entre eux dans le futur.

4.2 Dynamique des informations dans le réseau (dynamique sur le réseau)

Le deuxième facteur important lié à l'analyse temporelle des informations dans le réseau social porte sur le partage et la diffusion d'informations. L'analyse de la dynamique des informations porte sur le modèle de diffusion d'informations, sur la recherche des nœuds influents dans le réseau mais aussi sur les techniques permettant de diffuser efficacement des informations dans le réseau social (Jiang et al., 2014). D'après plusieurs travaux de recherche, la structure du réseau a beaucoup d'impact sur la dynamique des informations dans le réseau. Les individus qui sont considérés proches dans le réseau ont une probabilité plus grande de diffuser des informations entre eux. Dans le cas inverse, la dynamique de la diffusion d'informations peut également être à son tour, un facteur pour le changement de la structure du réseau social (Stattner et al., 2013; Weng et al., 2013).

Plusieurs travaux sur l'évolution du réseau social ne considèrent pas le changement d'informations dans le réseau comme un facteur important pour l'analyse du réseau. Toutefois, dans le contexte des médias sociaux dont les membres créent et partagent une masse d'informations importante, les données circulant dans le réseau deviennent volatiles et évoluent rapidement. Il existe des facteurs comme les « Buzz », les campagnes de marketing et les événements remarquables qui peuvent provoquer des mouvements sociaux. De tels facteurs peuvent augmenter la croissance de la diffusion de l'information dans les médias sociaux mais aussi la création de nouveaux liens entre les personnes qui s'intéressent à l'événement. Pourtant, les informations partagées sont souvent temporaires et disparaissent quand l'événement se termine, par exemple, la guerre civile libyenne, la révolution égyptienne de 2011 (Gomez Rodriguez et al., 2013). Les techniques existantes de traitement des informations doivent être adaptées aux caractéristiques de ces nouvelles sources d'informations afin d'obtenir des informations qui soient les plus représentatives de l'utilisateur. Il paraît donc important de prendre en compte non seulement la dynamique de la structure du réseau mais également la dynamique des informations dans l'analyse de l'évolution du réseau.

Les caractéristiques dynamiques des médias sociaux nous amènent à considérer qu'il n'est pas certain que les individus du réseau social de l'utilisateur ainsi que les

informations qu'ils partagent soient toutes pertinentes et à jour. Notre proposition prend en compte ces remarques dans le processus de dérivation de la dimension sociale du profil utilisateur.

5. Proposition

Notre travail utilise des critères temporels pour prendre en compte la dynamique de la structure du réseau mais aussi la dynamique des informations échangées dans le processus de construction/dérivation de la dimension sociale l'utilisateur. Nous nous appuyons sur les travaux de (Tchunte 2013). Nous expliquons brièvement le processus de dérivation de la dimension sociale (CoBSP) proposé par ces travaux qui se fait en 4 étapes successives :

1) La première étape consiste à extraire, depuis le réseau égocentrique d'un utilisateur, les communautés de ce réseau en utilisant l'algorithme iLCD proposé par (Cazabet et al., 2012).

2) La deuxième étape consiste à calculer le profil de chaque communauté détectée dans la première étape. Le profil d'une communauté peut être calculé en utilisant les informations de tous les individus qui en font partie (ici, les poids associés aux éléments présents dans la dimension utilisateur de chaque individu).

3) La troisième étape consiste à caractériser chaque communauté en se basant sur une caractérisation structurelle et/ou une caractérisation sémantique. La caractérisation structurelle se base sur la centralité de degré (degree centrality). Selon cette mesure, la communauté qui possède le plus grand nombre de connexions directes dans le réseau est caractérisée comme la plus importante par rapport aux autres communautés. La caractérisation sémantique d'une communauté consiste à rechercher sa spécificité par rapport aux autres communautés en se basant sur la mesure de pondération tf-idf. Les deux caractérisations peuvent être combinées pour obtenir une caractérisation unique (sémantico-structurelle). La caractérisation sémantico-structurelle de chaque élément e du profil d'une communauté $C1$ sera calculée par la formule suivante :

$$\text{Caractérisation finale}(e) = \alpha \text{Struct}(C1) + (1-\alpha) \text{Sem}(C1) \quad (1)$$

Le paramètre α (valeur comprise entre $[0,1]$) dans la formule permettra de juger et de faire varier l'importance des mesures de structure dans la dérivation de la dimension sociale du profil de l'utilisateur par rapport aux mesures sémantiques. La valeur de α est déterminée de manière empirique et sera donc fixée lors des expérimentations.

4) La quatrième étape permet de dériver les intérêts de la dimension sociale par combinaison des différents poids associés à un intérêt à partir de toutes les communautés en utilisant une fonction linéaire.

Nous proposons dans ce travail de prendre en compte le critère temporel dans l'étape de pondération des intérêts de l'utilisateur. Notre travail se situe après l'étape 3 de (Tchunte, 2013) où nous intégrons un poids temporel (**structuro-sémantico-temporel**) dans l'étape de calcul de la pondération des intérêts de l'utilisateur.

Notre proposition se divise en 2 parties. La première partie consiste à attribuer aux informations, un poids prenant en compte le critère temporel (poids temporel) (qui correspond à la prise en compte de la dynamique de la structure du réseau et des informations dans le réseau). La deuxième partie explique comment exploiter ce poids temporel, lors de l'étape de construction de la dimension sociale de l'utilisateur.

5.1 Calcul du poids structuro-semantic-temporel

Le poids structuro-semantic-temporel est calculé en combinant d'une part, le poids structurel du réseau égocentrique qui prend en compte la dynamique de la structure du réseau (appelé poids structuro-temporel : StrTem) et d'autre part le poids de pertinence de l'information utilisée pour extraire les intérêts (appelé semantic-temporel : SemTem). Nous détaillons dans la suite, le calcul de ces deux poids ainsi que leur combinaison.

5.1.1 Calcul du poids structuro-temporel (StrTem) avec une méthode de prédiction de liens temporelle

Il s'agit ici de la prise en compte de la dynamique de la structure du réseau. Comme les liens entre l'utilisateur et les individus dans son réseau égocentrique peuvent varier au fil de temps, nous considérons que les individus qui ont les relations les plus récentes avec l'utilisateur ont une probabilité plus grande de partager les mêmes intérêts avec l'utilisateur (plus significatifs). Nous donnons donc un poids plus important aux informations partagées par ces individus par rapport aux individus ayant des liens moins récents. Il s'agit de calculer la pertinence d'un individu par rapport à l'utilisateur central en prenant en compte les informations temporelles de ses liens avec l'utilisateur central (date de création de liens avec l'utilisateur central, durée de la relation, ...).

La technique de prédiction de liens temporelle nous semble être adaptée à ce problème. Le principe de la prédiction de liens est de calculer la similarité entre les deux nœuds en utilisant des critères tels que la topologie du réseau par exemple. Selon le score de similarité obtenu, on obtient la probabilité que ces nœuds se connectent dans le futur. Contrairement à son utilisation initiale, nous utilisons dans notre travail, la prédiction de lien pour calculer le poids de similarité entre deux nœuds déjà connectés afin de représenter la persistance du lien entre ces nœuds dans le futur. La plupart des travaux s'intéressent à la prédiction de liens sans tenir compte de l'évolution du réseau. Cependant, comme introduit précédemment, la prise en compte du temps est essentielle pour un réseau social numérique. Nous nous intéressons donc à la prédiction de liens temporelle qui prend en compte les informations structurelles et temporelles lors du calcul du score de similarité entre deux nœuds.

Nous nous appuyons sur le travail de (Tylenda et al., 2009) qui proposent d'appliquer des critères temporels (durée d'existence de la dernière collaboration entre deux nœuds étudiés.) aux méthodes de base de prédiction de liens comme Adamic/Adar, introduit dans (Liben-Nowell et Kleinberg, 2003). Adamic/Adar se base sur le nombre de voisins communs. Pour une paire de nœuds (x,y), la formule du calcul de score de similarité de Adamic/Adar est :

$$AA(x, y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log|\Gamma(z)|} \quad (2)$$

où $\Gamma(x)$, $\Gamma(y)$ sont les ensembles des voisins de x et y respectivement, z est un noeud voisin commun de x et y et $\Gamma(z)$ l'ensemble de ses voisins. La formule de Adamic/Adar qui prend en compte les informations temporelles (Tylenda et al., 2009) est :

$$AA(x, y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{w(x,z) \cdot w(z,y)}{\log|\Gamma(z)|} \quad (3)$$

où $w(x,y)$ représente le poids temporel entre deux nœuds x et y (selon la date de leur dernière interaction). Pour calculer ce poids $w(x,y)$, nous utilisons le travail de (Ding et Li, 2005) qui pondère une information selon sa date de publication (plus l'information est récente, plus elle est importante). Mais nous appliquons cette fonction temporelle à la dernière interaction entre les deux nœuds.

$$f(t) = e^{-\lambda t} \quad (4)$$

Pour chaque t_i ($i \in \mathbb{N}$), t_0 est considéré comme l'instant le plus récent et ainsi de suite (ex : t_0 pour l'année 2014, t_1 pour l'année 2013, ...). La valeur λ , fixée de manière empirique lors des expérimentations, est le taux de décroissance des valeurs. Plus λ est petit, plus les informations récentes sont importantes. La valeur de $w(x,y)$ est calculée en fonction de $f(t)$, t représentant l'estampille de la dernière interaction entre w et y .

$$w(x, y) = f(t) \quad (5)$$

5.1.2 Poids sémantico temporel (SemTem)

Nous donnons également de l'importance à la fraîcheur des informations. Comme nous considérons que les intérêts de l'utilisateur doivent évoluer selon les informations partagées dans son réseau, les plus récentes auront plus d'importance que les plus anciennes. A partir des informations que partage chaque individu du réseau égocentrique de l'utilisateur, nous proposons de pondérer leur poids de pertinence en intégrant des informations temporelles. Nous pondérons les intérêts extraits de cette information avec la mesure de pondération tf-idf qui est une mesure de fréquence de terme et nous appliquons la même fonction temporelle exponentielle (5) au poids des informations. Nous appliquons cette fonction au poids de chaque information selon sa fraîcheur.

$$\text{SemTem}(i) = \text{score_tf-idf} * \text{score_temporel}(i) \quad (6)$$

5.1.3 Poids final (Poids structuro-sémantico temporel)

Pour obtenir le poids final d'une information, le poids de la caractérisation structurelle et le poids de la caractérisation sémantique sont combinés. Le poids d'un intérêt i extrait depuis un individu **ind** du réseau égocentrique d'un utilisateur sera calculé par la formule suivante :

$$\text{Poids final (i)} = \gamma(\text{StrTem}(\text{ind})) + (1-\gamma)(\text{SemTem}(i)) \quad (7)$$

Le paramètre γ (compris entre $[0,1]$) dans la formule permettra de juger et de faire varier l'importance du poids structuro-temporel dans la dérivation de la dimension sociale du profil de l'utilisateur par rapport au poids sémantico-temporel. Plus la valeur de γ est importante plus on donne de l'importance au poids structuro-temporel. La valeur de γ sera fixée de manière empirique lors des expérimentations.

5.2 Construction de la dimension sociale de l'utilisateur selon le poids de pertinence temporelle des sources et des informations

Pour construire le profil social de l'utilisateur en appliquant les techniques de prédiction de liens temporelle et de pondération d'informations décrites précédemment, nous utilisons le même principe (CoSPK) proposé par (Tchuente, 2013) mais dans l'étape d'extraction des intérêts de l'utilisateur, nous pondérons un terme extrait avec le poids temporel associé calculé par l'équation (7). Cette méthode consiste à ajouter une nouvelle signification aux poids des centres d'intérêt du profil de l'utilisateur avant de les exploiter dans la construction/enrichissement du profil de l'utilisateur (il s'agit d'augmenter ou de diminuer l'importance des centres d'intérêt). Pour ce faire, premièrement, nous calculons le poids de pertinence comme indiqué dans 5.1, deuxièmement, nous classifions les intérêts de manière croissante par rapport à ce poids, enfin nous utilisons la formule de combinaison linéaire pour pondérer chaque intérêt comme indiqué dans l'étape 4 du processus CoSPK.

Pour étudier l'impact de la prise en compte de la dynamique des informations par rapport à la dynamique de la structure du réseau, nous proposons deux algorithmes différents associés à cette démarche de construction du profil social de l'utilisateur. Nous appelons CoSPKTP, l'algorithme qui utilise seulement le poids structuro-temporel pour pondérer les informations traitées et nous appelons CoSPKTPI, l'algorithme qui utilise le poids structuro-sémantico temporel pour pondérer les informations traitées.

6. Expérimentation

Pour valider notre proposition, nous avons comparé la pertinence de notre approche par rapport à l'approche existante. Pour cela, nous comparons la pertinence des dimensions sociales construites par ces approches. Le domaine d'expérimentation choisi concerne les réseaux d'auteurs de publications scientifiques (DBLP et Mendeley). Dans ces réseaux, les noeuds représentent les auteurs. Deux noeuds peuvent être reliés par un lien s'ils publient ensemble. Nous exploitons le réseau des auteurs DBLP pour lesquels on peut calculer les centres d'intérêt à partir des titres de leurs publications pour construire le réseau égocentrique de chaque auteur et dériver la dimension sociale de son profil. Pour calculer la pertinence de la **dimension sociale** créée, nous la comparons avec la **dimension utilisateur** qui contient les intérêts réels de l'utilisateur que nous créons à partir des intérêts indiqués explicitement par l'utilisateur sur son profil Mendeley. L'objectif d'une telle expérimentation est de

montrer que la dimension sociale construite indépendamment des informations propres à l'utilisateur est suffisamment représentative de l'utilisateur et pourra donc être utilisée en particulier lorsque la dimension utilisateur n'existe pas ou est incomplète.

Nous partons des travaux existants pour construire les dimensions sociales et la dimension utilisateur du profil de l'utilisateur. Comme présenté dans la figure 1, nous recherchons dans un premier temps, les auteurs de publications scientifiques dans Mendeley ayant plus de 6 centres d'intérêt. Ensuite, nous partons de ces auteurs pour rechercher leurs profils dans DBLP à condition qu'ils aient au moins 50 co-auteurs dans ce dernier. Notre échantillon de test porte sur 79 auteurs. Nous construisons ensuite pour un auteur, les différentes dimensions sociales à partir de ses données DBLP (issues des différentes approches de construction de la dimension sociale) et la dimension utilisateur à partir des intérêts indiqués dans son profil Mendeley. Enfin nous évaluons les dimensions sociales construites par les mesures de rappel et de précision.

6.1 Construction de la dimension sociale du profil utilisateur

Dans cette étape, nous construisons 3 dimensions sociales avec les différents algorithmes (CoSPK pour l'approche existante, CoSPKTP et CoSPKTPI pour l'approche proposée).

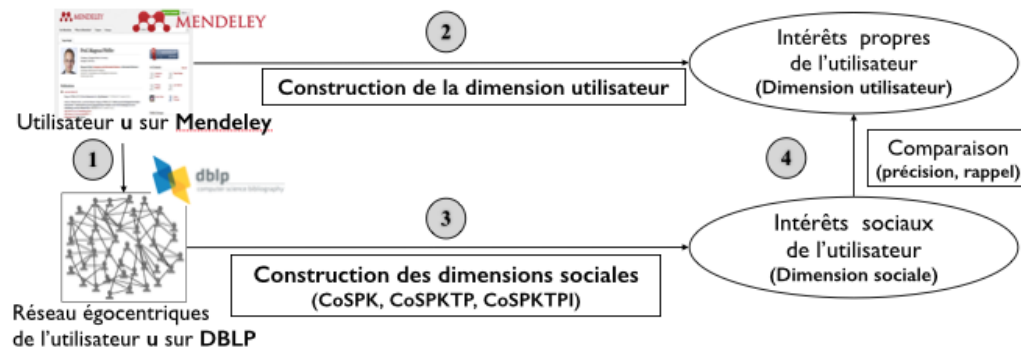


Figure 1. Illustration de la construction de la dimension sociale et de la dimension utilisateur depuis des données DBLP et Mendeley.

6.1.1 Construction de la dimension sociale avec l'algorithme CoSPK

Comme décrit précédemment, le réseau égo-centrique d'un auteur (ego) est construit dans DBLP à partir des relations entre ses co-auteurs. La première étape consiste donc à récupérer les co-auteurs de cet ego. Nous stockons ainsi la liste des publications de tous les auteurs qui font partie de son réseau égo-centrique. Ceci permet de récupérer le titre des publications des communautés. Nous analysons ensuite les titres de publications pour extraire des termes significatifs qui seront considérés comme les centres d'intérêt de l'utilisateur. Les dictionnaires utilisés dans ce

processus sont les dictionnaires de synonymes regroupant tous les synonymes d'un mot. Ces synonymes seront ensuite considérés comme des occurrences de ce mot. Les filtres permettent de sélectionner les mots les plus importants et d'en exclure certains moins importants et moins significatifs tels que des mots vides. Seuls les mots retenus seront considérés comme les centres d'intérêt de l'utilisateur. Ensuite, nous exploitons les mesures sémantiques et les mesures structurelles et représentons les centres d'intérêt de cette dimension sociale par un vecteur de termes pondérés (Tchuenté 2013).

6.1.2 Construction de la dimension sociale en prenant en compte les informations temporelles (CoSPKTP et CoSPKTPI)

Pour construire les deux dimensions sociales, nous utilisons le même principe que celui de la construction de la dimension sociale CoSPK. Par contre, dans l'étape d'extraction des intérêts de l'utilisateur, le poids d'un terme extrait correspond pour CoSPKTP, à son poids sémantico-structurel calculé par la formule (3) et pour CoSPKTPI, à son poids temporel final calculé par la formule (7).

6.2 Construction de la dimension utilisateur du profil utilisateur à partir de Mendeley

Pour construire la dimension utilisateur, nous utilisons les intérêts indiqués explicitement par l'utilisateur dans son profil Mendeley. En appliquant le même traitement que celui de la construction de la dimension sociale, nous utilisons les dictionnaires et les filtres pour extraire des termes significatifs qui seront considérés comme les centres d'intérêt dans la dimension utilisateur. A la différence de la dimension sociale, la pondération de chaque mot est définie uniquement par la mesure de fréquence t_f dans le texte entier analysé. Les centres d'intérêt de cette dimension seront également représentés par un vecteur de termes pondérés.

6.3 Evaluation

Après l'étape de construction de la dimension sociale et de la dimension utilisateur du profil utilisateur, nous obtenons 3 dimensions sociales différentes construites par les algorithmes CoSPK, CoSPKTP, CoSPKTPI. Les dimensions sociales construites seront comparées à la dimension utilisateur via les mesures de précision et de rappel. Dans notre contexte d'évaluation, la précision d'un algorithme de dérivation de la dimension sociale est évaluée par le nombre de centres d'intérêt prédits qui sont trouvés dans la dimension utilisateur par rapport au nombre total des centres d'intérêt calculés dans la dimension sociale.

$$\text{Précision} = \frac{\text{nombre de centres d'intérêt prédits dans la dimension utilisateur}}{\text{nombre de centres d'intérêt calculés dans la dimension sociale}} \quad (8)$$

Le rappel d'un algorithme de dérivation de la dimension sociale quant à lui est évalué par le nombre de centres d'intérêt trouvés dans la dimension utilisateur par rapport au nombre de centres d'intérêt de la dimension utilisateur.

$$\text{Rappel} = \frac{\text{nombre de centres d'intérêt prédits dans la dimension utilisateur}}{\text{nombre total de centres d'intérêt dans la dimension utilisateur}} \quad (9)$$

Pour calculer la précision et le rappel, nous nous intéressons uniquement aux centres d'intérêt les plus pertinents renvoyés par chaque algorithme de dérivation de la dimension sociale. Si la dimension utilisateur d'un profil possède n centres d'intérêt, la précision et le rappel de chaque algorithme de dérivation de la dimension sociale seront calculés à partir du top $n+m$ premiers centres d'intérêt, donc un vecteur de taille $n+m$ pour représenter la dimension sociale.

6.4. Résultats

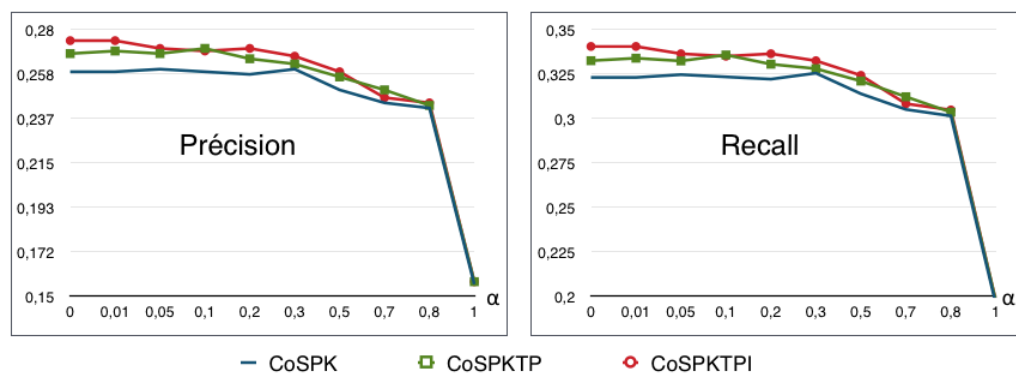


Figure 2 : Comparatif de la pertinence des dimensions sociales construites avec CoSPK, CoSPKTP et CoSPKTPI avec la dimension utilisateur

La figure 2 représente les résultats permettant de comparer la pertinence des différentes dimensions sociales avec le top 10 des centres d'intérêt de l'utilisateur. Dans cette figure, la valeur α représente la proportion entre la mesure sémantique et la mesure structurelle comme présentée dans la formule (1). Pour obtenir les résultats de la figure 2, après plusieurs expérimentations avec différentes valeurs de γ et λ , nous avons calculé les poids temporels dans la formule (7) avec la valeur de $\gamma = 0,75$ et $\lambda = 0,2$ calculée avec la formule $(1/((0,5)*f(0)))$ où $f(0) = 10$. Nous pouvons observer que l'algorithme CoSPKTP produit de meilleurs résultats que l'algorithme CoSPK en termes de précision et de rappel. Le meilleur résultat peut être observé quand $\alpha = 0,1$ avec successivement 4,3 % et 3,8% de taux de gain en terme de précision et rappel par rapport à l'algorithme CoSPK. Cela nous montre dans un premier temps, l'intérêt d'utiliser une méthode de prédiction de liens temporelle dans le processus de construction de la dimension sociale du profil utilisateur à partir de son réseau social.

Quand nous appliquons l'algorithme CoSPKTPI, il produit de meilleurs résultats par rapport à l'algorithme CoSPKTP et donne en conséquence un écart plus important comparé à l'algorithme CoSPK, en particulier quand $\alpha \in [0 ; 0,01]$, avec successivement 5,8 % et 5,4% de taux de gain en terme de précision et de rappel par

rapport à l'algorithme CoSPK et 1,9% et 2,0 % par rapport à l'algorithme CoSPKTP. Cela nous montre l'importance de la prise en compte de critères temporels non seulement dans la sélection de sources d'informations liées à la structure du réseau social de l'utilisateur mais aussi dans le traitement des informations partagées dans le réseau. Ces résultats nous permettent également de renforcer l'importance de la prise en compte de la dynamique du partage et de la diffusion d'informations tout comme de la dynamique de la structure du réseau dans l'étude de l'évolution du réseau social de l'utilisateur.

Algorithme	CoSPK		CoSPKTP	
	Précision	Rappel	Précision	Rappel
CoSPKTP ($\alpha = 0,1$)	4,3%	3,8%	-	-
CoSPKTPI ($\alpha \in [0 ; 0,01]$)	5,8%	5,4%	1,9%	2,0%

Tableau 1. Taux de gain (écarts entre les courbes) des algorithmes proposés (CoSPKTP et CoSPKTPI) par rapport l'algorithme existant (CoSPK)

6. Conclusion et perspectives

Dans ce travail, nous proposons de prendre en compte la caractéristique dynamique des réseaux sociaux dans l'étape d'extraction des intérêts du processus de construction de la dimension sociale du profil de l'utilisateur. Dans ce processus, nous intégrons des critères temporels (poids temporel) pour pondérer les individus et les informations du réseau social de l'utilisateur afin de pouvoir extraire des intérêts de l'utilisateur pertinents et à jour. Ce poids temporel est calculé, à partir du poids de pertinence temporelle des individus et du poids de pertinence temporelle des informations que ces individus partagent. Le poids de pertinence des individus est calculé en appliquant une méthode de prédiction de liens temporelle afin de sélectionner les individus ayant les liens les plus actifs avec l'utilisateur. Le poids de pertinence temporelle des informations partagées est calculé en prenant en compte la fraîcheur des informations. Les résultats des expérimentations nous ont permis de montrer l'efficacité de cette approche par rapport au processus de construction de la dimension sociale qui ne tient pas compte de critères temporels.

A court terme, nous envisageons d'appliquer cette approche dans d'autres types de réseaux sociaux où la caractéristique dynamique est très importante (Facebook, Twitter, ...) pour effectuer une évaluation de notre proposition à plus grande échelle. Nous souhaitons également étudier d'autres algorithmes de prédiction de liens temporelle ainsi que d'autres fonctions temporelles afin de tenter d'améliorer la performance de notre approche. A plus long terme, nous envisageons la mise à jour continue du profil utilisateur afin d'avoir un profil pertinent et à jour à tout moment de son utilisation par des mécanismes d'adaptation de l'information par exemple.

Remerciements

Projet Subventionné par la Communauté de travail pyrénéenne



Bibliographie

- Bennett, P. N., R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, et X. Cui. (2012). Modeling the Impact of Short- and Long-term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 185–194. New York, NY, USA: ACM.
- Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics* 87, 597-620.
- Carmel, D., N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, et al. (2009). Personalized Social Search Based on the User's Social Network. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p. 1227–1236. New York, NY, USA: ACM.
- Cheng, Y., G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, et C. Chen. (2008). Model Bloggers' Interests Based on Forgetting Mechanism. In *Proceedings of the 17th International Conference on World Wide Web*, p. 1129–1130. New York, NY, USA: ACM.
- Crabtree, B., S. Soltysiak, M. Pp, et I. Re. (1998). Identifying and tracking changing interests. *International Journal on Digital Libraries* 2, 38–53.
- Ding, Y. et X. Li. (2005). Time Weight Collaborative Filtering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, p. 485–492. New York, NY, USA: ACM.
- Gomez Rodriguez, M., J. Leskovec, et B. Schölkopf. (2013). Structure and Dynamics of Information Pathways in Online Media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, p. 23–32. New York, NY, USA: ACM.
- Hasan, M. A. et M. J. Zaki. (2011). A Survey of Link Prediction in Social Networks. In C. C. Aggarwal (Ed.), *Social Network Data Analytics*, p. 243-275. Springer US.
- Jiang, C., Y. Chen, et K. J. R. Liu. (2014). Modeling information diffusion dynamics over social networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1095-1099.
- Kacem, A., M. Boughanem, et R. Faiz. (2014). Time-Sensitive User Profile for Optimizing Search Personalization. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, et G.-J.

- Houben (Eds.), *User Modeling, Adaptation, and Personalization*, p. 111-121. Springer International Publishing.
- Kumar, R., J. Novak, et A. Tomkins. (2006). Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 611–617. New York, NY, USA: ACM.
- Leskovec, J., L. Backstrom, R. Kumar, et A. Tomkins. (2008). Microscopic Evolution of Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 462–470. New York, NY, USA: ACM.
- Liben-Nowell, D. et J. Kleinberg. (2003). The Link Prediction Problem for Social Networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, p. 556–559. New York, NY, USA: ACM.
- Li, D., P. Cao, Y. Guo, et M. Lei. (2013). Time Weight Update Model Based on the Memory Principle in Collaborative Filtering. *Journal of Computers* 8.
- Maloof, M. A. et R. S. Michalski. (2000). Selecting Examples for Partial Memory Learning. *Machine Learning* 41, 27-52.
- Massa, P. et P. Avesani. (2007). Trust-aware Recommender Systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, p. 17–24. New York, NY, USA: ACM.
- Spiliopoulou, M. (2011). Evolution in Social Networks: A Survey. In C. C. Aggarwal (Ed.), *Social Network Data Analytics*, p. 149-175. Springer US.
- Stattner, E., M. Collard, et N. Vidot. (2013). D2SNet: Dynamics of diffusion and dynamic human behaviour in social networks. *Computers in Human Behavior* 29, 496-509.
- Tan, B., X. Shen, et C. Zhai. (2006). Mining Long-term Search History to Improve Search Accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 718–723. New York, NY, USA: ACM.
- Tchunte, D. (2013, janvier 28). *Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques* (phd). Université de Toulouse, Université Toulouse III - Paul Sabatier.
- Tylenda, T., R. Angelova, et S. Bedathur. (2009). Towards Time-aware Link Prediction in Evolving Social Networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, p. 9:1–9:10. New York, NY, USA: ACM.
- Weng, L., J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, et al. (2013). The Role of Information Diffusion in the Evolution of Social Networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 356–364. New York, NY, USA: ACM.
- Zheng, N. et Q. Li. (2011). A recommender system based on tag and time information for social tagging systems. *Expert Systems with Applications* 38, 4575-4587.

Détection des intérêts d'un utilisateur par l'exploitation du comportement d'annotation de son réseau égocentrique

**Manel Mezghani^{1,2}, André Péninou², Corinne Amel Zayani¹,
Ikram Amous¹, Florence Sèdes²**

1. Laboratoire MIRACL, Université de Sfax, 3021 SFAX, Tunisie

mezghani.manel@gmail.com, {corinne.zayani, ikram.amous}@isecs.rnu.tn

*2. Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse,
CNRS, INPT, UPS, UT1, UT2J, 31062 TOULOUSE Cedex 9, France*

{mezghanni.manel, andre.peninou, florence.sedes}@irit.fr

RÉSUMÉ. Les médias sociaux fournissent un environnement d'échange et reposent principalement sur leurs utilisateurs dont le rôle est de créer, d'annoter le contenu des ressources et de construire des relations avec d'autres utilisateurs. Nous nous concentrons sur l'analyse de ces environnements sociaux afin de détecter les intérêts des utilisateurs qui sont des éléments clés pour améliorer l'accès à l'information. L'originalité de notre approche est basée sur la proposition d'une nouvelle technique de détection des intérêts en analysant, d'une part le réseau égocentrique d'un utilisateur, et, d'autre part, la précision du comportement d'annotation des utilisateurs dans le but de sélectionner les tags qui reflètent réellement des intérêts. L'approche proposée a été testée et évaluée sur la base de données sociales Delicious. Une évaluation comparative avec une méthode classique (basée sur les tags) de détection des intérêts montre que notre approche donne de meilleurs résultats.

ABSTRACT. Social media provide an environment of information exchange. They principally rely on their users to create, to annotate content and to make on-line relationships. We focus on analysing this social environment to detect user interests which are the key elements for improving adaptation. The originality of our approach is based on the proposal of a new technique of interests detection by analysing the egocentric network of the user and the accuracy of the tagging behaviour of a user in order to figure out the tags which really reflect the content of the resources. The proposed approach has been tested and evaluated in the Delicious social database. A comparative evaluation with the classical tag-based approach of interests detection shows that the proposed approach is better.

MOTS-CLÉS : Profil utilisateur, intérêts, réseaux sociaux, indexation, comportement d'annotation.

KEYWORDS: User profile, interests, social network, indexation, tagging behaviour.

1. Introduction

L'avènement du web 2.0, centré utilisateur, a fait émerger une quantité très importante d'informations. Souvent partagées dans les médias sociaux, ces informations constituent un moyen pour guider les autres utilisateurs vers l'information recherchée. Cet aspect collaboratif de partage d'information a servi dans plusieurs applications comme le e-commerce, le e-learning, etc.

Plusieurs techniques ont été développées afin de mieux se servir de cette connaissance collective. Parmi ces techniques, citons l'adaptation qui permet de fournir à l'utilisateur une information adaptée à ses besoins. L'adaptation est donc un processus fortement lié à l'utilisateur puisqu'on adapte l'information selon ses besoins spécifiques. Un profil utilisateur qui reflète les intérêts réels de l'utilisateur permet d'améliorer l'adaptation et ainsi d'éviter la surcharge cognitive et la désorientation de l'utilisateur pendant son accès à l'espace d'information. Nous considérons dans cet article qu'un profil utilisateur est un vecteur de mots-clés (intérêts).

De manière classique, les intérêts d'un utilisateur sont extraits de son propre profil (par exemple de l'attribut intérêt). Toutefois, l'utilisateur généralement ne donne pas toutes les informations relatives à ses intérêts. Donc, le profil explicite de l'utilisateur ne peut jamais être considéré comme entièrement connu par un système. Ainsi, il est difficile de s'appuyer sur la seule analyse du profil pour détecter les intérêts réels (Tchunte *et al.*, 2013).

Notre approche est construite à partir de l'hypothèse que l'environnement social, et en particulier les personnes proches d'un individu, fournissent une information à partir de laquelle les intérêts de cet individu peuvent être extraits. Cette hypothèse a été prouvée dans le contexte de dérivation des profils utilisateurs des réseaux sociaux (Tchunte *et al.*, 2013). Les personnes proches peuvent être celles partageant certains comportements communs (par exemple en visitant ou en annotant la même ressource), le réseau égocentrique¹, les utilisateurs appartenant à une même communauté, etc. Donc, nous analysons les personnes proches afin de détecter les intérêts les plus pertinents pour chaque utilisateur. Nous considérons dans ce papier les personnes proches comme celles qui appartiennent au réseau égocentrique d'un utilisateur.

Nous nous concentrons sur le comportement d'annotation des personnes du réseau égocentrique, qui, reflète l'opinion de ces utilisateurs sur une ressource (Astrain *et al.*, 2010). Ce comportement est défini par l'action d'annoter une ressource par un utilisateur. Il est représenté sous la forme de tuples <Utilisateur, Tag, Ressource>. Cette information a prouvé son utilité pour détecter les intérêts des utilisateurs (Meo *et al.*, 2010) (Kim *et al.*, 2011). Un tag est une annotation sociale générée par un utilisateur, qui reflète son intérêt sur une ressource. Malgré l'importance de cette information, elle peut être ambiguë. Nous utilisons aussi le contenu des ressources annotées afin d'éliminer des tags ambigus (opinions, spam, etc.).

L'approche proposée traite principalement les ressources textuelles (semi-structurées, texte, etc.) qui sont présents dans presque tous les principaux médias sociaux tels que :

1. Réseau social d'un utilisateur (égo) réduit aux utilisateurs avec qui il (l'égo) est en relation directe.

Delicious en analysant les URLs annotées, *Twitter* en analysant les *tweets*, etc. Notre approche ne traite pas les autres médias (les images dans le cas de *Flickr*, par exemple). Notre approche est expérimentée sur la base de données sociales *Delicious*. Nos résultats sont comparés à l'approche utilisant directement les tags fournis par les utilisateurs (approche classique basée sur les tags).

Le reste de cet article est structuré comme suit. Dans la deuxième section, nous présentons notre positionnement par rapport aux éléments d'information sur le comportement d'annotation (utilisateur, tag et ressource). Dans la troisième section, nous présentons et décrivons l'approche proposée qui s'appuie principalement sur un filtrage des tags à partir des ressources et le réseau égocentrique de chaque utilisateur. Dans la quatrième section, nous présentons et commentons les résultats de notre expérimentation sur la base sociale *Delicious*. Dans la dernière section, nous concluons et présentons les perspectives de notre travail.

2. Positionnement

Un système d'adaptation social efficace essaie de détecter les intérêts des utilisateurs à l'aide de données sociales pertinentes. Mais les intérêts estimés d'un utilisateur peuvent être considérés comme non pertinents, en raison de l'information inappropriée utilisée pour les détecter. Pour surmonter ce problème, notre approche fait une utilisation sélective de l'information disponible dans l'objectif de construire une liste d'intérêts précise pour chaque utilisateur. Nous détaillons notre analyse par rapport aux éléments du comportement d'annotation ci-dessous :

– **Utilisateur** : Afin de développer notre approche, nous analysons le comportement d'annotation des personnes proches (le réseau égocentrique) de chaque utilisateur. Ce choix est motivé par :

a) les études qui favorisent la connaissance collective pour détecter les intérêts des utilisateurs (Meo *et al.*, 2010), (Kim *et al.*, 2011), (Tchuente *et al.*, 2013), (On-at *et al.*, 2014) et (Zhou *et al.*, 2010).

b) l'absence des informations dans le profil utilisateur explicite. En effet, l'utilisateur ne fournit pas toutes les informations relatives à ses intérêts. Donc son profil ne peut jamais être considéré comme une information suffisante pour connaître ses intérêts (Tchuente *et al.*, 2013).

c) l'inefficacité de l'information déduite de son comportement classique. En fait, ce dernier ne reflète pas toujours les vrais intérêts de l'utilisateur. Par exemple, l'analyse du comportement de navigation de l'utilisateur selon (Ma *et al.*, 2011) : i) conduit à l'analyse du comportement antérieur qui peut ne pas refléter ses intérêts actuels, ii) n'est pas toujours un indicateur efficace puisque l'utilisateur peut accéder à une page web sans avoir un intérêt pour son contenu.

– **Tag** : Dans la plupart des travaux analysant le comportement d'annotation, les intérêts sont détectés à partir des tags. Cette détection est basée sur des mesures de popularité des tags ou d'analyse des tags (par analyse de la sémantique des tags, par

exemple) (Milicevic *et al.*, 2010) (Mezghani *et al.*, 2012). Ces analyses peuvent fournir des tags pertinents pour l'utilisateur. Mais, selon (Milicevic *et al.*, 2010), le problème lié à certains tags est qu'ils sont spécifiques à l'utilisateur. En effet, ces tags ne décrivent pas le document mais plutôt l'avis de l'utilisateur comme par exemple "j'aime", "sympa", "nul", etc. Selon (Milicevic *et al.*, 2010), l'ambiguïté d'un tag est qu'un seul tag a de nombreuses significations et peut faussement donner l'impression que deux ressources sont similaires quand elles sont en fait sans rapport. Ainsi, le filtrage des tags pourrait être une solution pour surmonter les tags ambigus. Donc, dans un but de détecter des intérêts pertinents, nous allons essayer de détecter des tags significatifs (compréhensibles) plutôt que les tags spécifiques à l'utilisateur.

– **Ressource** : Généralement, les approches traitant cette information utilisent des techniques telles que le *clustering*, le traitement sémantique, etc. (Ma *et al.*, 2011). Cependant, analyser seulement le contenu de la ressource ne permet pas d'avoir des informations sur les intérêts de l'utilisateur (Ma *et al.*, 2011). Le contenu des ressources peut permettre de détecter la nature des tags qui lui sont associés. Cependant, la plupart des recherches ne considèrent pas l'exactitude des tags avec le contenu de la ressource selon (Zhou *et al.*, 2010). Contrairement à la plupart de ces recherches, nous nous concentrons sur l'analyse de l'exactitude des tags par rapport au contenu des ressources pour surmonter les problèmes liés à la nature des annotations sociales.

Pour résumer, notre approche tente de combiner les informations utilisateur, tag et ressource d'une manière qui cherche à garantir une détection des intérêts pertinents. Notre approche utilise les tags des personnes proches et les traite en fonction du contenu de leurs ressources respectives. Les tags considérés comme des intérêts pertinents sont ceux qui reflètent le contenu des ressources auxquelles ils ont été associés.

3. Description de l'approche de détection des intérêts

Le processus de détection des intérêts que nous proposons analyse, d'une part le réseau égocentrique d'un utilisateur, et, d'autre part, la précision du comportement d'annotation d'un utilisateur dans le but de sélectionner les tags qui reflètent réellement le contenu des ressources. Le filtrage des tags qui reflètent réellement le contenu des ressources est fait en plusieurs étapes :

- i) Pour chaque tag de chaque utilisateur, construction de l'ensemble des ressources pertinentes pour ce tag. Cette construction est effectuée par l'analyse de l'ensemble de toutes les ressources pour un tag donné.
- ii) Pour chaque tag, attribution d'un score à ces ressources et sélection des top-k ressources.
- iii) Filtrage des tags : si une ressource associée au tag est dans le top-k des ressources pertinentes pour ce tag, alors le tag est retenu.

Le fait d'utiliser toutes les ressources existantes pour chercher les ressources pertinentes pour un tag (et non pas les seules ressources auxquelles ce tag est associé) doit permettre de réellement analyser la pertinence des tags. Un tag ne sera retenu que s'il

est associé à une ressource qui appartient à l'ensemble des ressources auxquelles il correspond le mieux (calcul des top-k ressources pour chaque tag). Nous cherchons ainsi à analyser la précision du comportement d'annotation d'un utilisateur dans le but de sélectionner les tags qui reflètent réellement le contenu des ressources et reflètent le mieux les intérêts de l'utilisateur.

L'approche de détection des intérêts est effectuée selon deux étapes principales détaillées ci-après. D'abord, nous préparons les données que nous allons utiliser. Ensuite, nous détaillons le processus de détection des intérêts (filtrage des tags).

Pour la suite de l'article, notons :

- $U = \{u_1, \dots, u_n\}$, l'ensemble des utilisateurs dans le réseau social, où n est le nombre d'utilisateurs.
- $R = \{r_1, \dots, r_m\}$, l'ensemble de toutes les ressources dans le réseau social, où m est le nombre de ressources.
- $T = \{t_1, \dots, t_h\}$, l'ensemble des tags, où h est le nombre des tags.
- $N_u = \{n_{u1}, \dots, n_{uj}\}$, l'ensemble des personnes proches de l'utilisateur u , où j est le nombre de personnes proches de l'utilisateur $u \in U$.
- $I_u = \{i_{u1}, \dots, i_{uk}\}$, l'ensemble d'intérêts pertinents pour l'utilisateur u , où k est le nombre des intérêts pertinents de l'utilisateur $u \in U$. Ceci est le résultat construit par notre algorithme.

3.1. Préparation des données

Avant d'expliquer notre processus de détection des intérêts, nous préparons les données utilisées comme une entrée pour détecter les intérêts des utilisateurs. La figure 1 illustre cette étape de préparation de données.

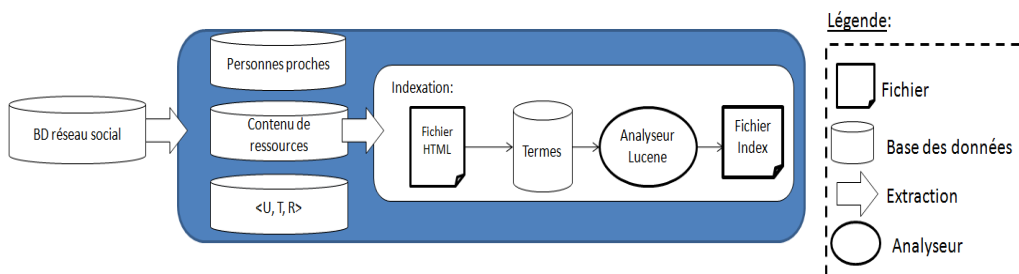


Figure 1. Préparation des données

Nous extrayons les données relatives au :

- Comportement d'annotation $\langle U, T, R \rangle$, qui est constitués par les tags appliquées aux ressources par les utilisateurs.
- Personnes proches N_u (le réseau égocentrique).
- Contenu des ressources.

Après avoir extrait les données, nous indexons les ressources extraites. L'indexation vise à décrire le contenu d'une ressource par des mots-clés. Les ressources sont indexées, en utilisant l'API *Lucene*². *Lucene* est capable de traiter de grands volumes de ressources grâce à sa puissance et à sa rapidité dues à l'indexation. *Lucene* est un outil d'indexation basé sur les champs. Cette caractéristique permet l'indexation des ressources selon un ou plusieurs champs. Par exemple, les champs peuvent être le *titre*, le *contenu*, le *URL*, etc.

Les étapes de cette indexation sont les suivantes (cf. figure 1) : *Lucene* indexe les ressources avec un parseur³ en les divisant en un certain nombre de termes en utilisant un analyseur. Puis, il stocke les termes dans un fichier d'index, où chaque terme est associé au contenu de la ressource.

L'index est composé de segments, pouvant être considéré comme des sous-index bien qu'ils ne soient pas entièrement indépendants. *Lucene* assigne à chaque ressource de l'index un identifiant unique. Les segments conservent les éléments suivants : 1) les noms des champs utilisés dans l'index, 2) un dictionnaire des termes : les termes contenus dans chaque champ, 3) la fréquence des termes : numéros de tous les ressources contenant ce terme et 4) proximité des termes : la position de chaque terme.

3.2. Processus de détection des intérêts

Le processus de détection des intérêts procède pour chaque personne proche ($n_{uj} \in N_u$) d'un utilisateur donné ($u \in U$) comme suit : D'abord, il génère les ressources pertinentes (R') pour chaque tag ($t_h \in T$). Cette génération a pour but de sélectionner les tags qui reflètent le contenu de la ressource. Puis, il score ces ressources afin de ne garder que les ressources les plus pertinentes (R''). Enfin, il sélectionne les tags pertinents qui sont associés à la fois à la ressource annotée par u et à la liste des ressources pertinentes (R''). Ces tags sont considérés comme étant des intérêts de l'utilisateur (I_u) puisqu'ils reflètent vraiment le contenu de la ressource. Ce processus est illustré dans la figure 2.

3.2.1. Détection de l'exactitude du tag par rapport à la ressource

Nous commençons par générer les ressources pertinentes R' pour chaque tag donné, où $R' = \{r'_1, \dots, r'_v\}$ est l'ensemble des ressources pertinentes et v est le nombre de ressources pertinentes et $R' \subseteq R$.

Cette étape interroge le fichier index (la sortie de l'étape d'indexation). Lorsqu'une requête est faite, elle est traitée par le même analyseur utilisé pour créer l'index et ensuite utilisé pour trouver le(s) expression(s) correspondante(s) dans l'index. Ceci fournit une liste de ressources correspondant à la requête. Dans notre contexte, une requête est considérée comme un tag dans le reste de ce papier.

2. <http://lucene.apache.org/core/>

3. analyseur syntaxique qui étiquette les mots d'un texte

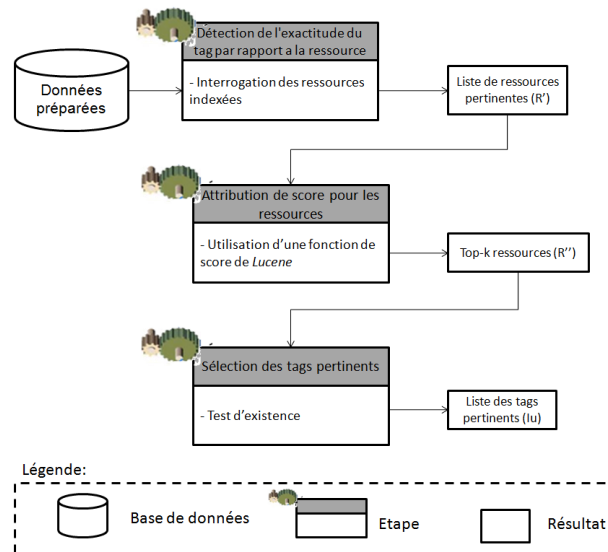


Figure 2. Processus de détection des intérêts pour un utilisateur

3.2.2. Attribution de score pour les ressources

Après la génération des ressources pertinentes (R') pour chaque tag (t_h), un score est attribué à chaque ressource pertinente. Le but de l'utilisation de tel score est de sélectionner les ressources les plus pertinentes liées à un tag. Ce score est le résultat d'une fonction de similarité qui prend en considération la ressource et le tag. De nombreuses fonctions de similarité existent dans la littérature telle que la fonction de similarité soutenue par *Lucene*. Nous choisissons une fonction prédéfinie⁴ de similarité qui est une variante du modèle de notation TF-IDF. Le choix d'un tel modèle est dû au fait que TF-IDF est un algorithme simple et efficace pour faire correspondre des mots dans une requête aux ressources qui sont pertinentes à cette requête (dans notre cas un tag).

Cette fonction fournit les top-k ressources pertinentes (R'') pour un tag, où $R'' = \{r''_1, \dots, r''_w\}$ est l'ensemble des top-k ressources pertinentes et w est le nombre de ressources pertinentes et $R'' \subseteq R'$.

La valeur de k est déterminante dans notre cas. Elle est aussi très dépendante de la base de test utilisée. Cette valeur sera choisie dans la partie expérimentation.

3.2.3. Sélection des tags pertinents

Après avoir généré les top-k ressources pertinentes pour un tag t_h (d'une personne proche), nous testons si la ressource annotée par t_h (la ressource annotée directement par l'utilisateur) existe dans le résultat top-k fourni par la fonction de score. Si c'est le cas, le tag t_h est considéré comme pertinent pour la ressource (puisque'il reflète vraiment son contenu).

4. http://lucene.apache.org/core/3_5_0/scoring.html

Cette étape génère une liste des intérêts pertinents (I_u) sous la forme d'une liste de tags qui décrivent au mieux le contenu de la ressource annotée. Cette liste est issue de l'analyse des personnes proches (le réseau égo-centrique) pour chaque utilisateur.

3.3. Processus de validation

Afin de valider notre approche, nous considérons les utilisateurs ayant un profil connu (qui ont déjà eu des activités dans le réseau). Dans une approche classique, nous considérons le profil utilisateur comme étant la liste de tags affectés par l'utilisateur. Donc, nous comparons les tags de l'utilisateur (issus de son profil) avec des tags fournis par notre approche (issus du réseau égo-centrique).

De notre analyse sociale nous avons construit, pour chaque utilisateur, une liste d'intérêts (tags). Cette liste est validée par sa comparaison avec les intérêts de cet utilisateur. Le processus d'évaluation est décrit à travers la figure 3.

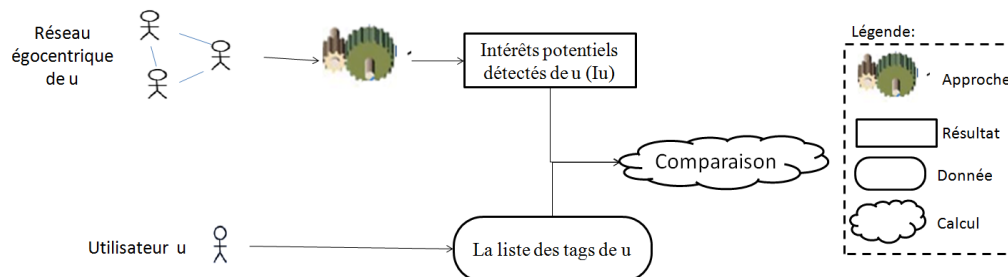


Figure 3. Processus de validation de l'approche de détection des intérêts

Pour chaque utilisateur cible $u \in U$, notre approche construit un ensemble $I_u = \{i_{u1}, \dots, i_{uk}\}$ d'intérêts potentiellement pertinents. Pour chaque $i_{uk} \in I_u$, nous analysons l'existence de l'intérêt i_{uk} dans le profil de l'utilisateur cible u . Les intérêts corrects sont considérés comme l'ensemble de tous les intérêts $C_u \subset I_u$, où $C_u = \{c_{u1}, \dots, c_{uy}\}$ et y est le nombre d'intérêts présents à la fois dans I_u et dans le profil utilisateur de u .

La validation de notre proposition, se fait par un test d'existence des intérêts des utilisateurs dans les intérêts potentiels calculés par notre approche. Ce test est effectué à travers deux méthodes :

- Par une simple comparaison des tags (comparaison exacte) : par exemple si le tag de l'utilisateur= "image" et tag de la personne proche= "image", alors "image" est considéré comme un tag pertinent. Nous appellerons cette technique dans le reste de cet article comme "simple comparaison".

- En prenant en compte les synonymes ou les mots reliés : par exemple si le tag de l'utilisateur= "image" et tag de la personne proche= "photo", alors "image" est considéré comme un tag pertinent. Les synonymes ou mots reliés sont détectés en

interrogeant Wordnet⁵. Nous appellerons cette technique dans le reste de cet article comme "avec synonymes ou mots reliés".

4. Expérimentations sur *Delicious*

Dans cette section, nous détaillons d'abord la base de test utilisée dans la section 4.1. Ensuite, nous présentons les mesures utilisées pour nos calculs dans la section 4.2. Enfin, nous détaillons les évaluations faites afin de tester l'efficacité de notre approche. En effet, notre approche est évaluée selon deux critères :

1. Nous évaluons dans la section 4.3 notre approche selon les deux méthodes de validation : i) comparaison simple ou ii) avec synonymes ou mots reliés. Nous avons testé aussi l'influence de la valeur de k qui sélectionne les top-k ressources pertinents à un tag. Nous conservons les valeurs qui donnent de meilleurs résultats pour faire le reste des évaluations.

2. Nous comparons dans la section 4.4 notre approche avec l'approche qui utilise les informations des tags du réseau égocentrique sans pré-traitement (approche classique basée tags).

Il est à noter que pour l'étape de l'indexation, nous avons pris en compte uniquement le champ *contenu*, vu la richesse de l'information présente.

4.1. Base de test

Nous avons évalué notre approche sur la base de test *Delicious*. Cette base de données est extraite de (Ivan *et al.*, 2011). La base de données *Delicious* contient le réseau égocentrique de chaque utilisateur, des marques-pages des utilisateurs et des tags des utilisateurs. Les utilisateurs U sont décrits par leur ID⁶ par exemple : *UserID=8*. Les ressources R sont décrites par leur ID, titre et l'URL par exemple : *1 IFLA - le site Web officiel des Internationaux Fédération d'Associations de Bibliothèque et Institutions <http://www.ifla.org/>*. Les tags T sont décrits par leur ID et valeur par exemple : *1 developpement*. La base de test contient :

- 1867 utilisateurs, 7668 relations bidirectionnelles et une moyenne de 8.236 relations par utilisateur.
- 69226 URLs dont 38581 URLs principales, ex. : www.delicious.com
- 53388 tags, 437593 tag *assignments* (tas), sous forme de tuples [user, tag, URL], et une moyenne de 234.383 tas par URL et une moyenne de 6.321 tas par tags
- 104799 bookmarks, une moyenne de 56.132 URLs annotées par utilisateur et une moyenne de 1.514 utilisateur annotant une URL.

5. <http://wordnet.princeton.edu/>

6. Identifiant

4.2. Mesures

Dans cette section, nous présentons les mesures utilisées dans notre évaluation.

- **Précision et précision moyenne** : Nous calculons la précision des intérêts détectés selon les intérêts produits par notre approche (cf. formule 1). La précision $Precision(u)$ pour chaque utilisateur $u \in U$ est calculée selon le nombre de tags précis ($C_u \subset I_u$), qui existent à la fois dans le profil de l'utilisateur et les profils des personnes proches (le réseau égocentrique), et le nombre total des tags trouvés (I_u) :

$$Precision(u) = \frac{|C_u|}{|I_u|} \quad (1)$$

Nous calculons également la précision moyenne pour tous les utilisateurs (cf. formule 2) fournie à partir de la formule de précision $Precision(u)$ (cf. formule 1) pour un utilisateur u , où n est le nombre d'utilisateurs (dans notre cas, $n = 1867$) :

$$Precision_moyenne = \frac{\sum_{i=1}^n P(u)}{n} \quad (2)$$

- **Boîtes de Tukey (box plot)** : Ces boîtes reflètent la distribution des valeurs de précision dans les résultats (selon quatre quartiles). Elles sont plus représentatives qu'une simple moyenne des précisions. Un exemple explicatif de boîtes de Tukey est présenté dans la figure 4. L'extrémité supérieure de la ligne continue représente la valeur maximale des valeurs obtenues, tandis que l'extrémité inférieure représente la valeur minimale. Concernant le rectangle, il récupère toutes les valeurs situées entre le premier et le troisième quartile (Q3). C'est les valeurs de 25 % des données qui sont situées en dessous du premier quartile (Q1), et 25 % des données qui sont situées au-dessus du troisième quartile. L'écart interquartile correspond donc à 50 % des valeurs situées dans la partie centrale de la distribution. Il est donc utilisé comme indicateur de la dispersion.

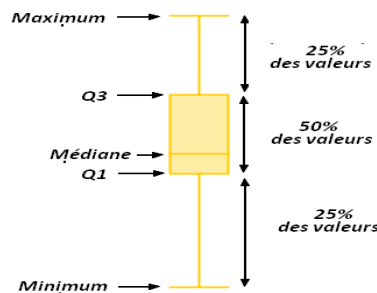


Figure 4. Un exemple de boîtes de Tukey

4.3. Évaluation de notre approche

Nous testons notre approche sur tous les utilisateurs de la base de données. Ces utilisateurs ont un nombre différent de personnes proches (entre 1 et 90 ami(s) explicite(s)). Le nombre de tags, de ressources et de comportement d'annotation (tas) est

différent pour chaque utilisateur. Ce nombre peut varier d'environ 3 à 800 pour les tags, de 10 à 450 pour les ressources, et de 20 à 500 pour les comportements d'annotation.

Notre approche a été testée avec différentes valeurs de k (qui sélectionnent les top- k ressources) tel que $k = 20$, $k = 50$ et $k = 100$. Nous calculons la précision moyenne de notre approche pour les deux méthodes d'évaluation : "simple comparaison" et "avec synonymes ou mots reliés" (cf. figure 5).

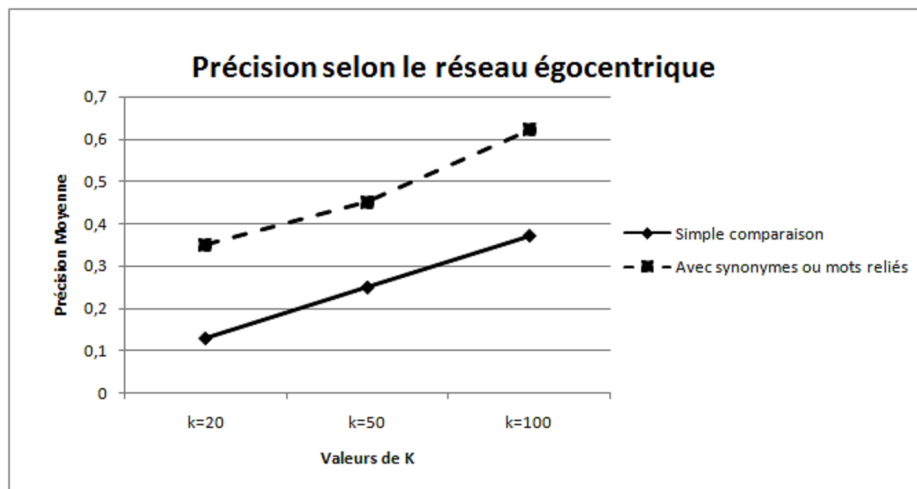


Figure 5. La précision moyenne selon $k = 20$, $k = 50$ et $k = 100$: (bleu) selon une simple comparaison, (rouge) selon les synonymes ou mots reliés

Nous voyons clairement que la précision qui prend en considération les synonymes ou mots reliés donne de meilleurs résultats que la technique avec simple comparaison (cf. figure 5). Ceci est un résultat attendu puisque les utilisateurs peuvent avoir les mêmes intérêts (tags), mais ils peuvent les décrire différemment en utilisant différents tags.

Nous choisissons $k = 100$ pour le reste de l'évaluation, car cette valeur donne de meilleurs résultats. Nous présentons le résultat de la précision moyenne selon les deux techniques d'évaluation (pour $k=100$) à travers une représentation en boîtes de Tukey (*box plot*) dans la figure 6. Nous remarquons que :

- Pour la précision selon les synonymes ou mots reliés, la distribution est presque centrée sur la moyenne. Ceci reflète que la plupart des utilisateurs ont la même précision moyenne.
- Pour la précision en fonction de la technique de simple comparaison, la distribution est inférieure à la répartition des synonymes ou mots reliés. Ceci reflète que la plupart des utilisateurs ont des valeurs de précisions assez basses.

Afin de comprendre mieux ces résultats, nous avons choisi de représenter un échantillon de 20 utilisateurs choisis au hasard (comme nous ne pouvions pas montrer la précision des 1876 utilisateurs dans une même figure). Dans cette figure, nous détaillons

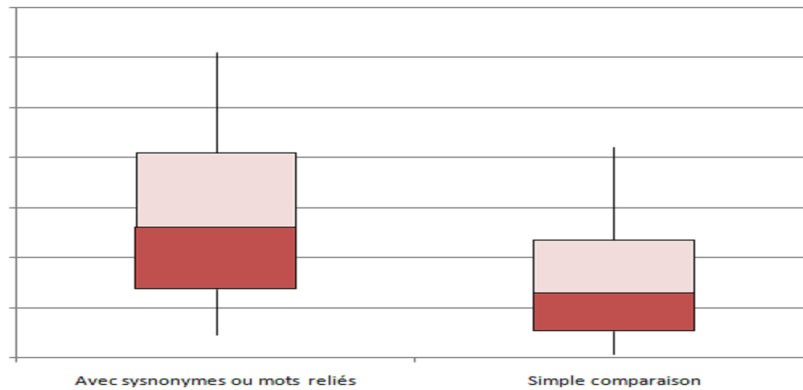


Figure 6. Répartition des précisions selon des boîtes de Tukey de notre approche en fonction du réseau égocentrique

les différentes valeurs de précision calculées par la technique de simple comparaison et aussi en tenant compte des synonymes ou mots reliés. La figure 7 montre les valeurs de précision pour cet ensemble de 20 utilisateurs.

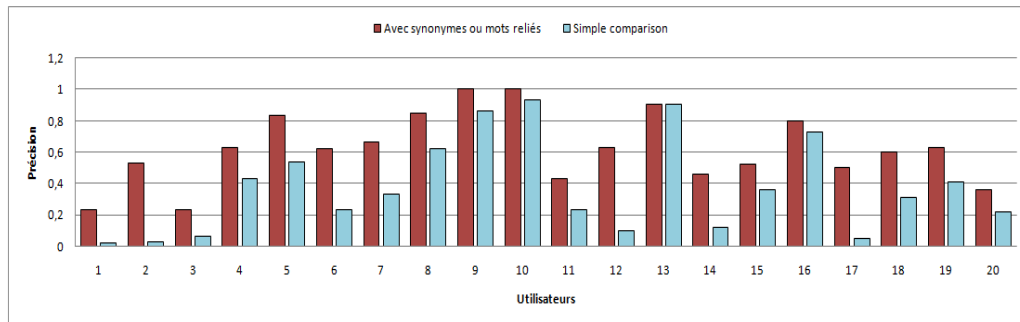


Figure 7. Précision des intérêts détectés pour un ensemble de 20 utilisateurs ($k = 100$)

De cette évaluation (cf. figure 7), nous voyons clairement que la précision qui prend en considération les synonymes ou mots reliés est généralement meilleure que la technique de simple comparaison.

De tout ensemble d'utilisateurs, nous remarquons que la précision (pour les deux méthodes de calcul) varie selon trois cas : i) la précision est plus élevée pour les utilisateurs actifs (ayant beaucoup de personnes proches et beaucoup de comportements d'annotation). Ces utilisateurs actifs ont une précision de 79% selon la méthode avec synonymes ou mots reliés et de 68% pour la méthode simple comparaison. ii) la précision est moins élevée pour les utilisateurs moins actifs. Ces utilisateurs moins actifs ont une précision de 26% selon la méthode avec synonymes ou mots reliés et de 18% pour la méthode simple comparaison. iii) la précision est égale à zéro lorsque l'écart du nombre des tags fourni par l'utilisateur par rapport à ses personnes proches est important. Par exemple, d'une part, le nombre de comportements d'annotation est faible (ex. : 20) pour un utilisateur donné. D'autre part, le nombre de comportements d'an-

notation de toutes ses personnes proches est important (ex. : 200). Cette différence contribue à réduire les taux de précisions.

De plus, nous avons voulu tester si notre approche a fourni des tags non ambigus. Nous jugeons qu'un tag est compréhensible (resp. ambigu) s'il existe (resp. s'il n'existe pas) dans WordNet. Ainsi nous notons que, les intérêts précis fournis par notre approche sont des mots-clés compréhensibles qui reflètent vraiment le contenu de la ressource comme "technology", "foursquare", "history", etc. Ceci est un avantage, car les tags sont des mots clés générés par les utilisateurs. Notre approche a filtré les tags ambigus (par exemple "gis") qui ne sont pas compréhensibles par d'autres utilisateurs. L'ambiguïté des tags a diminué (pour cet ensemble d'utilisateurs) de 35 % à 10 % selon WordNet. Ainsi, l'écart d'ambiguïté des tags entre les données d'origine (avant traitement) et les résultats (après traitement) est égal à 71,25 %.

4.4. *Évaluation selon l'approche classique basée tags (tag-based)*

En utilisant le même ensemble d'utilisateurs que dans la section précédente, nous avons comparé notre approche avec l'approche classique basée sur les tags. Cette dernière considère les tags des utilisateurs comme représentant de ses intérêts (Astrain *et al.*, 2010) (Li *et al.*, 2008).

Nous comparons le résultat fourni par notre approche avec le résultat de l'approche qui utilise tous les tags du réseau égocentrique (sans tenir compte de leur pertinence par rapport aux ressources associées). Nous comparons selon $k = 100$ de notre approche (fournit les meilleurs résultats). De plus, nous comparons en prenant en considération que les synonymes ou mots reliés (puisque elle est meilleure que la simple comparaison). Nous calculons la précision moyenne de tous les utilisateurs dans la base de données et nous la comparons avec la précision moyenne fournie par notre approche. Nous avons obtenu une précision moyenne égale à **0.6038** pour notre approche et une précision moyenne égale à **0.3459** pour l'approche classique basée tags. Ceci, montre que notre approche permet de surmonter l'approche classique basée sur les tags en terme de précision. Cela est dû à l'examen du contenu des ressources analysées pour la sélection des tags pertinents. Le processus de sélection filtre implicitement les tags ambigus qui peuvent ne pas être compréhensibles pour les autres utilisateurs.

De même que dans la section précédente, nous présentons les boîtes de Tukey de ces résultats selon les valeurs de précision dans la figure 8.

Cette répartition des valeurs de précision s'explique par :

- Pour la précision de notre approche selon les synonymes ou mots reliés, la distribution est presque au milieu. Ceci reflète que la plupart des utilisateurs ont la même précision moyenne.
- Pour la précision selon l'approche classique basée les tags, la distribution est en dessous de la distribution des synonymes ou des mots reliés. Ceci reflète que la plupart des utilisateurs ont des valeurs de précisions assez basses.

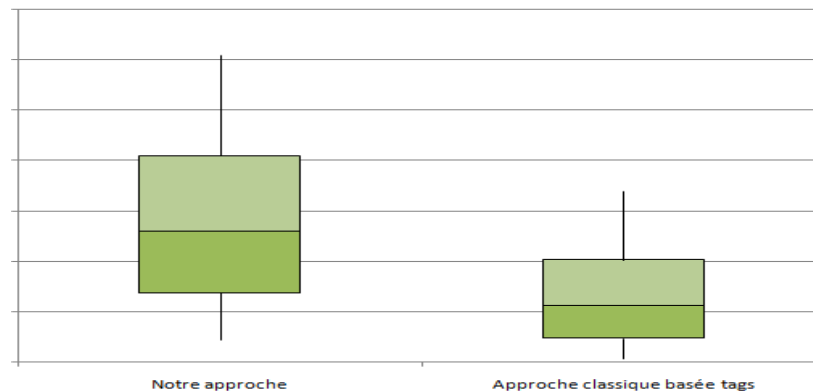


Figure 8. Répartition des précisions selon des boîtes de Tukey de notre approche et de l'approche basée tags en fonction du réseau égocentrique

Afin de mieux comprendre ces résultats, nous avons choisi pour représenter un échantillon de 20 utilisateurs choisis au hasard (comme nous ne pouvions pas montrer la précision des 1876 utilisateurs dans une même figure). En fait, les résultats de notre approche ont fourni des échantillons similaires. La figure 9 compare notre approche proposée avec la précision fournie par l'approche basée sur les tags.

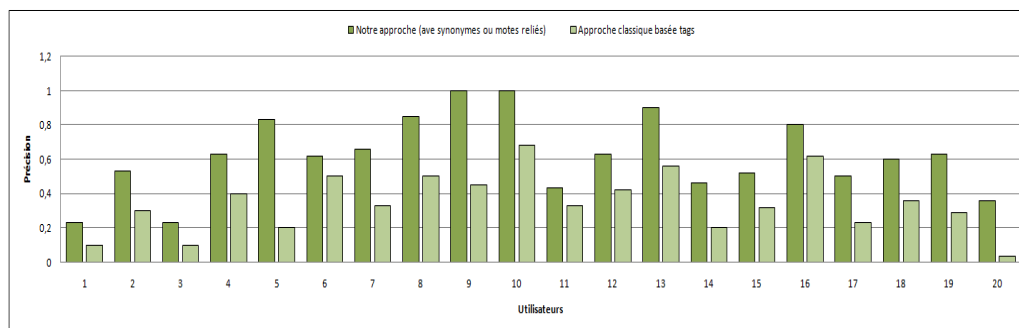


Figure 9. Comparaison de la précision de notre approche proposée avec la précision fournie par l'approche classique basée sur les tags

Nous remarquons que notre approche est généralement plus performante que l'approche basée sur les tags (cf. figure 9). En fait, la précision est 88,10 % plus élevée pour notre approche avec les synonymes ou mots reliés que l'approche basée sur les tags (pour tous les utilisateurs). De plus, nous constatons que les meilleurs résultats sont liés à des utilisateurs actifs.

5. Conclusion

Dans ce papier, nous avons proposé une approche pour détecter les intérêts des utilisateurs précis en se basant sur l'environnement social. Le but est de déduire les intérêts des utilisateurs à partir des tags. Pour cela, nous utilisons le contenu des ressources annotées afin de filtrer les tags reflétant vraiment la thématique des ressources

et censés représenter les intérêts de l'utilisateur.

L'originalité de notre approche est basée sur la proposition d'une nouvelle technique de détection des intérêts en cherchant à prendre en compte la précision du comportement d'annotation du réseau égocentrique d'un utilisateur. Cela se fait par l'application d'une technique d'indexation suivie d'une fonction qui score les tags attribués aux ressources. Ce score reflète la pertinence de la ressource par rapport au tag. Ensuite, nous sélectionnons les ressources les plus pertinentes (top-k) pour un tag donné. Si le tag est attribué par l'utilisateur à une ressource qui est dans ce top-k, alors le tag est considéré comme un intérêt précis, c'est à dire il décrit le contenu des ressources auxquelles il a été attribué.

Les résultats ont montré que notre approche est plus performante lorsqu'elle est complétée par la prise en compte des synonymes ou mots reliés.

De plus, les résultats ont prouvé que la prise en compte des ressources annotées pour détecter les intérêts des utilisateurs concernés (notre approche) est meilleure que de considérer directement les tags attribués par les utilisateurs (approche classique basée sur les tags). En fait, notre approche traite implicitement l'ambiguïté des tags et donc elle fournit de meilleurs résultats.

Les expérimentations ont montré que notre approche fournit un ensemble compréhensible d'intérêts. Par conséquent, notre approche pourrait être utilisée à des fins d'adaptation (par exemple la recommandation), car elle offre une solution pour détecter les intérêts des utilisateurs concernés.

Les limites dégagées qui feront l'objet de travaux futurs sont les suivantes.

En ce qui concerne le choix de la valeur de k (qui sélectionne les top-k ressources), nous envisageons un calcul expérimental par apprentissage afin d'automatiser le choix de cette valeur. En ce qui concerne la fonction de score (pour le calcul des top-k ressources pour un tag), la limitation principale de ce modèle est qu'il ne prend pas en compte les relations entre les mots (par exemple, les synonymes). Nous pouvons prévoir donc une prise en compte de la sémantique lors du calcul de score afin de voir l'influence de cette caractéristique sur les valeurs de précisions. En ce qui concerne les utilisateurs, nous avons montré que notre approche est moins efficace pour les utilisateurs non actifs. Il s'agira de trouver des solutions plus efficaces que des techniques classiques comme leur attribuer les tags les plus populaires et/ou les plus récentes comme étant des intérêts.

Nous comptons également tester notre approche sur d'autres bases sociales, afin de voir son efficacité dans d'autres contextes.

Remerciements

Ce travail a été soutenu financièrement par le programme "PHC Utique" du ministère français des affaires étrangères et le ministère de l'enseignement supérieur et la recherche et le ministère tunisien de l'enseignement supérieur et la recherche scientifique sous le numéro de projet CMCU 30540XK.

Bibliographie

- Astrain J. J., Cordoba A., Echarte F., Villadangos J. (2010). An algorithm for the improvement of tag-based social interest discovery. In *Semapro '10: Proceedings of the fourth international conference on advances in semantic processing*, p. 49–54. Consulté sur http://www.thinkmind.org/index.php?view=article&articleid=semapro_2010_3_10_50021
- Ivan C., Peter B., Tsvi K. (2011). *HetRec '11: Proceedings of the 2Nd international workshop on information heterogeneity and fusion in recommender systems*. New York, NY, USA, ACM.
- Kim H.-N., Alkhalidi A., El Saddik A., Jo G.-S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. *Expert Systems with Applications*, vol. 38, n° 7, p. 8488–8496. Consulté sur <http://www.sciencedirect.com/science/article/pii/S0957417411000686>
- Li X., Guo L., Zhao Y. E. (2008). Tag-based social interest discovery. In *Proceedings of the 17th international conference on world wide web*, p. 675–684. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/1367497.1367589>
- Ma Y., Zeng Y., Ren X., Zhong N. (2011). User interests modeling based on multi-source personal information fusion and semantic reasoning. In *Proceedings of the 7th international conference on active media technology*, p. 195–205. Berlin, Heidelberg, Springer-Verlag. Consulté sur <http://dl.acm.org/citation.cfm?id=2033896.2033923>
- Meo P. D., Quattrone G., Ursino D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, vol. 20, n° 1, p. 41–86. Consulté sur <http://link.springer.com/article/10.1007/s11257-010-9072-6>
- Mezghani M., Zayani C. A., Amous I., Gargouri F. (2012). A user profile modelling using social annotations: A survey. In *Proceedings of the 21st international conference companion on world wide web*, p. 969–976. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2187980.2188230>
- Milicevic A. K., Nanopoulos A., Ivanovic M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, vol. 33, n° 3, p. 187–209. Consulté sur <http://link.springer.com/article/10.1007/s10462-009-9153-2>
- On-at S., Canut C. M., Péninou A., Sèdes F. (2014). Deriving user's profile from sparse egocentric networks: Using snowball sampling and link prediction. In *Ninth international conference on digital information management, ICDIM 2014, phitsanulok, thailand, september 29 - oct. 1, 2014*, p. 80–85. Consulté sur <http://dx.doi.org/10.1109/ICDIM.2014.6991421>
- Tchuenta D., Canut M.-F., Jessel N., Peninou A., Sèdes F. (2013). A community-based algorithm for deriving users' profiles from egocentric networks: experiment on facebook and DBLP. *Social Network Analysis and Mining*, vol. 3, n° 3, p. 667–683. Consulté sur <http://link.springer.com/article/10.1007/s13278-013-0113-0>
- Zhou T. C., Ma H., Lyu M. R., King I. (2010). Userrec: A user recommendation framework in social tagging systems. In M. Fox, D. Poole (Eds.), *Aaai*. AAAI Press.

Session Modèles et méthodes (1)

Concept de Famille de Méthodes : qu'en pensent les utilisateurs ?

Rébecca Deneckère¹, Elena Kornyshova²

1. Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
90, Rue de Tolbiac, 75013 Paris, France
rebecca.deneckere@univ-paris1.fr

2. CEDRIC, Conservatoire National des Arts et Métiers
2, rue Conté, 75003, Paris, France
elena.kornyshova@cnam.fr

RESUME. L'ingénierie des méthodes situationnelles a pour objectif de fournir des approches de construction de méthodes en fonction des projets sur lesquels elles seront appliquées. L'un des enjeux majeurs de ce domaine est de réussir à convaincre les professionnels de l'utilité de telles approches. En effet, les entreprises sont en général convaincues que ces approches sont trop complexes et trop coûteuses en temps et en ressources. Le concept de famille de méthode a été introduit pour répondre à cet enjeu et propose une approche de construction de méthodes configurables et adaptables en fonction du contexte du projet. Ce concept a déjà été appliqué à différents domaines et nous proposons dans ce papier une évaluation de cette approche auprès des utilisateurs pour identifier si elle permet réellement de répondre à cet enjeu.

ABSTRACT. Situational Methods Engineering aims to provide different approaches for constructing methods in function of projects on which they will be applied. One of the major issues in this area is to convince the professionals of the utility of such approaches. Indeed, companies are generally convinced that these approaches are too complex and too time- and resource-consuming. The method family concept has been introduced to meet this challenge and offers an approach of constructing configurable and adaptable methods based on the project context. This concept has already been applied to various fields and we propose in this paper an evaluation of this approach by users to identify if it really can meet this challenge.

MOTS-CLES : Ingénierie des méthodes situationnelles, Famille de méthodes, Evaluation, Questionnaire

KEYWORDS: Situational method engineering, Method family, Evaluation, Questionnaire

1. Introduction

Les méthodes de développement de Systèmes d'information (SI) se sont de plus en plus complexifiées, ceci dans le but d'obtenir une méthode adaptée à toutes les situations de projets rencontrées par les professionnels. Cependant, il est maintenant admis que le concept de méthode universelle n'est pas un concept viable (Henderson-Sellers, 2014) et plusieurs travaux se sont penchés sur la manière d'adapter les méthodes à la situation au moment de leur construction, ce qui permet d'obtenir directement la meilleure méthode pour le projet en cours. Ces méthodes adaptées à la situation sont appelées des *méthodes situationnelles* et reposent sur le principe que les méthodes sont décomposables en blocs indépendants appelés *composants de méthodes* et qui peuvent être réassemblés pour former une nouvelle méthode. Ces composants sont en général sélectionnés selon les besoins en fonction de l'adéquation de leur contexte et de celui du projet (Kornysheva et al 2011b). Cependant, ces approches de construction sont souvent assez complexes. L'ingénierie des méthodes est considérée comme une démarche comme trop coûteuse, consommatrice du temps et exige un niveau considérable d'expertise. Le concept de *Famille de méthodes* a été proposé pour répondre à la demande d'une approche de construction plus simple et facile à appliquer (Kornysheva et al, 2011a).

Déjà utilisée dans plusieurs domaines (ingénierie des besoins, analyse de données, prise de décision, etc.), l'approche des Familles de méthodes n'avait pas encore reçu de véritable évaluation par les utilisateurs. Nous proposons dans ce papier une évaluation de ce concept auprès d'un ensemble d'utilisateurs (professionnels et académiques) pour tenter de valider la compréhension et l'utilité de la proposition. Cette évaluation a été effectuée auprès de populations diverses évoluant autour des systèmes d'information et des méthodes. Des ateliers explicatifs ont été organisés autour du concept puis les participants ont répondu à un questionnaire qui nous a permis de comprendre le ressenti des utilisateurs par rapport à cette nouvelle approche de construction de méthodes. Certaines corrélations ont également été faites selon l'habitude des participants à utiliser plusieurs méthodes, selon le projet ou dans un même projet. Les résultats, encourageants, sont présentés dans cet article.

Après une explication du cadre théorique (section 2), nous présentons notre méthode de recherche (section 3) et les résultats de l'évaluation (section 4). Nous décrivons les travaux connexes (section 5) et terminerons notre article en proposant quelques perspectives sur les travaux futurs (section 6).

2. Cadre théorique

2.1. Ingénierie des méthodes

La maîtrise des développements des SI passe par l'emploi des méthodes d'ingénierie. Les méthodes permettent de transcrire les besoins d'une organisation en un SI approprié, grâce à un ensemble d'idées, d'approches, de techniques et d'outils. Cependant, cette complexité croissante des SI engendre également de nouveaux besoins sans cesse renouvelés en terme de méthode. Il n'existe malheureusement pas de

méthode universelle pouvant d'adapter à toutes les situations de projet (Kornysheva et al, 2011b). L'ingénierie des méthodes situationnelles (IMS) tente de répondre à cette problématique en offrant diverses techniques permettant de créer ou d'adapter des méthodes selon la situation du projet en cours (Henderson-Sellers et al, 2014). En effet, il est maintenant admis que la situation d'ingénierie de chaque projet est différente et engendre donc un besoin de support méthodologique différent (Kumar, 1992). Pour permettre une adaptation efficace, l'IMS propose de réutiliser les méthodes en les considérant comme des ensembles de composants, ceux-ci étant réutilisables et peuvent être combinés les uns avec les autres. Ces composants sont stockés dans une base de méthodes, ce qui permet de les retrouver plus facilement. Plusieurs approches ont été proposées sur cette notion de composant modulaire : Approches d'assemblage de composants (Brinkkemper et al, 1998) (Firesmith and Henderson Sellers, 2002) (Ralyté and Rolland, 2001) ; Approches de configuration (Karlsson and Agerfalk, 2004) ; Approches d'adaptation (Rossi et al, 2004) ; Approches dirigées par les modèles (Cervera et al, 2012) ; Approches orientées services (Guzelian et al, 2007) (Iacovelli 2012).

Cependant, l'usage de ces approches d'IMS n'est pas très répandu dans les entreprises car leur mise en œuvre est difficile. L'un des enjeux de l'ingénierie des méthodes est de vaincre cette réticence des entreprises à l'utilisation de méthodes (collectif INFORSID, 2012). (Deneckere et al, 2014) s'interroge sur les causes de ce désamour : approches peu interopérables, techniques de sélection des composants complexes et pouvant nécessiter des recherches dans différentes bases, pas de processus organisationnel pour la gestion des composants, prise en compte de la caractérisation des projets et des composants mal formalisée. Cependant, la difficulté principale concernant l'intégration des méthodes situationnelles en entreprise est l'idée que ces approches consomment trop de temps et sont trop chers en termes de connaissances et de ressources. Les professionnels exigent des outils simples et agiles, faciles à mettre en œuvre et ne requérant pas de vastes connaissances en ingénierie des méthodes (Deneckere et al, 2014).

2.2. Famille de méthodes

Nous avons proposé dans (Deneckere et al, 2011), (Kornysheva, 2013) (Deneckere et al, 2014) le concept de Famille de méthodes. Comme les autres approches d'IMS, les familles considèrent les méthodes comme des ensembles de composants qui peuvent être assemblés de différentes manières pour créer de nouvelles méthodes adaptées à la situation en cours. La grande originalité de cette approche réside dans la manière dont sont organisés les composants et qui facilite leur sélection et leur réutilisation. Chaque famille est définie pour un domaine spécifique et cette approche a été utilisée avec succès dans plusieurs domaines tels que la prise de décision (Kornysheva, 2011), la définition des besoins (Deneckere et al, 2011), l'analyse de données dans les sciences humaines et sociales (Ammar et al, 2014), la co-création de services (Ralyté, 2013), les projets agiles (Deneckere, 2015), etc.

Cette approche est inspirée du concept de lignes de produit logiciels (Clements, 2001) (Weiss, 1999). L'ingénierie des lignes de produits repose sur le développement des produits logiciels basé sur de la personnalisation massive. Le concept repose sur la spécification des similitudes et de différences entre les variantes d'une

ligne de produits (Pohl 2005). Ce concept de variabilité a été réutilisé dans le concept de Famille de méthodes : certains composants sont communs et d'autres sont différents ; des dépendances de variabilité pouvant également être définies entre les différents composants (Kornyshova *et al*, 2013). L'organisation des composants dans la famille permet de spécifier la variabilité des différents composants des méthodes similaires. Il est en conséquence beaucoup plus facile de sélectionner les composants selon la situation du projet en cours. Cette organisation des composants offre une très grande flexibilité aux utilisateurs en leur offrant un processus composé d'alternatives adaptées à leur situation.

Les familles de méthodes sont graphiquement représentées avec le formalisme MAP (Rolland *et al.*, 1999). La particularité de ce formalisme est qu'il permet de spécifier des processus de manière flexible en se focalisant sur la partie intentionnelle du processus. Ce formalisme a déjà été utilisé avec succès dans le domaine de l'IMS (Ralyté and Rolland, 2001) (Deneckère, 2001) (Ralyté *et al*, 2003). La représentation formelle des familles permet également d'intégrer les contraintes de variabilité et de dépendances inhérentes à chaque famille (Deneckere *et al*, 2014). Les approches de construction de méthodes situationnelles se doivent de prendre en compte plusieurs éléments : offrir un processus de construction, des techniques de prise en compte du contexte, la configuration de la méthode selon le contexte. Le concept de famille de méthodes propose un processus de construction spécifique basé sur des composants de méthodes modulaires (Deneckere *et al*, 2014). La prise en compte du contexte se fait à la fois au niveau des composants et à celui du projet (Kornyshova *et al*, 2011b). Le processus de configuration des familles utilise la contextualisation des composants et du projet pour aider l'utilisateur à sélectionner les composants pertinents de manière plus efficace pour définir une ligne de méthode adaptée (Deneckere *et al*, 2011). Le concept de famille de méthodes a été utilisé avec succès dans plusieurs domaines. Une explication succincte de ces différentes familles est donnée dans la partie 5.

3. Méthode de recherche

Les familles de méthodes ont déjà été appliquées à plusieurs domaines. Ces familles ont en général été validées qualitativement sur des cas d'études théoriques ou avec des utilisateurs ciblés précisément par le domaine concerné. L'une d'elle a également bénéficiée d'une évaluation quantitative (Deneckere *et al*, 2015) mais le concept de famille en tant que tel n'a pas encore été validé auprès des utilisateurs.

3.1. Conception du questionnaire

Pour effectuer notre évaluation, nous avons choisi de réaliser une enquête. Une enquête a pour objectif de produire de la connaissance (De Singly, 2012). Elle peut être réalisée par des questionnaires, des entretiens, etc. L'intérêt de l'enquête par questionnaire réside aussi dans le fait que l'on peut interroger une population plus grande. Nous avons donc choisi de créer un questionnaire (sur le web) et de demander à plusieurs populations concernées par les méthodes de développement de le remplir : des professionnels (chefs de projets, développeurs, etc.) ainsi que des cher-

cheurs, des doctorants et des étudiants en ingénierie des SI. Cette enquête est descriptive. Son objectif est de comprendre comment les répondants envisagent le concept de famille alors qu'ils le rencontrent pour la première fois. L'enquête a été évaluée par des spécialistes externes pendant l'été 2014 et a été conduite auprès des répondants entre Septembre 2014 et Janvier 2015.

Le questionnaire satisfait les besoins de la directive 'Data Protection Directive' de l'Union Européenne (Directive 95/46/EC du parlement européen et du conseil du 24 octobre 1995 sur la protection des individus en regard de l'utilisation des données personnelles et sur la liberté de mouvement de ces données.). Ce document établit que les organisateurs de l'enquête doivent respecter la vie privée des répondants ainsi que les droits sur la protection des données qui leur sont confiées. Pour cela, nous avons respecté les points suivants :

- Nous avons conçu un protocole spécifique d'évaluation (article 6).
- Nous avons obtenu l'accord des répondants sur l'utilisation de leurs réponses (article 7).
- Nous avons distribué un document aux répondants pour expliquer l'objectif principal de l'enquête et les données qui allaient être utilisées dans l'analyse (article 10 et 8). Bien entendu, toutes les réponses étaient anonymes.
- Nous avons assuré la sécurité des données (article 17).
- Nous avons donné le droit aux répondants d'accéder à leurs propres informations (articles 10 et 12).
- Nous nous sommes assurées que tous les répondants aient un droit de regard sur la collecte et l'utilisation de leurs données (article 14).

3.2. *Questions de recherche*

L'évaluation doit permettre d'avoir un aperçu de ce que pensent les utilisateurs du concept de famille de méthodes. Nous avons notamment décidé d'interroger les participants en terme de complexité du concept mais également en terme d'utilité. En effet, les principales critiques opposées aux différentes approches de construction de méthodes situationnelles sont essentiellement portées sur ces deux points. Ces deux objectifs d'évaluation sont évalués de manière quantitative par le biais de plusieurs questions. Nous avons également choisi de poser quelques questions ouvertes pour obtenir une évaluation qualitative des répondants sur le concept.

Pour répondre à cette problématique, nous avons défini les questions de recherche suivantes (QR) :

QR 1 : Le concept de famille de méthodes semble-t-il complexe du point de vue des utilisateurs ?

Pour cette question de recherche nous avons posé trois questions (Q) :

- Q1.1 : Pensez-vous que le concept de famille est facile à comprendre ?
- Q1.2 : Pensez-vous que l'utilisation de ce concept de famille de méthodes nécessite une phase d'apprentissage ?
- Q1.3 : Seriez-vous prêt à expliquer le concept de famille de méthodes à quelqu'un d'autre ?

QR2 : Le concept de famille de méthodes paraît-il utile du point de vue des utilisateurs ?

Les deux questions permettent de répondre à cette question de recherche :

- Q2.1 : Seriez-vous prêt à utiliser le concept de famille de méthodes dans votre travail ?
- Q2.2 : Recommanderiez-vous le concept de famille à d'autres personnes ?

QR3 : Quels sont les avantages et les inconvénients à utiliser une famille de méthodes ?

Pour cette question, nous avons demandé aux participants :

- Q3.1 : Quels pourraient être les avantages de l'utilisation d'une famille ?
- Q3.2 : Quels pourraient être les désavantages de l'utilisation d'une famille ?

Pour approfondir l'analyse, nous avons ajouté deux questions permettant de comprendre les habitudes des répondants en termes d'utilisation de plusieurs méthodes de développement sur un ou plusieurs projets. L'aspect multi-méthodes étant important pour le concept de famille de méthodes, l'objectif est de savoir si l'expérience d'utilisation de plusieurs méthodes différentes impacte l'avis des participants sur le concept de famille de méthodes. Ces deux questions sont :

- Q0.1 : Appliquez-vous des méthodes différentes selon le projet ?
- Q0.2 : Avez-vous déjà appliqué plusieurs méthodes (ou plusieurs parties de méthodes différentes) dans un même projet ?

Les résultats du questionnaire sont restitués pour toute la population et en fonction des réponses à ces deux questions.

3.3. Protocole

Le processus de l'enquête est illustré à la figure 1. Nous avons conduit l'enquête en combinant deux techniques : l'organisation d'ateliers pour les groupes et un questionnaire en ligne. Un document a été remis à tous les répondants pour expliquer l'objectif de l'évaluation et les principaux concepts de l'approche. Après avoir pris connaissance de ce document, les participants ont répondu au questionnaire. La première partie du questionnaire concernait les répondants eux-mêmes. La deuxième partie mettait en exergue la compréhension du concept de famille par les répondants. Après avoir reçu les réponses, nous avons nettoyé les données. Les réponses comportant des données incomplètes ont été supprimées. L'étape finale de l'enquête a été d'analyser les résultats pour répondre aux questions de recherche.



Figure 1. Illustration du protocole de l'enquête.

3.4. Validité

Nous avons tenté de toucher les différents types de population avec ce questionnaire, ce qui assure une certaine diversité dans les réponses. Tous les répondants évoluent dans le domaine des systèmes d'information, soit en tant qu'étudiants de Master, soit en tant que professionnels (consultants, chefs de projets, développeurs) ou encore chercheurs (doctorants ou enseignants-chercheurs). Tous les répondants avaient donc une certaine connaissance et une certaine sensibilité en ingénierie des méthodes, ce qui assurait également une certaine justesse, les répondants étant tous concernés par la problématique soulevée. Le questionnaire a été évalué et commenté par deux experts du monde professionnel, avant la mise en œuvre auprès des répondants, ce qui nous a permis de faire quelques améliorations et de mieux cibler nos questions. Il aurait pourtant été souhaitable d'avoir un plus grand nombre de réponses en réussissant à toucher un plus grand nombre de répondants, ce qui aurait assuré une meilleure validité quantitative des réponses obtenues.

4. Résultats du questionnaire

Nous avons obtenu 48 réponses au total. Suite à une vérification des données, deux réponses ont été supprimées car elles comportaient des données incomplètes. La structure de la population enquêtée (Tableau 1 et 2) contient 28 hommes (61%) et 18 femmes (39%). La majorité des personnes participant au questionnaire sont des étudiants de Master en Informatique (50%) et des salariés du secteur informatique (35%). Le reste de la population est constitué d'enseignants-chercheurs et de doctorants, les deux groupes appartenant également au domaine informatique.

Tableau 1. Répartition des sexes

Sexe	Nombre	%
Homme	28	61%
Femme	18	39%

Tableau 2. Répartition des métiers

Métier	Nombre	%
Etudiant de Master	23	50%
Salarié	16	35%
Enseignant-Chercheur	3	7%
Doctorant	4	9%

Les tableaux 3 et 4 reprennent les informations sur les habitudes des participants concernant l'utilisation d'une ou plusieurs méthodes. Nous souhaitons savoir si le fait d'avoir déjà appliqué des méthodes différentes - soit dans un même projet, soit dans des projets différents - peut avoir un impact sur le ressenti des utilisateurs par rapport à la famille de méthodes dont la vocation est d'unifier plusieurs méthodes en une seule.

Tableau 3. Q0.1 : Appliquez-vous des méthodes différentes selon le projet ?

Réponse	Nombre	%
Oui	23	50%
Non	23	50%

Tableau 4. Q0.2 : Avez-vous déjà appliqué plusieurs méthodes (ou plusieurs parties de méthodes différentes) dans un même projet ?

Réponse	Nombre	%
Oui	18	39%
Non	28	61%

Nous présentons ci-dessous les résultats de l'évaluation pour les trois sous-questions de recherche. Pour plus de clarté, les résultats mentionnés dans le texte sont grisés dans les tableaux correspondants.

4.1. QR 1 : Le concept de famille de méthodes semble-t-il complexe du point de vue des utilisateurs ?

Le tableau 5 présente les réponses des participants à la première question concernant la facilité de compréhension du concept. Ces résultats sont analysés par rapport aux réponses des questions Q0.1. et Q0.2. (tableaux 3 et 4) présentés précédemment. La plupart des répondants - 67% (61%+6%) - sont d'accord sur le fait que le concept de famille de méthodes est facile à comprendre. Seuls 11% des participants n'adhèrent pas à cette idée. On peut cependant remarquer que les utilisateurs ayant l'habitude de manipuler plusieurs méthodes sont beaucoup plus conscients de la difficulté d'exécution que peut représenter une telle approche : 61% (57%+4%) pour une méthode différente selon le projet et 55% (44%+11%) pour plusieurs méthodes dans un même projet au lieu des 67% sur la population entière. Cela montre que même si ces participants, plus touchés par le cœur de l'approche, sont légèrement plus réticents que les autres envers la proposition, ils sont quand même majoritairement en faveur de celle-ci.

Tableau 5. Q1.1 : Pensez-vous que le concept de famille est facile à comprendre ?

Réponse	Nbr	%	Q0.1.				Q0.2.			
			OUI		NON		OUI		NON	
			Nbr	%	Nbr	%	Nbr	%	Nbr	%
Tout à fait d'accord	3	6%	1	4%	2	9%	2	11%	1	4%
Plutôt d'accord	28	61%	13	57%	15	65%	8	44%	20	71%
Sans opinion	10	22%	6	26%	4	17%	5	28%	5	18%
Plutôt pas d'accord	5	11%	3	13%	2	9%	3	17%	2	7%
Pas du tout d'accord	0	0%	0	0%	0	0%	0	0%	0	0%

Le tableau 6 présente les réponses concernant la phase d'apprentissage nécessaire à l'utilisation du concept. La majorité des participants (76%) estiment que le concept de famille ne nécessite qu'une courte phase d'apprentissage. 15% pensent quand à eux que la phase d'apprentissage est longue. Pour quelques autres répondants, la durée d'apprentissage devrait dépendre de la personne concernée. La corrélation entre l'utilisation de méthodes différentes et la perception de la phase d'apprentissage n'est pas flagrante. Les participants ayant utilisé différentes méthodes sur des projets différents pensent plus que la phase d'apprentissage est courte que ceux qui ne sont pas dans ce cas (83% contre 70%). Pour la deuxième question, la tendance est inversée puisque ceux qui n'utilisent qu'une seule méthode par projet sont majoritairement pour une estimation de phase d'apprentissage courte (82% contre 67%). Les participants qui ont utilisé plusieurs méthodes pour un seul projet sont également plus nombreux à considérer qu'il faut avoir une longue phase d'apprentissage par rapport à la population entière (28% contre 15%).

Tableau 6. Q1.2 : Pensez-vous que l'utilisation de ce concept de famille de méthodes nécessite une phase d'apprentissage ?

Réponse	Nbr	%	Q0.1.				Q0.2.			
			OUI		NON		OUI		NON	
			Nbr	%	Nbr	%	Nbr	%	Nbr	%
Une longue phase d'apprentissage	7	15%	3	13%	5	22%	5	28%	3	10%
Une courte phase d'apprentissage	35	76%	19	83%	16	70%	12	67%	23	82%
Pas d'apprentissage	1	2%	0	0%	1	4%	0	0%	1	4%
Autre	3	7%	1	4%	1	4%	1	5%	1	4%

Le tableau 7 présente les résultats concernant l'estimation des utilisateurs à avoir suffisamment compris le concept pour pouvoir l'expliquer à quelqu'un d'autre. 61% des répondants pourraient réexpliquer le concept. Concernant les réponses aux questions Q0.1. et Q0.2, la seule inférence marquante serait que les répondants ayant déjà appliqué des différentes méthodes au sein d'un même projet auraient mieux compris le concept et seraient plus enclins à expliquer le concept à d'autres (72%).

Tableau 7. Q1.3 : Seriez-vous prêt à expliquer le concept de famille de méthodes à quelqu'un d'autre ?

Réponse	Nbr	%	Q0.1.				Q0.2.			
			OUI		NON		OUI		NON	
			Nbr	%	Nbr	%	Nbr	%	Nbr	%
Oui	28	61%	13	56%	15	65%	13	72%	15	53%
Non	15	33%	8	35%	7	31%	5	28%	10	36%
Autre	3	6%	2	9%	1	4%	0	0%	3	11%

4.2. QR2 : Le concept de famille de méthodes paraît-il utile du point de vue des utilisateurs ?

Tableau 8. Q2.1 : Seriez-vous prêt à utiliser le concept de famille de méthodes dans votre travail ?

Réponse	Nbr	%	Q0.1.				Q0.2.			
			OUI		NON		OUI		NON	
			Nbr	%	Nbr	%	Nbr	%	Nbr	%
Oui, cela serait utile	21	46%	12	53%	10	44%	5	27%	16	57%
Non, cela complexifierait trop mon travail	6	13%	4	17%	2	9%	3	17%	3	11%
Non, je n'en n'ai pas l'utilité	11	24%	4	17%	7	30%	3	17%	8	29%
Autre	8	17%	3	13%	4	17%	7	39%	1	3%

Le tableau 8 comporte les résultats concernant l'utilité que pourrait représenter le concept de famille dans le travail des répondants. 46% des participants pensent que le concept leur serait utile alors que 37% ne sont pas prêt à l'utiliser (24% + 13%). La relation avec les questions Q0.1 et Q0.2 montre que les personnes utilisant des mé-

thodes différentes selon le projet seraient un peu plus enclines à utiliser le concept de famille (53% contre 44%). Le rapport avec les personnes utilisant plusieurs méthodes dans un même projet est beaucoup plus faible (27% contre 57%) mais la différence se compense avec le taux de réponses de type « autre » (39%). En effet, les réponses ouvertes ont été beaucoup utilisées pour cette question car ce sont ces personnes qui ont déjà rencontré les problèmes de composition de méthodes ou de parties de méthodes. Elles sont plus critiques face à l'utilité des familles de méthodes.

Le tableau 9 introduit les réponses concernant l'appréciation positive du concept de famille en demandant aux répondants s'ils pourraient le recommander. 66% des répondants seraient d'accord pour recommander le concept à d'autres personnes. Il est à noter qu'une partie de ces répondants (9 personnes) n'utiliseraient pas cette méthode pour leurs projets en cours. Ici également les réponses ouvertes étaient plutôt positives, en proposant des adaptations selon les situations et des essais préalables en situation concrète. La corrélation la plus frappante montre qu'une grande majorité (83%) des personnes utilisant des méthodes différentes selon le projet serait prête à effectuer cette recommandation.

Tableau 9. Q2.2 : Recommanderiez-vous le concept de famille à d'autres personnes ?

Réponse	Nbr	%	Q0.1.				Q0.2.			
			OUI		NON		OUI		NON	
			Nbr	%	Nbr	%	Nbr	%	Nbr	%
Oui	30	66%	19	83%	11	48%	12	66%	18	64%
Non	8	17%	3	13%	5	22%	3	17%	5	18%
Autre	8	17%	1	4%	7	30%	3	17%	5	18%

4.3. QR3 : Quels sont les avantages et les inconvénients à utiliser une famille de méthodes ?

Pour cette question de recherche, nous n'avons pas voulu influencer les utilisateurs en posant des questions avec des réponses pré-formatées trop précises. Les questions étaient donc complètement ouvertes.

Q3.1 : Quels pourraient être les avantages de l'utilisation d'une famille ?

Nous avons obtenu 25 réponses sur les avantages potentiels de l'utilisation des familles de méthodes. Le tableau 10 résume les réponses données par les participants. L'avantage cité le plus souvent est la mise en œuvre d'une meilleure organisation du travail (17%) ce qui est compréhensible puisque l'approche offre une organisation des composants permettant de décrire le processus à utiliser. 15% ont trouvé que l'approche est flexible grâce à la configuration des méthodes (possibilité de choisir les techniques désirées en fonction des besoins, avant ou pendant le déroulement du projet). 15% citent la complétude du processus offert, considéré comme un recueil de bonnes pratiques sur un domaine et permettant la découverte potentielle de techniques alternatives à des techniques connues ou déjà pratiquées. 13% à cette question ont également appréciée la bonne vision du processus qui est offert par cette approche, que ce soit au niveau de la description, standardisation ou généralisation. Le guidage offert par l'approche n'a été que marginalement apprécié par 7% des répondants.

Tableau 10 : Avantages des familles de méthodes.

Types d'avantages cités par les participants	Nbr	%
Meilleure organisation du travail	8	17%
Flexibilité et configuration	7	15%
Complétude du processus	7	15%
Bonne vision du processus	6	13%
Guidage dans la prise de décision	3	7%
Simplicité d'utilisation	1	2%

Q3.2 : Quels pourraient être les désavantages de l'utilisation d'une famille ?

20 participants ont mentionné les inconvénients de l'approche de familles (Tableau 11). 15% des répondants ont cité comme principal inconvénient la complexité de l'approche : beaucoup de composants différents dans le même processus, ce qui peut engendrer une difficulté de compréhension, et une coordination pouvant être rendue difficile entre les différents acteurs. 11% ont estimé que l'application de l'approche engendrerait une perte de temps, tant au niveau de la phase d'apprentissage que lors de la mise en place de la famille. 7% des répondants ont pensés que le guidage pourrait être amélioré pour faciliter le choix entre les différentes alternatives proposées dans une famille. De manière marginale, 4% des répondants ont estimés que cela augmenterait leur charge de travail (cette réponse étant fréquemment cité avec le coût en temps) et pour 4%, l'approche n'était pas adaptée à tous les cas de figures. Un répondant (2%) a estimé que proposer un processus générique contenant toutes les techniques pouvait engendrer une perte de créativité et de curiosité envers les techniques existant en dehors de la famille.

Tableau 11 : Inconvénients des familles de méthodes.

Types d'inconvénients cités par les participants	Nbr	%
Complexité	7	15%
Coût en temps	5	11%
Guidage difficile	3	7%
Augmentation de la charge de travail	2	4%
Pas adapté à tous les cas	2	4%
Perte de créativité	1	2%

5. Travaux connexes sur les familles de méthodes

Cette section a pour but de montrer dans quelles situations l'approche des familles de méthodes a été utilisée et comprend une brève description des familles de méthodes existantes ainsi qu'une étude sur l'évaluation d'une famille de méthodes construite pour la phase de lancement des projets agiles.

5.1. Familles de méthodes existantes

Plusieurs familles de méthodes ont déjà été développées avec cette approche, chacune dans un domaine très précis. Cette section donne un aperçu des différentes familles déjà existantes.

Définition de scénarios en ingénierie des besoins (Deneckere et al, 2011). Cette famille prend en compte des composants venant des deux méthodes ; L'Écritoire (Rolland et al, 1998 - SE) (Rolland et Benachour, 1998 - DKE) (Tawbi et Souveyet, 1999) et SAVRE (Sutcliffe et al, 1998) (Maiden, 1998). La méthode L'Écritoire propose des directives de découverte des besoins fonctionnels des systèmes exprimés sous forme d'intentions et conceptualise ces besoins comme des scénarios permettant d'atteindre ces intentions. La méthode SAVRE propose des directives pour découvrir les dysfonctionnements d'un système causés par des erreurs humaines en générant des scénarios correspondants aux besoins du système et en les analysant pour découvrir de potentielles exceptions dues à des erreurs humaines.

Prise de décision (Kornysheva et al, 2011a). MADISE est une famille de méthodes permettant de guider les utilisateurs à travers les activités de prise de décision. La création de cette famille a permis de remplir plusieurs objectifs : elle permet de clarifier et d'organiser les concepts de prise de décision en apportant une vision générique de ceux-ci ; elle permet de comparer les différents modèles existants pour une meilleure sélection, en valide la complétude et permet d'en créer de nouveaux. Les méthodes utilisées pour construire la famille dédiée à la prise de décision étaient essentiellement des méthodes dites multi-critères telles que MAUT (Multiattribute Utility Theory) (Keeney et al., 1993), AHP (Analytic Hierarchy Process) (Saaty, 1980), des méthodes de surclassement (Roy, 1996), des méthodes de pondération (Keeney, 1999), et des méthodes floues (Fuller et al., 1996).

Co-création de services transdisciplinaires (Ralyté, 2013). Cette famille explore comment les contextes de collaboration inter-organisationnels affectent le processus d'ingénierie de nouveaux services d'information, en particulier ceux dédiés au support de la collaboration entre partenaires de secteurs différents (public ou professionnel). Ces services sont appelés des services d'information transdisciplinaires. Différentes techniques et méthodes ont été utilisées pour créer les composants de la famille de méthodes, entre autres le modèle e3value (Gordijn & Akkermans, 2001), i* (Yu, 1995), certaines techniques de pensées créatives (Michalko, 2006, 2011) ou encore quelques jeux d'innovation (Gray et al., 2010) (Hohmann, 2009).

Analyse de données en Sciences Humaines et Sociales (SHS) (Ammar et al, 2014). Les chercheurs en SHS ont chacun leur propre manière de travailler les données (collection, codage, analyse et interprétation). Ces manières sont très hétérogènes et nécessitent une représentation unique et partagée. De plus, beaucoup d'analyses de données sont faites avec des méthodes statistiques d'analyse, pour trouver des corrélations entre les événements ou faire des prédictions ou des hypothèses. La famille de méthodes dédiée à ce problème propose un processus flexible et organisé, composé de composants de méthodes tirés de plusieurs méthodes statistiques (par exemple, ACP - Analyse en Composantes Principales) et offrant une vision globale des diverses possibilités d'analyse de données offertes à l'utilisateur.

Lancement de projets agiles (Deneckere et al, 2014). Les méthodes agiles sont en général définies comme des ensembles de bonnes pratiques et beaucoup de documents tentent de les expliquer. Ceux-ci manquent cependant d'explications claires sur le processus à appliquer, sur l'observation que l'on se doit d'être agile si l'on veut faire

de l'agile. Cependant, dans la jungle des tâches, principes, conseils et recommandations, les nouveaux utilisateurs de méthodes agiles sont parfois perdus dans les différentes alternatives qui leur sont offertes, sans argumentation claire pour choisir l'une envers les autres. La famille de méthodes agiles propose un ensemble de composants organisés dans un processus clair et facile à comprendre spécialisé dans la phase de lancement des projets agiles. Les méthodes utilisées sont SCRUM (Schaber et Beedle, 2001), XP (Beck, 1999), Crystal Clear (Cockburn, 2002) et DSDM (Stapleton, 1995).

5.2. Evaluation de la famille de méthodes « Lancement des projets agiles »

Cette dernière famille de méthodes a été évaluée récemment avec une centaine d'utilisateurs (Deneckere et al, 2015) dans une enquête effectuée à l'aide d'un questionnaire. Nous avons tenté de répondre à quatre questions principales : L'usage de plusieurs méthodes agiles dans le même projet est-il utile et pertinent ? La famille offre-t-elle une meilleure compréhension des méthodes agiles ? La famille facilite-t-elle le lancement d'un projet agile ? La famille est-elle cohérente avec les pratiques habituelles des méthodes agiles ? Nous avons organisé des ateliers pour expliquer la famille de méthodes et répondre aux différentes questions soulevées. Ensuite nous avons demandé aux participants de répondre à un questionnaire qui nous a permis de faire les constats suivants.

- Les utilisateurs sont conscients qu'appliquer plusieurs méthodes dans le même projet est pertinent et peut s'avérer très utile. En effet, 63% des personnes ayant déjà été confronté à cette situation en sont satisfaits et 45% de ceux qui n'ont jamais eu à appliquer plusieurs méthodes dans le même projet pensent que cela pourrait être une excellente idée.
- La majorité des utilisateurs (76%) sont d'accord sur le fait que la famille de lancement des projets agiles facilite la compréhension de cette phase.
- 65% des participants à l'enquête s'accordent sur le fait que la famille est utile pour améliorer la qualité des méthodes de développement agiles. La plupart d'entre eux (54%) reconnaissent également que l'utilisation de la famille permet d'éviter les erreurs dans le processus de lancement.
- La majorité des répondants (71%) sont d'accord sur le fait que la famille représente très bien le processus de lancement de projets agiles tel qu'il est actuellement pratiqué.

Cette enquête a montré que cette famille est pertinente et facilite la phase dédiée au lancement des projets agiles. Elle nous a également donné quelques indications sur ce que les utilisateurs pensaient du concept général de famille de méthodes.

6. Conclusion et les travaux futurs

Les familles permettent d'organiser un ensemble de composants de méthodes (venant de méthodes différentes) pour un même domaine, selon une même unité téléologique (un même objectif). Leur but principal est de faciliter la réutilisation et l'adaptation des composants. Au lieu de créer de toutes pièces une méthode adaptée au contexte, comme le suggère la majorité des approches de l'IMS, l'utilisateur de famille de méthodes peut tout simplement configurer sa ligne de méthode à partir de la fa-

mille. L'effort accordé à la phase de création de la méthode sera considérablement réduit alors que l'offre de composants sera elle beaucoup plus large.

Le concept de famille de méthodes a fait ses preuves dans plusieurs domaines. La famille de méthodes dédiée au lancement de projets agiles a été évaluée de manière très positive lors d'une enquête préalable. L'évaluation présentée ici nous a permis d'étudier le ressenti des utilisateurs face au concept même de famille, quel que soit le domaine pris en compte. Effectuée sur 46 utilisateurs, les répondants ont largement compris l'intérêt de l'approche, la trouvent plutôt intuitive et certains envisagent même de l'utiliser au niveau professionnel. Le ressenti des répondants est donc très positif sur le concept de famille de méthodes. Le questionnaire précédent sur la famille dédiée au lancement de projets agiles a soulevé une demande sur le processus de la phase du développement agile, notre prochain travail sera donc destiné à étudier cette phase plus complexe pour en déduire des composants de méthodes et les organiser sous forme de famille. Nous élargirons cette famille en y intégrant des composants d'autres méthodes agiles, comme Lean (Poppendieck et Poppendieck, 2003) ou RAD (Macmillan, 1991).

Les résultats de l'évaluation du concept de famille de méthodes sont plutôt encourageants. L'aspect multi-méthodes est au cœur du concept de famille de méthodes et l'un de nos objectifs était de savoir si l'expérience d'utilisation de plusieurs méthodes différentes impactait l'avis des participants sur le concept de famille de méthodes. Le résultat de l'évaluation montre que ces répondants étaient particulièrement intéressés par l'approche, tout en restant très conscients des difficultés engendrées par l'utilisation de plusieurs méthodes en parallèle.

Les résultats nous permettent également d'identifier des axes d'amélioration. Par exemple, il apparaît clairement que le guidage des familles n'a pas été complètement compris par les participants. Un travail de fond doit donc être accompli sur ce point pour améliorer la description et l'explication du guidage lors de l'utilisation de familles.

Bibliographie

- Ammar A., Hug H., Deneckere R. (2014) Intentional Process Modeling of Statistical Analysis Methods. *Computer Applications and Quantitative Methods in Archaeology (CAA)*, France.
- Beck K. (1999). *Extreme programming explained: Embrace change*. Addison-Wesley, UK
- Brinkemper S., Saeki M., Harmsen F. (1998). Assembly Techniques for Method Engineering. *CAiSE 1998*, pp. 381-400, LNCS 1413, Springer Verlag.
- Cervera M., Albert M., Torres V., Pelechano V. (2012). A Model-Driven Approach for the Design and Implementation of Software Development Methods. *International Journal of Information System Modeling and Design (IJISMD)*, vol. 3(4), pp. 86-103
- Clements P., Northrop L. (2001). *Software Product Lines: Practices and Patterns*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Cockburn A. (2002). *Agile software development*. Addison-Wesley, London, UK
- Collectif INFORSID (2012). La recherche en SI et ses nouvelles frontières. *Revue Ingénierie des Systèmes d'information (RSTI-ISI)*, Vol 17, n°3, pp 9-68.

- Deneckere R. (2001). *Approche d'extension de méthodes fondée sur l'utilisation de composants génériques*. Thèse de doctorat, Université de Paris 1-Sorbonne, 2001.
- Deneckere R., Kornysheva E., Rolland C. (2011) Method Family Description and Configuration. *ICEIS 2011*, Beijing, China. pp.384-387.
- Deneckere R., Kornysheva E., Ralyte R. (2014). Famille de méthodes : la flexibilité au cœur du processus de construction de méthode, *Revue Ingénierie des Systèmes d'information (RSTI-ISI)*, 19/1, pp.67-95
- Deneckere R., Kornysheva E. (2015) Agility in agile methods: A survey on the application of the method family approach, *soumis à CBI 2015*
- De Singly F. (2012). *L'enquête et ses méthodes – Le questionnaire*, Armand Colin.
- Firesmith D.G., Henderson-Sellers, B. (2002). *The OPEN Process Framework: An Introduction*. Addison-Wesley, London, UK, 330p.
- Fuller R., Carlsson C., Fuzzy multiple criteria decision making: Recent developments, *Journal Fuzzy Sets and Systems*, 78, (1996), pp. 139-153
- Gordijn, J. & Akkermans, H. (2001). E3-value: Design and Evaluation of e-Business Models. *IEEE Intelligent Systems*, 16(4). 11-17.
- Gray, D., Brown, S. & Macanufo, J. (2010). *Gamestorming: A Playbook for Innovators, Rulebreakers, and Changemakers*. O'Reilly Media, Inc, USA
- Guzélian G., Cauvet C. (2007). SO2M : Towards a service-oriented approach for method engineering, *IKE'07*, Las Vegas, Nevada, USA.
- Henderson-Sellers B., Ralyté J., Agerfalk, P. Rossi M. (2014) *Situational Method Engineering*, Springer
- Hohmann, L. (2009). *Innovation Games: Creating Breakthrough Products Through Collaborative Play*. Addison-Wesley.
- Kornysheva E., Ralyte J., Deneckere R. Constructing Method Families Based on the Variability Analysis. *RCIS 2013 Forum*, May 2013, Paris, France. pp.1.
- Kornysheva E. (2011). *MADISE: Method Engineering-based Approach for Enhancing Decision-Making in Information Systems Engineering*, thèse de doctorat de l'université Paris 1 Panthéon-Sorbonne, Paris, France.
- Kornysheva E., Deneckère R., Rolland C. (2011a). Method Families Concept: Application to Decision-Making Methods. *Int. conf. EMMSAD*, pp. 413–427, LNBIP 81, Springer.
- Kornysheva K., Deneckere R., Claudepierre B. (2011b). Towards Method Component Contextualization. *International Journal of Information System Modeling and Design (IJISMD)*, IGI Global, 2 (4), pp.49-81.
- Karlsson F., Ågerfalk P.J. (2004). Method Configuration: Adapting to Situational Characteristics while Creating Reusable Assets, *Journal IST*, Vol.46 (9).
- Keeney R.L., Raiffa H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*, Cambridge University Press
- Keeney R.L. (1999). *Foundations for Making Smart Decisions*, IIE Solutions, 31, No. 5, pp. 24-30
- Kumar K., Welke R.J. (1992). Methodology EngineeringR: A Proposal for Situation Specific Methodology Construction. *Challenges and Strategies for Research in Systems Development*, Cotterman, W. and J. Senn (eds.), J. Wiley, Chichester, UK, pp. 257-266.

- Iacovelli A. (2012). *Approche orientée service pour la configuration de méthodes outillées*. Thèse de doctorat, Université Paris 1 Panthéon-Sorbonne, Paris, France
- Maiden N.A.M. (1998). CREWS-SAVRE: Scenarios for Acquiring and Validating Requirements, *Journal of Automated Software Engineering*
- Macmillan J.M. (1991). *Rapid Application Development*, Coll. Div
- Michalko, M. (2006). *Thinkertoys: A Handbook of Creative-Thinking Techniques*. Ten Speed Press.
- Michalko, M. (2011). *Creative Thinkering: Putting Your Imagination to Work*. New World Library.
- Pohl K., Böckle G., van der Linden F. (2005). *Software product line engineering: foundations, principles and techniques*, Springer, Berlin Heidelberg New York.
- Poppendieck M., Poppendieck T. (2003). *Lean Software Development: An Agile Toolkit Paperback*, Eds Addison-Wesley
- Ralyté, J. (2013). Towards a Method Family Supporting Information Services Co-creation in the Transdisciplinary Context, *IJISMD*, Vol. 4(3), IGI-Global
- Ralyté J., Rolland C. (2001). An Assembly Process Model for Method Engineering. *CAISE 2001*, LNCS 2068, Springer, Berlin, pp. 267-283.
- Rolland C., Prakash N., Benjamin A. (1999). A Multi-Model View of Process Modelling” *Requirements Engineering Journal (REJ)*, vol 4(4), pp. 169–187.
- Rolland C., Souveyet C., Ben Achour C. (1998). Guiding Goal Modelling Using Scenarios, *IEEE Transactions on Software Engineering, special issue on Scenario Management*, Vol. 24, No. 12, pp 1055-1071
- Rolland C., Ben Achour C. (1998). Guiding the construction of textual use case specifications, *Data & Knowledge Engineering Journal*, P. Chen, R.P. van de Riet (eds), North Holland, Elsevier Science Publishers, Vol. 25 N° 1, pp. 125-160.
- Rossi M., Ramesh B., Lyytinen K., Tolvanen J-P. (2004). Managing evolutionary method engineering by method rationale, *Journal of the AIS*, 5(9): 356-391.
- Roy B. (1996). *Multicriteria Methodology for Decision Aiding*, Dordrecht, Kluwer Academic Publishers
- Sutcliffe A.G., Maiden N.A.M., Minocha S., Manuel D. (1998). Supporting Scenario-based Requirements Engineering, *IEEE Transactions on Software Engineering: Special Issue on Scenario Management*, Vol. 24, No. 12
- Tawbi M., Souveyet C. (1999). Guiding Requirement Engineering with a Process Map, *MFPE'99*, Gammarth, Tunisia
- Saaty T.L. (1980). *The Analytic Hierarchy Process*, NY, McGraw Hill
- Schwaber K., Beedle M. (2001). *Agile software development with scrum*. Prentice Hall PTR, Australia
- Stapleton J. (1995). *DSDM – Dynamic system development method*. Addison-Wesley, UK
- Weiss D.M., Lai C.T.R. (1999). *Software product-line engineering: a family-based software development process*. Addison-Wesley.
- Yu E. (1995). Modelling Strategic Relationships for Process Reengineering, thèse de doctorat, Université de Toronto

Simulation des Systèmes d'Information des Smart Grids

Une approche par points de vue

**Rachida Seghiri^{1,2}, Frédéric Boulanger², Claire Lecocq³,
Vincent Godefroy¹**

1. EDF R&D

Clamart, France

prénom.nom@edf.fr

2. LRI, Supélec

Gif-sur-Yvettes, France

prénom.nom@supelec.fr

3. Institut Mines-Télécom/Télécom SudParis, SAMOVAR CNRS UMR 5157

Evry, France

prénom.nom@telecom-sudparis.fr

RÉSUMÉ. Nous proposons de simuler les Systèmes d'Information des Smart Grids afin de valider/critiquer les choix de modélisation des experts métier. Nous proposons une approche par points de vue selon deux aspects information et dynamique tout en ajoutant un point de vue intégration pour le maintien de cohérence. Nous mettons à contribution les techniques l'Ingénierie Dirigée par les Modèles, en particulier la transformation de modèle. Enfin, nous éprouvons notre démarche sur un cas métier Smart Grid.

ABSTRACT. We propose to simulate the Information Systems of Smart Grids to validate / criticize the modeling choices of domain experts. We propose a viewpoint based approach regarding two aspects : information and dynamics. We add an integration viewpoint to maintain consistency and we use Model Driven Engineering techniques. Finally, we experience our approach on a Smart Grid use case.

MOTS-CLÉS : Simulation, Ingénierie Dirigée par les Modèles, Smart Grid, Système d'Information

KEYWORDS: Simulation, Model Driven Engineering, Smart Grid, Information System

1. Introduction

Un Smart Grid est un réseau électrique intelligent permettant d'optimiser la production, la distribution et la consommation d'électricité grâce à l'introduction des technologies de l'information et de la communication sur le réseau électrique¹. Les Systèmes d'Information (SI) doivent donc intégrer pleinement les Smart Grids pour les superviser et les piloter. Dans ce contexte, nous proposons de simuler ces SI en adoptant différents points de vue (métier, fonctionnel, applicatif) afin de faire valider ces simulations par les experts impliqués dans l'implantation des Smart Grids. Nous identifions principalement trois verrous à lever pour y parvenir :

1. Tout d'abord, le Smart Grid implique plusieurs expertises car il combine le domaine électrotechnique, celui des télécommunications et les SI. Les modèles à simuler doivent respecter deux critères : (1) s'appuyer sur des standards pérennes, (2) être compréhensibles par les différents experts métier. Nous expérimentons les approches issues de l'Ingénierie Dirigée par les Modèles (IDM) pour concevoir des modèles respectant ces critères. En effet, l'IDM est à l'origine de nombreux travaux facilitant la prise en compte de standards comme UML dans les processus de développement (Authosserre-Cavarero *et al.*, 2012). De plus, ces standards sont effectivement adoptés par les experts Smart Grids (Uslar *et al.*, 2012) ;

2. Ensuite, pour les simuler, les modèles construits doivent être exécutables. Par la simulation, les experts peuvent valider/critiquer leurs choix de modélisation au plus tôt dans le processus de développement. Des langages comme BPMN² et UML possèdent désormais des sémantiques d'exécution formalisées permettant le développement d'outils de simulation ;

3. Finalement, une approche par points de vue est indispensable à l'appréhension de systèmes complexes (France, Rumpe, 2007) tels que les SI des Smart Grids mais il faut absolument maintenir la cohérence entre les vues du système. (Lahna *et al.*, 2005) propose de décrire des correspondances au niveau du métamodèle UML enrichi. Cette approche peut être généralisée en définissant des liens de cohérence entre les vues indépendamment du langage de modélisation.

Nous proposons donc une démarche intégrative par points de vue pour la simulation des SI des Smart Grids en profitant des techniques issues de l'IDM comme la transformation de modèle et en adoptant des modèles exécutables. La deuxième section de cet article décrit le contexte industriel dans lequel s'inscrivent nos travaux. La section 3 présente un état de l'art des concepts et des techniques de l'IDM, des framework d'architecture SI et de la simulation de modèles pour le SI. Notre démarche, ainsi que son application à un cas métier, sont présentées dans la section 4. La dernière section conclut cet article et donne nos perspectives de recherche.

1. www.smartgrids-cre.fr

2. Business Process Management Notation

2. Contexte industriel

Dans cette partie, nous présentons la problématique industrielle dans laquelle s'inscrivent nos travaux avant d'introduire le cas métier sur lequel nous souhaitons éprouver notre démarche.

2.1. Problématique industrielle

L'équilibre entre la consommation et la production est essentiel au bon fonctionnement des réseaux électriques. Les Smart Grids sont d'autant plus incontournables pour préserver cet équilibre avec l'arrivée massive des véhicules électriques dont la recharge est imprévisible et mobile, ainsi que l'implantation massive des énergies intermittentes (produites par exemple par les éoliennes, les panneaux solaires). En effet, les Smart Grids assurent une gestion automatique et en temps réel des réseaux électriques grâce à de multiples capteurs et points de contrôle commandés à distance.

Les Smart Grids apportent, par essence, de profonds changements au niveau des Systèmes d'Information qui les pilotent : nouveaux flux d'information provenant du réseau électrique, entrée en jeu de nouveaux acteurs tels que les producteurs décentralisés (éolien, photovoltaïque), nouveaux équipements communicants comme le compteur Linky³, nécessaire conformité aux nouvelles réglementations et directives européennes⁴, nouveaux usages (véhicule électrique, maison connectée).

Afin d'appréhender ces paradigmes naissants, il est indispensable d'élaborer des scénarios métier nécessitant d'être éprouvés et validés avant leur adoption finale. Plusieurs démonstrateurs physiques ont été déployés sur le terrain⁵. Ces projets pilotes permettent de mener des expérimentations en conditions réelles pour tester des fonctions et des services comme par exemple le démonstrateur InfiniDrive⁶ pour le pilotage des infrastructures de recharge de véhicules électriques ou le démonstrateur Venteea⁷ pour l'intégration de forte capacité éolienne dans un réseau rural. Cependant, les démonstrateurs nécessitent que le gestionnaire de réseau de distribution recrute des clients industriels et/ou domestiques qui acceptent d'avoir du matériel à tester chez eux. De plus, leur exploitation reste limitée par les réglementations en cours. Enfin, leur mise en place se révèle souvent longue et coûteuse.

En plus de ces démonstrateurs, des réseaux de distribution d'expérimentation grandeur nature comme Concept Grid⁸, implanté à EDF Lab, permettent de tester les nouveaux équipements avant leur installation sur les réseaux du distributeur ERDF⁹. Ces réseaux ont l'avantage de permettre la conduite de stress tests en conditions perturbées, impossibles à réaliser dans le cadre de démonstrateurs, ceux-ci impliquant de vrais

3. www.erdf.fr/Linky

4. www.horizon2020.gouv.fr

5. www.erdf.fr/Carte_demonstrateurs_Smart_Grids

6. avem.fr/actualite-erdf-et-le-groupe-la-poste-lancent-le-projet-infini-drive-a-nice-3450.html

7. www.venteea.fr

8. chercheurs.edf.com

9. Électricité Réseau Distribution France

clients. Cependant, la taille réduite de ces réseaux reste limitante.

Pour palier toutes ces limitations, une troisième voie est la simulation. Cette simulation intègre les trois domaines qui composent les Smart Grids : l'infrastructure électrique (transformateurs, lignes, charges, sources), l'infrastructure télécom (réseaux mobiles, CPL¹⁰) et enfin les SI qui les pilotent. Des simulateurs spécialisés dans la simulation de réseaux électriques (EMTP-RV, Dymola, PowerFactory, Eurostag, etc.) ainsi que des simulateurs télécoms (OPNET, NS-3, OMNeT ++, etc.) ont déjà validé l'apport de la simulation dans leurs domaines respectifs. Toutefois, la couche SI est souvent reléguée à de simples modèles de calcul de consigne souvent développés en Matlab ou en C++ (Palensky *et al.*, 2014). Nous souhaitons donc simuler aussi le domaine SI et faire valider les simulations obtenues par les experts impliqués dans le déploiement des Smart Grids.

2.2. Cas métier : gestion d'une flotte de véhicules électriques d'entreprise

De nouveaux usages de consommation d'électricité émergent dont, notamment, la mobilité électrique. En France, les pouvoirs publics ont annoncé leur ambition de miser sur deux millions de véhicules électriques en circulation en 2020. Ce parti pris découle de l'objectif, dit du « facteur 4 », visant à diviser par quatre les émissions de gaz à effet de serre à l'horizon 2050 par rapport à son niveau de 1990.

Dans ce contexte, plusieurs projets R&D visent à gérer au mieux l'impact des véhicules électriques sur les réseaux de distribution d'électricité. En effet, la recharge complète d'un véhicule électrique ayant 150 km d'autonomie est équivalente en terme d'appel de puissance à :

- un chauffe-eau si la recharge s'effectue en 8h (recharge normale) ;
- un immeuble si la recharge s'effectue en 1 h (recharge accélérée) ;
- un quartier urbain si la recharge s'effectue en 3 min (recharge rapide).

Les moyens sollicités pour maintenir l'équilibre offre/demande et garantir la qualité de l'électricité fournie pendant les pics de consommation sont coûteux et émetteurs de CO₂ (démarrage de centrales de charbon et de fioul). Des expérimentations en cours testent de nouvelles méthodes pour éviter les contraintes réseau. Par exemple, la Poste et ERDF ont mis en place le démonstrateur InfiniDrive visant à concevoir un système de recharge optimal pour une flotte de véhicules électriques d'entreprise. Ce système propose de piloter la recharge en minimisant la concentration de recharge rapide et accélérée sur les heures de pointe (risque de mise en contrainte du réseau voire du parc de production et de panne sur le réseau), tout en prenant en compte les incitations tarifaires favorisant la recharge en heure creuse.

Par ailleurs, la mobilité électrique implique un changement de paradigme pour le gestionnaire de flotte de l'entreprise. D'une part, le véhicule électrique est limité par son autonomie et ne peut donc pas effectuer n'importe quelle tournée. D'autre part, la recharge d'un véhicule électrique ne se fait pas comme le plein d'essence d'un véhicule

10. Courant Porteur en Ligne

thermique (temps de recharge, disponibilité des bornes).

Dès lors, le processus d'affectation de véhicules aux tournées des agents, la gestion de la flotte de véhicules dans son ensemble et donc le SI qui l'implante sont fortement impactés par l'arrivée massive des véhicules électriques. Nous proposons de modéliser les SI impactés et de les simuler. Cette simulation permettra de valider/critiquer les choix de modélisation et d'anticiper l'éventuel dimensionnement de la flotte. Par exemple, si une forte proportion des tournées implique une distance effectuée supérieure à l'autonomie des véhicules électriques sans possibilité de recharge en cours de route (pas de borne à disposition au cours de la tournée), la simulation permet de trouver la proportion de véhicules thermiques à garder a minima dans une flotte. L'affectation doit aussi privilégier l'utilisation des véhicules électriques car la rentabilité d'un parc de véhicules électriques est proportionnelle au nombre de kilomètres effectués par ces véhicules.

Ce cas métier fait intervenir aussi bien un SI centralisé (celui du pilotage de la flotte de véhicules et du calcul du programme de recharge optimisé), qu'un SI réparti (celui embarqué dans les bornes qui exécutent le programme de recharge calculé en identifiant le véhicule électrique concerné).

3. État de l'art

Les Smart Grids sont, par leur nature même, des systèmes complexes (Monti, Ponci, 2010). L'IDM a démontré à travers de nombreux travaux sa capacité à adresser de tels systèmes (France, Rumpe, 2007). La première partie de cet état de l'art traite de l'IDM. De plus, l'objet de nos travaux est de simuler le domaine SI des Smart Grids. La partie 3.2 présente ainsi les principaux points de vue adoptés dans l'architecture des SI et la partie 3.3 précise les objectifs de la simulation SI et les moyens d'y parvenir.

3.1. IDM

L'IDM prône le paradigme du « tout est modèle » (Bézivin, 2005). En effet, c'est une approche orientée modèle qui couvre tout le cycle de vie d'un système logiciel en utilisant des modèles « productifs » (Bézivin *et al.*, 2004), c'est-à-dire, des modèles exploitables informatiquement (simulation, validation, vérification, génération de code, etc.). Les modèles sont donc des artefacts de premier ordre dans l'IDM. Bien que la définition du concept de modèle ait soulevé de nombreux débats, une définition communément admise est donnée par (Bézivin, Gerbé, 2001) : « *un modèle est une abstraction d'un système, construite selon le bon point de vue, qui permet de répondre à des questions prédéfinies sur ce système en lieu et place de celui-ci* ». De cette définition découle la première relation fondamentale de l'IDM, nommée ReprésentationDe, qui lie le modèle et le système qu'il représente. En outre, l'IDM permet de capturer la logique métier d'un domaine au niveau des métamodèles. En effet, un métamodèle décrit l'ensemble des concepts métier manipulés ainsi que leurs relations. Pour être exploitable, tout modèle utilisé au cours du cycle de vie du système à implanter doit se conformer à cette logique métier et donc au métamodèle. Ceci nous amène à la

deuxième relation fondamentale de l'IDM qui relie un modèle à un métamodèle et qui est la relation « ConformeÀ » (Favre, 2004).

L'intérêt majeur de l'IDM est de produire des modèles dont l'exploitation est automatisable. Ceci est rendu possible grâce à la transformation de modèle. La transformation génère un ou plusieurs modèles cible à partir d'un ou plusieurs modèles source, selon une description établie de la transformation (Mens, Van Gorp, 2006). Comme l'illustre la FIGURE 1, la description de la transformation se fait au niveau des métamodèles. Un moteur de transformation exécute cette description.

La transformation de modèle est au cœur d'une démarche dirigée par les modèles. Il est possible d'y recourir sur tout le cycle de vie d'un système pour par exemple le raffinement, la composition, l'analyse ou encore la simulation de modèle. Ainsi, l'IDM promeut l'utilisation de modèles pour construire des SI évolutifs, s'alignant rapidement avec les besoins métier grâce aux techniques de transformation de modèle et facilitant la capitalisation du savoir métier dans les métamodèles élaborés.

Cependant, la multitude de modèles produits et manipulés doit former un ensemble cohérent et représentatif du système final quel que soit le point de vue adopté. C'est pourquoi le maintien de la cohérence entre modèles et transformation de modèles fait l'objet de recherche active dans une discipline comme l'IDM. (Steel, Jézéquel, 2007) propose de typer les modèles en entrée des transformations en utilisant le Model Typing. Celui-ci permet de contrôler les modèles manipulés par les transformations. Autrement dit, une transformation ne prend en entrée qu'un certain type de modèle. Ainsi, tous les modèles conformes à ce type peuvent être manipulés par la transformation en question. Ceci a, de plus, l'avantage d'augmenter la réutilisabilité des transformations écrites en mettant en évidence les caractéristiques communes des modèles. Cette technique de typage est en outre outillée et intégrée dans le langage de méta-modélisation Kermeta¹¹ qui est basé sur EMOF¹² dans un environnement Eclipse.

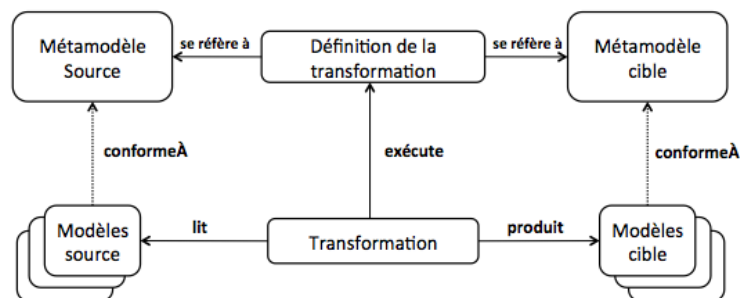


Figure 1. Composants d'une transformation de modèles

11. www.kermeta.org

12. Ecore Meta Object Facility

3.2. Architecture SI

Une définition communément admise du SI est donnée par (Reix *et al.*, 1995) : « *un Système d'Information est un ensemble organisé de ressources : matériel, logiciel, personnel, données, procédures, permettant d'acquérir, traiter, stocker, communiquer des informations (texte, images, sons, etc.) dans et entre des organisations* ». Cette définition du SI a l'avantage de mettre l'accent sur les acteurs impliqués et les informations manipulées outre les moyens techniques mis en œuvre pour les traiter, évitant ainsi de réduire le SI au seul système informatique. Une modélisation appropriée du SI reflète donc les besoins des acteurs (expert du domaine métier, architecte fonctionnel, architecte technique, etc.) en explicitant les informations manipulées.

Étant donnée la complexité des SI actuels, en particulier ceux des Smart Grids, et la multitude des acteurs concernés, une vision monolithique est inappropriée à la construction de SI évolutifs et adaptés aux différents acteurs.

Plusieurs cadres d'architecture adoptent une approche par points de vue. Un point de vue formalise la perspective d'un acteur particulier du SI. Une vue est conforme à ce point de vue. C'est le cas de RM-ODP (Raymond, 1995), de TOGAF^{13 14}, de la méthode « 4+1 » de Kruchten (Kruchten, 1995) ainsi que de la norme ISO/IEC/IEEE 42010¹⁵. Le SGAM¹⁶ (Uslar *et al.*, 2012) adresse l'architecture du Smart Grid en englobant les trois domaines : SI, réseau électrique et réseau de télécommunication.

En outre, ces frameworks organisent hiérarchiquement les différentes vues en appliquant « *IT follows business* » comme principe : commencer par le point de vue métier et le dériver progressivement jusqu'à l'infrastructure technique déployée en passant par les fonctions et les applications. Souvent, ces cadres d'architecture distinguent quatre points de vue principaux :

Point de vue métier : ce point de vue reflète la vision métier. On y retrouve les objectifs métier de l'entreprise, les processus, ainsi que les acteurs ;

Point de vue fonctionnel : ce point de vue organise le SI en blocs fonctionnels de manière à garantir son évolutivité tout en répondant aux besoins métier de l'entreprise. A l'échelle du SI d'une entreprise, cette structuration devient vite complexe à cause du caractère étendu et transverse des processus métier impactés ;

Point de vue applicatif : ce point de vue structure le SI en blocs applicatifs, chacun implantant un ou plusieurs blocs fonctionnels ;

Point de vue technique : ce point de vue correspond à l'infrastructure technique du SI (matériel informatique et réseaux télécom).

D'une part, ces frameworks proposent une architecture orientée composants (macros processus, blocs fonctionnels, blocs applicatifs, etc.). Les informations sont modélisées soit implicitement et de manière diffuse à l'intérieur des vues (TOGAF), soit

13. The Open Group Architecture Framework

14. www.opengroup.fr/togaf

15. <http://www.iso-architecture.org/>

16. Smart Grid Architecture Model

séparément dans une vue dédiée et décorrélée des autres vues (RM-ODP, Kruchten, SGAM).

D'autre part, les modèles utilisés pour les vues — métier, fonctionnel, applicatif — sont souvent des modèles « contemplatifs », destinés à la simple communication ou à la documentation et sont, de ce fait, désincarnés de l'implantation informatique. L'utilisation de modèles exécutables pour la modélisation de ces vues assure une meilleure cohérence des vues du SI et en facilite l'appréhension par les différents acteurs en corrigeant/validant la simulation de ces modèles au plus tôt.

3.3. *Simulation du SI*

(Shannon, 1975) définit la simulation comme étant « *le processus consistant à modéliser un système réel et à mener des expérimentations sur le modèle obtenu dans le but de comprendre le comportement du système et/ou d'évaluer différentes stratégies concernant son fonctionnement* ». Quel qu'en soit le domaine d'application, la simulation est un moyen d'apprécier les choix des concepteurs sur la structure et le comportement du système modélisé. Elle peut se traduire par l'animation d'un modèle (représentant notre perception du système, qu'il soit existant ou à construire) et l'étude du comportement de ce modèle en fonction des variables en entrée.

La simulation des SI des Smart Grids est d'autant plus cruciale que ces derniers sont en constante et rapide transformation : évolution des cadres législatifs, apparition de nouveaux partenaires, hétérogénéité des interactions avec les clients finaux (compteurs intelligents, web, téléphones, tablettes, etc.). Ainsi, le recours aux modèles exécutables dès les premières phases du cycle de vie des SI des Smart Grids augmente leur évolutivité en apportant une aide supplémentaire à leur validation. La capacité d'exécution et donc de simulation des modèles facilite leur exploration par les experts métier et peut lever les ambiguïtés engendrées par les modèles purement contemplatifs.

L'exécution des modèles est rendue possible par la définition d'une sémantique exécutoire du langage dans lequel ils sont exprimés. La sémantique d'un langage correspond au sens que peuvent prendre les concepts manipulés et leurs agencements lorsqu'ils sont instanciés au niveau des modèles (Jézéquel *et al.*, 2012). La définition de la sémantique du langage dépend de l'objectif poursuivi : simulation, génération de code, vérification, compilation, etc. L'expression de la sémantique d'un langage fait l'objet d'intenses recherches en ingénierie des langages, et en particulier sous la thématique des langages formels (Kleppe, 2007).

Comme en atteste le manifeste d'IBM (Chesbrough, Spohrer, 2006), les axes principaux de l'IDM sont (1) les standards ouverts, (2) l'automatisation et (3) la représentation directe. Compte tenu de ces axes, pour modéliser les différentes vues du SI, nous préconisons l'utilisation de langages standardisés, exécutables et compréhensibles par les acteurs du SI des Smart Grids. Nous identifions plusieurs langages pouvant satisfaire ces critères :

- Un sous-ensemble de UML limité au diagramme de classes et au diagramme d'activité possède désormais une sémantique d'exécution décrite par le standard

fUML¹⁷. Les diagrammes de classes conviennent pour la description des modèles d'information tandis que les diagrammes d'activité sont adaptés à la description de la dynamique d'un modèle et au comportement attendu des fonctions ;

- Le standard BPMN est un langage de modélisation graphique permettant de décrire tous les aspects d'un processus métier à l'aide d'un seul type de diagramme. Ce formalisme présente l'avantage d'avoir une sémantique d'exécution bien définie permettant le développement d'outils pour la simulation de modèles de processus métier. BPMN est parfaitement adapté à la description de processus métier ;

- Le standard OCL¹⁸, un langage textuel standard d'expression de contraintes, a été ajouté à UML pour exprimer les propriétés difficiles à capturer dans des diagrammes UML. L'exécution d'OCL se fait à travers la transformation de modèle en ciblant un langage d'expression de contrainte de plus bas niveau qui soit exécutable comme MiniZinc (Nethercote *et al.*, 2007), ou à travers son utilisation au niveau du métamodèle avec OCLinEcore¹⁹.

Par la suite, nous présentons la démarche générale que nous proposons pour la modélisation et la simulation des SI des Smart Grid en adoptant une approche par points de vue et en mettant à contribution les techniques issues de l'IDM telles que la transformation de modèle et le Model Typing.

4. Démarche proposée et application au cas métier

Dans cette partie, nous formalisons tout d'abord notre démarche générale avant de la valider sur le cas métier de la gestion d'une flotte de véhicules électriques.

4.1. Démarche générale

Toute simulation d'un système commence par sa modélisation. Pour modéliser des systèmes complexes comme les SI des Smart Grid, nous adoptons une approche par points de vue. Celle-ci facilite la conception des modèles par les acteurs impliqués en séparant leurs préoccupations respectives. Elle permet également de présenter les modèles obtenus, ainsi que les résultats de simulation de manière plus compréhensible à ces acteurs, car chaque point de vue n'utilise que les concepts métier propres à chaque acteur, selon sa perspective sur le SI.

L'approche que nous proposons identifie quatre points de vue selon les acteurs, spécialistes et interlocuteurs intervenant dans la modélisation et la validation par simulation du SI. Cette approche est illustrée FIGURE 2. Les trois premiers points de vue — métier, fonctionnel et applicatif — se retrouvent dans les approches d'architecture d'entreprise existantes. Cependant, ces approches traitent essentiellement l'aspect *dynamique* selon le point de vue abordé. Notre approche intègre ces points de vue selon

17. Foundational UML

18. Object Constraint Language

19. wiki.eclipse.org/OCL/OCLinEcore

deux aspects : *information* et *dynamique*. Nous nous limitons à trois points de vue — métier, fonctionnel et applicatif — dans un premier temps mais souhaitons étendre nos futurs travaux au point de vue technique également. Ainsi nous retrouvons :

1. L'aspect *dynamique* d'un point de vue métier qui reflète la perception des experts métier. On y retrouve les processus métier de l'entreprise, décrits en utilisant les concepts métier, sans référence aux détails d'implémentation. Nous préconisons l'utilisation de formalismes standard pour la modélisation de processus métier qui soient exécutables, tels que les diagrammes d'activité UML ou les diagrammes BPMN dans une perspective de simulation. Des langages spécifiques à un domaine (DSML) peuvent également être utilisés ;

2. L'aspect *dynamique* d'un point de vue fonctionnel qui reflète la perception de l'architecte fonctionnel et décrit les fonctions qui manipulent les données métier. Ces fonctions sont regroupées en blocs. Chaque objet métier identifié dans l'aspect Information de la vue métier correspond à un unique bloc fonctionnel. Ceci garantit la construction de blocs fonctionnels fortement décorrélés, avec une forte cohésion interne. Dans chaque bloc fonctionnel, on retrouve les opérations correspondant à une activité donnée du processus qui impacte l'objet métier impliqué ;

3. L'aspect *dynamique* d'un point de vue applicatif qui reflète la perception de l'architecte applicatif. Il aboutit à l'architecture applicative rassemblant les modules logiciels qui implantent les blocs fonctionnels. Dans un premier temps, il est conseillé de dresser un inventaire de l'existant applicatif et d'en extraire les modules capables de réaliser les opérations des blocs fonctionnels. Ensuite, si aucune application ou module existant ne peut répondre au besoin des nouveaux processus métier, l'architecte technique fait le choix des nouveaux composants applicatifs à mettre en place. En plus d'identifier les composants applicatifs existants ou à développer, l'architecte applicatif spécifie leurs interconnexions (échange de messages, synchronisation de données, transfert de fichiers périodique, etc.).

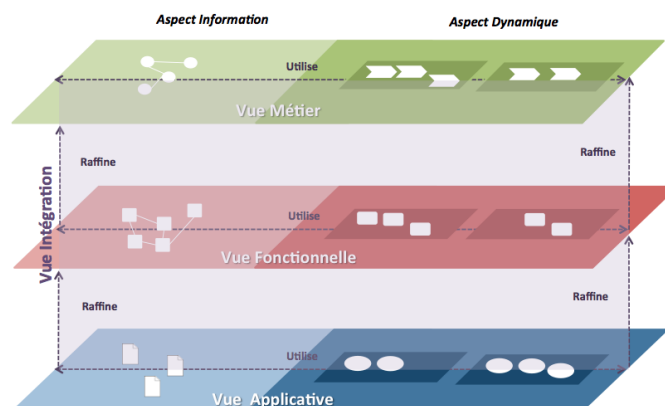


Figure 2. Démarche proposée

Nous proposons d'étendre chacune des trois vues classiques par l'aspect *information*

qui décrit les données qu'utilisent les modèles de l'aspect *dynamique*. L'aspect *information* permet d'avoir un modèle explicite des données utilisées dans chacune des vues métier, fonctionnelle et applicative :

1. L'aspect *information* d'un point de vue métier établit le « modèle de données métier » qui décrit les entités métier majeures manipulées par le processus. Ce modèle est peu sujet au changement, sauf évolution importante des pratiques métier. Il est aussi à l'origine du découpage en bloc par entité métier de la vue fonctionnelle ;

2. L'aspect *information* d'un point de vue fonctionnel établit le « modèle de données fonctionnelles » qui donne une description fine des données utilisées par les blocs fonctionnels nécessaires à la réalisation de processus métier. Elle décrit leurs caractéristiques et leurs relations sous forme de diagrammes de classes par exemple ;

3. L'aspect *information* d'un point de vue applicatif établit le « modèle de données applicatives », qui dépend fortement des applications choisies : il décrit les formats de données compatibles avec les modules applicatifs.

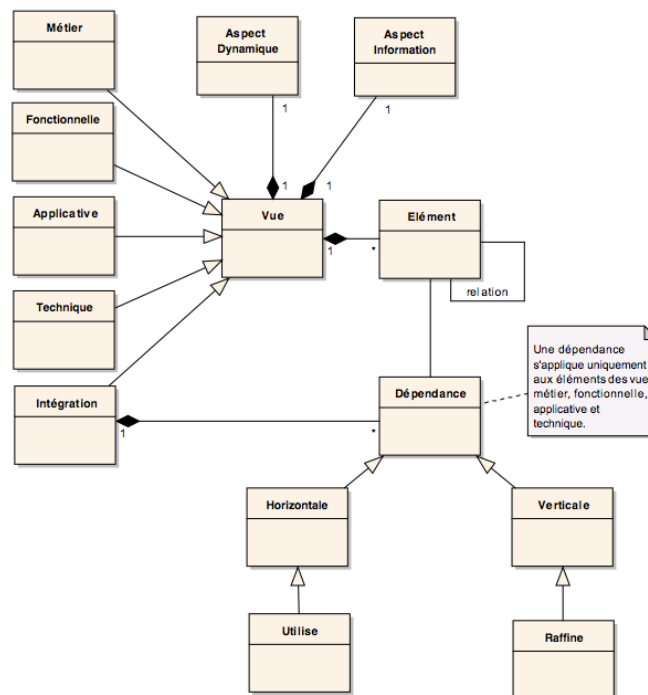


Figure 3. Métamodèle de la démarche proposée

Notre approche se distingue des démarches classiques par l'ajout du **point de vue intégration**. Ce point de vue permet une intégration « verticale » (entre les vues) et « horizontale » (à l'intérieur d'une seule vue). La FIGURE 3 donne un métamodèle de notre démarche en explicitant les concepts abordés et leurs relations. L'intégration verticale décrit les relations entre les trois points de vue précédents, étendus à leur aspect information. Ce point de vue intégration est donc transverse et permet de mo-

définir explicitement les notions de raffinement à travers la classe « Raffine », tant sur les aspects *dynamique* (processus, fonctions, applications) que sur les aspects *information* (représentation des données métier par des données fonctionnelles, codage des données fonctionnelles dans des formats applicatifs). Par exemple, elle permet de vérifier qu'une application implémente bien tous les blocs fonctionnels nécessaires au déroulement d'un processus métier. Cette vue donne ainsi accès aux informations de traçabilité qui permettent de déterminer l'impact d'une modification ou d'une défaillance d'un module applicatif sur les processus métier. Elle permet aussi de vérifier qu'un format applicatif permet d'encoder les données fonctionnelles qui représentent les données métiers, et de déterminer les éventuelles transformations de modèle nécessaires au déploiement. De la même façon, la vue intégration permet des vérifications « horizontales » à l'intérieur de chacune des vues. La classe « Utilise » assure donc la compatibilité des données échangées entre les tâches d'un processus métier, les fonctions d'un bloc fonctionnel ou entre les modules d'une application.

4.2. Mise en œuvre de la démarche sur le cas métier

Dans cette partie, nous éprouvons notre démarche au cas métier de la gestion d'une flotte de véhicules électriques. Nous construisons les modèles adéquats pour les vues métier, fonctionnelle et applicative en adoptant des langages exécutable. La cohérence est modélisée dans la vue intégration. L'architecture globale du cas métier est illustrée dans la FIGURE 4.

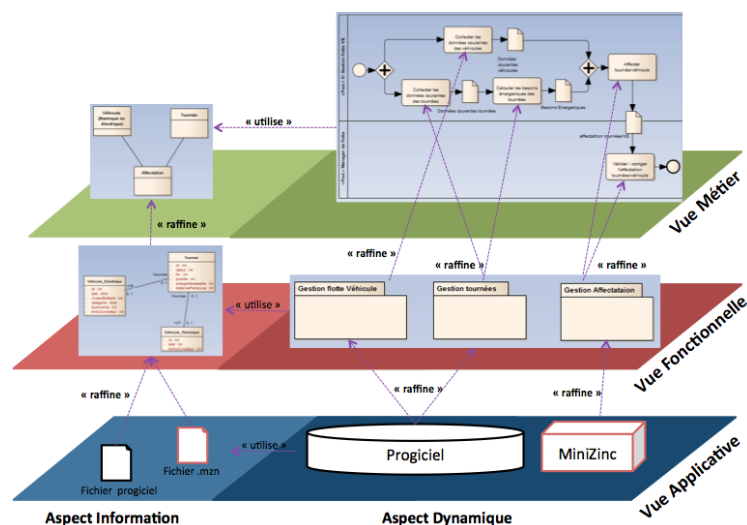


Figure 4. Architecture Globale du SI de gestion de flotte de véhicules électriques selon notre démarche

Vue métier

Pour cette vue nous préconisons l'utilisation de BPMN comme langage de modélisation exécutable. Nous modélisons et simulons le processus métier d'affectation de véhicule (électrique ou thermique) à une tournée avec l'outil Enterprise Architect²⁰. Nous utilisons donc un modèle de ce processus métier écrit en BPMN pour décrire l'aspect *dynamique* de la vue métier (FIGURE 4). Il s'agit de collecter les données relatives aux véhicules ainsi qu'aux tournées à effectuer, de calculer l'énergie nécessaire à chaque tournée et l'affectation véhicule/tournée avant de faire valider cette dernière par le manager de flotte. Pour l'aspect *information* de la vue métier, nous modélisons les objets métier sous la forme d'un diagramme de classes UML représentant les concepts de véhicule, de tournée et d'affectation tournée/véhicule. Le choix des langages de modélisation et de l'outil de simulation est motivé par les pratiques du domaine. En effet, la Commission Électrique Internationale a adopté Enterprise Architect comme outil pour maintenir et distribuer le CIM²¹ (Uslar *et al.*, 2012), un modèle d'information commun pour le domaine électrique²².

Vue fonctionnelle

Nous identifions trois blocs fonctionnels : un bloc pour la gestion de la flotte de véhicules (électriques et thermiques), un bloc pour la gestion des tournées, un bloc pour la gestion de l'affectation. Ces blocs servent à modéliser l'aspect *dynamique* de la vue fonctionnelle (FIGURE 4). En effet, les blocs fonctionnels sont responsables des traitements nécessaires à la réalisation du processus métier. Par exemple, nous modélisons l'affectation véhicule/tournée sous la forme de contraintes OCL (FIGURE 5) : pour affecter un véhicule à une tournée, il faut que l'énergie nécessaire à celle-ci soit inférieure à l'autonomie de la batterie.

```
context Tournee
inv: self.VE.autonomie * self.VE.niveauBatterie > self.distanceParcourue

context Tournee
inv: self.VE <> undefined xor self.VT <> undefined

context Tournee::kmElec(): int
body: (Tournee::allInstances()->collect(t.VE <> undefined |t.distanceParcourue))->sum()
```

Figure 5. Contraintes OCL pour la gestion d'affectation (vue fonctionnelle)

Nous considérons un premier cas où il n'est pas possible de recharger la batterie au cours de la tournée. L'aspect *information* de la vue fonctionnelle prend la forme de données fonctionnelles modélisées par un diagramme de classes sur lequel s'appliquent les contraintes OCL (FIGURE 4). OCL est adapté aux diagrammes de classes. De plus, OCL est un langage standardisé et exécutable pour l'expression de contraintes. Il est possible de modéliser les autres algorithmes de traitement (calcul des tournées à partir de bon de travaux, calcul de l'énergie nécessaire à une tournée, etc.) à l'aide de diagrammes d'activité exécutables.

20. www.sparxsystems.com.au

21. Common Information Model

22. www.sparxsystems.com.au/press/articles/iec.html

Vue applicative

Dans cette vue, nous identifions les applications nécessaires à l’implantation des blocs fonctionnels. Dans notre cas, le patrimoine applicatif de l’entreprise dispose déjà d’applications pour la gestion de tournées (calcul de tournées optimisé à partir de bons de travaux) et la gestion de véhicules (administration, maintenance, etc.). Nous choisissons MiniZinc pour modéliser les contraintes au niveau applicatif (FIGURE 6).

```
constraint forall (i in Tournées) (
%affectation de vehicule electrique si l autonomie l autorise
constraint forall(i in Tournées, j in VehiculesElec)
(tourneeVehicule[i]= identifiantVE[j] -> (autonomie[j]*niveauBatterie[j] > distanceTournee[i] /\
kmElec[i]=distanceTournee[i]));

% affectation d un vehicule thermique et dans ce cas kmElec est nul
constraint forall(i in Tournées, k in VehiculesTherm)
(tourneeVehicule[i]= identifiantVT[k] -> kmElec[i]=0);

%maximiser le nombre de km de tournee fait par les vehicules electriques
solve maximize sum(i in Tournées) (kmElec[i]);
```

Figure 6. Modélisation des contraintes en MiniZinc (vue applicative)

MiniZinc est un langage de modélisation et de résolution de contraintes de niveau intermédiaire qui a pour vocation de devenir un langage de modélisation standard dans le domaine de la programmation par contraintes. Pour l’aspect *information*, nous retrouvons les fichiers contenant les données nécessaires aux différentes applications (FIGURE 7).

```
nbreVE = 4;
identifiantVE = [1, 2, 3, 4];
nbreVT = 4;
identifiantVT = [11, 22, 33, 44];
autonomie = [1500, 1200, 1500, 1200];
niveauBatterie = [8, 6, 5, 2];
nbreTournées = 4;
distanceTournee = [1000, 700, 800, 1600];
```

Figure 7. Fichier de données MiniZinc (vue applicative)

Vue intégration

Nous avons modélisé les dépendances entre les vues par la relation «raffine» et entre aspects d’une même vue par la relation «utilise». Nous proposons un modèle de ces relations dans la FIGURE 8. Nous envisageons d’implémenter ces dépendances en utilisant le Model Typing pour assurer la cohérence entre ces vues et entre les aspects *information* et *dynamique* au sein d’une même vue. En plus de ces dépendances, la transformation peut jouer le rôle de garant de la cohérence. Par exemple, en transformant les contraintes OCL en contraintes MiniZinc tout en les raffinant par l’ajout de la fonction de maximisation de la distance parcourue par les véhicules électriques. ATL²³ permet d’écrire ce genre de transformation (Chenouard *et al.*, 2010).

23. <https://eclipse.org/atl/>

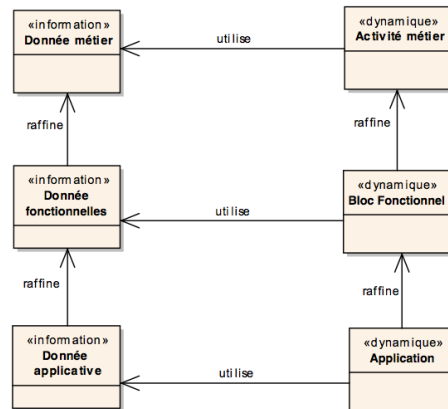


Figure 8. Modèle de la vue intégration

5. Conclusion et perspectives

Dans cet article, nous avons démontré l'importance de la simulation des SI des Smart Grids. Pour y parvenir, nous proposons une approche de modélisation et de simulation par points de vue. Nous retenons les vues métier, fonctionnel et applicatif. Outre l'aspect dynamique de ces vues, nous explicitons la modélisation de l'aspect informationnel. Nous proposons ensuite de rajouter une vue intégrative permettant de maintenir la cohérence horizontale (entre l'aspect dynamique et l'aspect information), ainsi que la cohérence verticale (entre les trois vues traditionnelles). Le Model Typing et les techniques de transformation de modèles sont de bons candidats pour la construction de cette vue. Pour la simulation des modèles, nous proposons l'utilisation de langages de modélisation exécutables en privilégiant les standards tels que fUML, BPMN ou encore OCL.

Nous avons ensuite commencé à valider notre démarche sur un cas métier Smart Grid : la gestion d'une flotte de véhicules incluant des véhicules électriques et en particulier l'affectation quotidienne des véhicules aux tournées.

Nos travaux portent maintenant sur la co-simulation des différentes vues tout en maintenant leur cohérence en mettant en œuvre le Model Typing et les transformations de modèle. Nous projetons aussi de généraliser notre démarche en intégrant la vue technique afin de construire une simulation complète du SI.

Bibliographie

- Authosserre-Cavarero A., Bertrand F., Blay-Fornarino M., Collet P., Dubois H., Ducasse S. *et al.* (2012). *Interopérabilité des systèmes d'information: approches dirigées par les modèles.* In Inforsid 2012, p. 11–30.
- Bézivin J. (2005). *On the unification power of models.* Software & Systems Modeling, vol. 4, n° 2, p. 171–188.

- Bézivin J., Blay M., Bouzhegoub M., Estublier J., Favre J.-M., Gérard S. et al. (2004). Rapport de synthèse de l'as cnrs sur le mda (model driven architecture). *CNRS, novembre*.
- Bézivin J., Gerbé O. (2001). Towards a precise definition of the omg/mda framework. In *Automated software engineering, 2001.(ase 2001). proceedings. 16th annual international conference on*, p. 273–280.
- Chenouard R., Granvilliers L., Soto R. (2010). Using atl to define advanced and flexible constraint model transformations. *arXiv preprint arXiv:1002.3078*.
- Chesbrough H., Spohrer J. (2006). A research manifesto for services science. *Communications of the ACM*, vol. 49, n° 7, p. 35–40.
- Favre J.-M. (2004). Towards a basic theory to model model driven engineering. In *3rd workshop in software model engineering, wisme*, p. 262–271.
- France R., Rumpe B. (2007). Model-driven development of complex software: A research roadmap. In *2007 future of software engineering*, p. 37–54.
- Jézéquel J.-M., Combemale B., Vojtisek D. (2012). *Ingénierie dirigée par les modèles: des concepts à la pratique...* Ellipses.
- Kleppe A. G. (2007). A language description is more than a metamodel.
- Kruchten P. B. (1995). The 4+ 1 view model of architecture. *Software, IEEE*, vol. 12, n° 6, p. 42–50.
- Lahna B., Roudies O., Giraudin J.-P. (2005). Une approche multivue pour la conception de systèmes d'information à base de composants. In *Inforsid*, p. 45–60.
- Mens T., Van Gorp P. (2006). A taxonomy of model transformation. *Electronic Notes in Theoretical Computer Science*, vol. 152, p. 125–142.
- Monti A., Ponci F. (2010). Power grids of the future: Why smart means complex. In *Complexity in engineering, 2010. compeng'10.*, p. 7–11.
- Nethercote N., Stuckey P. J., Becket R., Brand S., Duck G. J., Tack G. (2007). Mini-zinc: Towards a standard cp modelling language. In *Principles and practice of constraint programming–cp 2007*, p. 529–543. Springer.
- Palensky P., Widl E., Elsheikh A. (2014). Simulating cyber-physical energy systems: challenges, tools and methods. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, n° 3, p. 318–326.
- Raymond K. (1995). Reference model of open distributed processing (rm-odp): Introduction. In *Open distributed processing*, p. 3–14. Springer.
- Reix R., Fallery B., Kalika M., Rowe F. (1995). *Systèmes d'information et management des organisations*. Vuibert.
- Shannon R. E. (1975). Systems simulation.
- Steel J., Jézéquel J.-M. (2007). On model typing. *Software & Systems Modeling*, vol. 6, n° 4, p. 401–413.
- Uslar M., Specht M., Dänekas C., Trefke J., Rohjans S., González J. M. et al. (2012). Standardization in smart grids: Introduction to it-related methodologies, architectures and standards. *Springer Science & Business Media*.

Session Web sémantique et ontologies

AlzFuzzyOnto : Une ontologie floue pour l'aide à la décision dans le domaine de la maladie d'Alzheimer

Firas Zekri¹, Emna Turki², Rafik Bouaziz¹

1. Université de Sfax, Faculté des Sciences Economiques et de Gestion de Sfax, Laboratoire MIR@CL

Route de l'Aéroport km 4, B.P. 1088, 3018 Sfax, Tunisie
firas3112@gmail.com ; raf.bouaziz@fsegs.rnu.tn

2. Université de Sfax, Faculté de médecine de Sfax, Hôpital Hedi Cheker
turki_em@yahoo.fr

RESUME. La lutte contre la Maladie d'Alzheimer (MA) est devenue un enjeu majeur. Nous visons à contribuer à cette lutte en cherchant à fournir des moyens logiciels adéquats pour aider les décideurs du domaine de la MA dans le choix de la décision optimale pour chaque situation donnée. Par ailleurs, il est maintenant reconnu que les ontologies floues sont des outils utiles pour la représentation des connaissances précises et floues et pour le raisonnement sur ces connaissances. Ainsi, nous proposons dans cet article une ontologie floue de domaine, appelée « AlzFuzzyOnto », relative aux concepts spécifiques de la MA. Cette ontologie permet une représentation sémantique des données médicales relatives au diagnostic et à la prise en charge de la MA, tout en prenant en considération les incertitudes et les imprécisions y afférentes. A cet effet, nous avons utilisé l'ontologie **Mind**, comme ontologie noyau de domaine initiale, dans le processus de construction de l'ontologie « AlzFuzzyOnto », que nous avons standardisée pour faciliter l'intégration des bases de règles.

ABSTRACT. The fight against Alzheimer's disease (AD) has become a major issue. We aim to contribute to this fight by seeking to provide adequate software to assist decision makers in the field of AD to choose the optimal decision for each situation. Moreover, it is now recognized that fuzzy ontologies are useful tools for the representation of crisp and fuzzy knowledge and reasoning on it. Thus, we propose in this paper a fuzzy ontology called "AlzFuzzyOnto", related to the AD concepts. This ontology enables semantic representation of medical data for diagnosis and support of AD, while taking into account the uncertainties and inaccuracies associated with this disease. To this end, we used the **Mind** ontology, as initial core ontology, in the building process of the ontology "AlzFuzzyOnto", which we have standardized to facilitate the integration of rule bases.

MOTS-CLÉS : Maladie d'Alzheimer, Ontologie floue, Logique floue, Mind, AlzFuzzyOnto.

KEYWORDS : Alzheimer Disease, Fuzzy ontology, Fuzzy logic, Mind, AlzFuzzyOnto.

1. Introduction

La Maladie d'Alzheimer (MA) est la plus fréquente des maladies neurodégénératives, c'est-à-dire les maladies liées à une détérioration irréversible du fonctionnement des cellules du système nerveux. C'est pourquoi la lutte contre la MA est devenue un enjeu majeur aussi bien dans les pays développés que dans les pays en cours de développement, où le progrès de la médecine a prolongé la durée de vie de la population. La MA affecte 10% de la population âgée (Pleckaityte, 2010) et touche aujourd'hui près de 35,6 millions de personnes dans le monde. Ce nombre doublera d'ici 2030 (Alzheimer Association, 2012). Notons par ailleurs qu'en Tunisie, son coût de traitement varie entre 3000 et 6000 dinars tunisien (entre 1400 et 2843 euros) par an et par personne¹. Alors, l'impact socio-économique de la MA est énorme et les tentatives pour réduire cet impact sont en cours de recherche (Monien *et al.*, 2009). Par conséquent, la détection de la MA et la recherche de techniques informatiques de diagnostics plus fiables et de thérapies plus efficaces constituent un axe de recherche important et récent.

En effet, l'informatique a su apporter une aide appréciable pour la prise de décision dans divers domaines, y compris celui de la médecine. En effet, la limitation des compétences humaines, l'explosion des connaissances médicales et la complexité et les incertitudes de la maladie contribuent à de grandes variations dans la pratique clinique, à des erreurs cliniques et à un mauvais respect des guidelines. Il est tout simplement irréaliste de croire que les soins des maladies (y compris la maladie d'Alzheimer) peuvent continuer à être délivrés sans l'adoption des sciences de l'information. Parmi ces sciences, les ontologies, et tout particulièrement les ontologies floues, sont devenues l'un des domaines prometteurs en raison de leur grande adaptabilité, robustesse et flexibilité, et des capacités de raisonnement qu'elles pourraient fournir aux outils informatiques pour aider à un diagnostic médical proche de la réalité et du raisonnement humain.

Dans cet article, nous définissons une ontologie, appelée « AlzFuzzyOnto », qui permet de représenter et manipuler les connaissances et les données relatives au diagnostic et à la prise en charge de la MA, tout en prenant en considération les aspects flous (imprécisions et incertitudes) qui peuvent être présents dans la démarche clinique de cette maladie.

La suite de l'article est structurée comme suit. La section suivante situe notre positionnement par rapport à l'état de l'art. Nous consacrons la section 3 à présenter la méthode de construction de l'ontologie « AlzFuzzyOnto ». Ensuite, nous présentons dans les sections 4 et 5 les différentes étapes de construction de notre ontologie. Enfin, dans la section 6, nous terminons par une conclusion et des perspectives futures.

1. Ajroudi H., « Pour une stratégie contre la MA », Article publié dans le journal quotidien tunisien « le temps » le 23 septembre 2010.

2. Etat de l'art et positionnement

Une ontologie est définie comme « une spécification explicite d'une conceptualisation » (Gruber, 1993). C'est un outil important pour la modélisation, le partage et la réutilisation des connaissances. Elle permet aux connaissances d'un domaine d'être représentées explicitement par les concepts et les relations entre eux, et donc de les manipuler automatiquement. Compte tenu de son caractère prometteur, elle est utilisée dans différents domaines de recherche en informatique, telles que la gestion des connaissances, l'intégration des données et la recherche d'information. En ce qui concerne la MA, un nombre important d'ontologies ont été présentées dans la littérature durant ces dernières années, touchant ses différents aspects. Nous pouvons classer ces ontologies en trois catégories :

- Ontologies relatives à la MA, destinées à la recherche d'information :

L'ontologie **OntoAD**, proposée dans (Dramé *et al.*, 2014), est une ontologie de maladie, bilingue (anglais-français), qui modélise des connaissances sur la MA et les syndromes apparentés. Elle est construite à partir de ressources textuelles et termino-ontologiques destinées à l'annotation sémantique et la recherche d'information des documents textuels. La méthode de construction combine deux approches : apprentissage d'ontologie à partir des textes et réutilisation des ressources terminologiques existantes. L'ontologie **ADO**, présentée dans (Malhotra *et al.*, 2014), est une ontologie représentant des caractéristiques cliniques, des traitements, des facteurs de risque, et d'autres aspects des connaissances actuelles dans le domaine de la MA. Son principal objectif est de développer un cadre sémantique pour une représentation interopérable et standardisée des connaissances dans le domaine d'Alzheimer. L'ontologie **CADRO** (*Common Alzheimer's Disease Research Ontology*), présentée dans (Refolo *et al.*, 2012), est un système de classification destiné à intégrer et analyser les différents portefeuilles de recherche de la MA à partir des organismes publiques et privés travaillant dans le domaine à travers le monde. Bien que les buts de cette ontologie soient plus larges que les nôtres, notre travail sur la modélisation des tests et de la prise en charge pourrait être inclus dans les catégories B (*Diagnosis, Assessment and Disease Monitoring*) et E (*Care, Support and Health Economics of Alzheimer's Disease*) de l'ontologie CADRO. De plus, en contactant les développeurs de cette ontologie, nous avons appris qu'elle n'est pas encore opérationnalisée.

OntoAD, ADO et CADRO sont, par nature, des ontologies spécialement orientées vers une représentation supérieure de la formalisation des termes, rendant leur utilisation une sorte de problématique dans un scénario du monde réel.

- Ontologies relatives à la MA, destinées à la standardisation des termes :

L'ontologie SNOMED CT (*Systematized Nomenclature of Medicine Clinical Terms*), élaborée par *International Health Terminology Standards Development Organisation* (IHTSDO) en 2007, est une terminologie clinique détaillée qui permet d'explicitier un contenu clinique et une expressivité pour la documentation et

les rapports cliniques. Elle présente le noyau général des terminologies cliniques pour le dossier de santé électronique (DSE) et décrit, d'une manière standard, les différents concepts cliniques, telles que les maladies et les procédures. SNOMED CT est utilisée pour des buts de normalisation, par exemple lors de l'intégration d'une ontologie récemment développée (ontologie de domaine) avec un standard sémantique, la réutilisation de l'ontologie de domaine par des organismes tiers est possible (Toro *et al.*, 2009).

UMLS (Unified Medical System Language) est principalement composé de deux ressources sémantiques : le Meta-thesaurus et le réseau sémantique (Bodenreider, 2004). UMLS Meta-thesaurus est un réseau sémantique contenant plus de deux millions de concepts, construit par l'intégration de 161 terminologies (dans la version 2012AA) et d'ontologies biomédicales (y compris SNOMED CT). Chaque concept Meta-thesaurus est un ensemble de termes synonymes, éventuellement défini dans différentes langues (le Meta-thesaurus est multilingue), avec un identifiant unique (Concept Identificateur - CUI) qui est parfois associé à des définitions textuelles. Ces concepts sont reliés par différents types de relations, dont la plupart sont dérivées de terminologies source. Ces relations sémantiques sont taxonomiques ou non taxonomiques, et certaines d'entre elles sont définies à partir d'un point de vue logique (c'est-à-dire, elles sont étiquetées comme *is_a*, *finding_site_of*, *part_of*, etc.).

Nous avons utilisé les ontologies **SNOMED CT** et **UMLS Meta-thesaurus** comme des standards, où chaque concept utilisé dans notre ontologie, est issu de ces dernières. Cette standardisation va permettre la réutilisation de notre ontologie de domaine par plusieurs organismes tiers, puisque elle va être compatible (en raison des termes standards utilisés) avec la plupart des moteurs d'inférences. Cependant, une liaison avec ces ontologies est un travail intéressant, et nous allons prochainement l'aborder.

- Ontologies relatives à la MA, destinées au stockage des connaissances cliniques et au raisonnement sur ces connaissances :

La seule ontologie trouvée concernant cette catégorie est l'ontologie **Mind** qui est développé dans le cadre du projet **MIND** par Sanchez *et al.* en 2011. Cette ontologie modélise les tests effectués sur les patients (par exemple, neuropsychologiques, neurologiques, radiologiques, métaboliques et génétique) afin de générer des résultats de diagnostics pour la MA. Cependant, elle est limitée seulement au diagnostic et elle ne comprend pas les connaissances relatives à la partie « prise en charge » de la maladie. En effet, la prise en charge constitue un stade important et nécessite une aide à la décision. De plus, et en raison de problèmes de propriété et de confidentialité du projet industriel, les auteurs ont présenté une brève description de l'ontologie **Mind** au lieu de discuter en profondeur de ses détails, ce qui rend difficile de l'utiliser comme une ontologie noyau par les chercheurs. C'est pourquoi nous n'avons utilisé que la version brève publiée de l'ontologie **Mind**, et nous avons essayé de l'étendre en une ontologie

plus vaste. En fait, nous voulons ajouter à cette ontologie un module capable d'assurer la représentation des connaissances relatives à la prise en charge de la MA. Le tableau 1 récapitule toutes les ontologies présentées dans l'état d'art.

Tableau 1. Un tableau récapitulatif comparatif des différentes ontologies

Ontologie	Points positifs	Points négatifs
OntoAD	<ul style="list-style-type: none"> - Ontologie bilingue - Elle combine deux approches : <ul style="list-style-type: none"> • apprentissage d'ontologie à partir des textes • réutilisation des ressources terminologiques existantes 	<ul style="list-style-type: none"> - Ne peuvent pas être utilisées dans un scénario du monde réel
ADO	<ul style="list-style-type: none"> - Elle représente : <ul style="list-style-type: none"> • des caractéristiques cliniques • des traitements • des facteurs de risque 	
CADRO	<ul style="list-style-type: none"> - Destiné à intégrer et analyser les différents portefeuilles de recherche de la MA 	<ul style="list-style-type: none"> - Ne peut pas être utilisée dans un scénario du monde réel - N'est pas encore opérationnalisée
SNOMED CT	<ul style="list-style-type: none"> - C'est une terminologie clinique détaillée - Elle présente le noyau général des terminologies cliniques pour le dossier de santé électronique (DSE) - Elle décrit, d'une manière standard, les différents concepts cliniques, telles que les maladies et les procédures 	<ul style="list-style-type: none"> - Elles sont utilisées seulement pour des buts de normalisation
UMLS	<ul style="list-style-type: none"> - Elle contient plus de deux millions de concepts. 	
MIND	<ul style="list-style-type: none"> - Elle modélise les tests effectués sur les patients - Elle permet de générer des résultats de diagnostics pour la MA 	<ul style="list-style-type: none"> - Elle est limitée seulement au diagnostic - elle ne comprend pas les connaissances relatives à la partie « prise en charge » de la maladie

3. Méthode de construction de l'ontologie « AlzFuzzyOnto »

Les concepts d'une ontologie peuvent appartenir à différents niveaux d'abstraction. Parmi ces niveaux, on peut trouver des ontologies noyaux de domaines et des ontologies de domaines.

- Ontologie noyau de domaine : à un niveau haut, les ontologies « noyaux » de domaines (Gangemi *et al.*, 2004) définissent, pour chaque domaine concerné, un ensemble minimal de concepts génériques et centraux (par exemple, dans le domaine médical : organe, maladie, inflammation).
- Ontologie de domaine : au niveau le plus spécifique, les ontologies noyaux de domaines sont à leur tour affinées pour définir, par spécialisation, des concepts spécifiques de domaines (par exemple, dans le domaine des maladies, maladie cardiaque, maladie pulmonaire).

Comme point de départ, (1) nous avons utilisé la version publiée de l'ontologie **Mind** comme une **ontologie noyau de domaine** initiale qui représente seulement les concepts génériques nécessaires pour le diagnostic de la MA. Deuxièmement, (2) nous avons étendu cette dernière selon deux étapes (cf. figure1) : (i) la première étape a consisté à ajouter des concepts et des relations supplémentaires de haut niveau permettant un diagnostic plus efficace de la MA, y compris la détermination du degré de sévérité de la maladie chez le patient en question et la prise en charge de la MA, ce qui permet une représentation sémantique des différentes thérapies possibles. Nous avons obtenu comme résultat une ontologie noyau de domaine étendue. Puis, (ii) nous avons ajouté des concepts et des relations spécialisant les concepts et les relations déjà définis. Ceci a résulté en une **ontologie de domaine étendue**. La troisième étape a consisté à (3) recenser, à l'aide d'experts du domaine, les points d'incertitude et d'imprécision présents dans chaque concept et chaque relation de l'ontologie pour générer des concepts flous et des relations floues permettant de représenter les informations et les données floues.

Nous sommes donc passés d'une ontologie de domaine étendue à une **ontologie de domaine étendue et floue**. Il faut noter que la collection des termes et des concepts liés à la MA a été généré, premièrement, par la numérisation de diverses sources de connaissances, y compris « Guideline for Alzheimer's Disease Management » (California Workgroup, 2008) et « Maladie d'Alzheimer – Enjeux scientifiques, médicaux et sociétaux » (Expertise collective, 2007) et deuxièmement, par les experts du domaine. De plus, tout concept utilisé dans l'ontologie « AlzFuzzyOnto » a été dégagé des deux ontologies SNOMED CT et UMLS Meta-thesaurus, pour aboutir à une ontologie standard et normalisée. En effet, nous avons vérifié chaque concept, ajouté à notre ontologie, est ce qu'il existe ou non dans ces deux dernières. Sinon, on utilise le terme existant équivalent à ce concept. La figure 1 présente la méthode suivie pour la construction de notre ontologie.

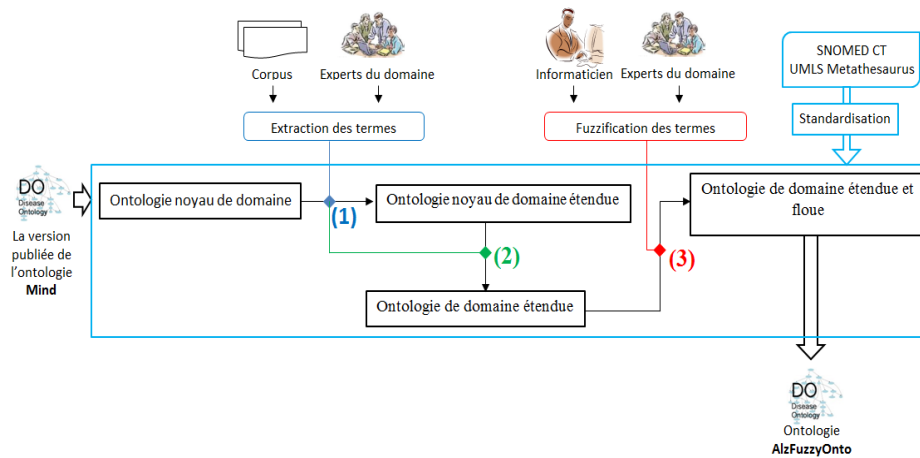


Figure 1. Méthode de construction de l'ontologie « AlzFuzzyOnto »

4. Extension de l'ontologie noyau de domaine Mind (Etape 1)

L'ontologie **Mind** est destinée à collecter et représenter sémantiquement les données cliniques relatives au diagnostic de la MA (Sanchez *et al.*, 2011). En effet, il existe sept classes dans cette ontologie qui sont : *Doctor*, *Patient*, *Diagnosis* (Diagnostic), *Enrollment* (Inscription), *FollowUp* (Suivi), *Test* et *TestValue* (cf. figure 2). La classe *Test* est la superclasse des différents tests appliqués, et tout nouveau test sera considéré comme une sous-classe de cette classe. La classe *Test* est liée à la classe *Patient* à travers la relation *correspondingPatient*, et avec la classe *Doctor* via la relation *orderingDoctor* et en même temps avec la classe *FollowUp* avec la relation *correspondingFollowUp*.

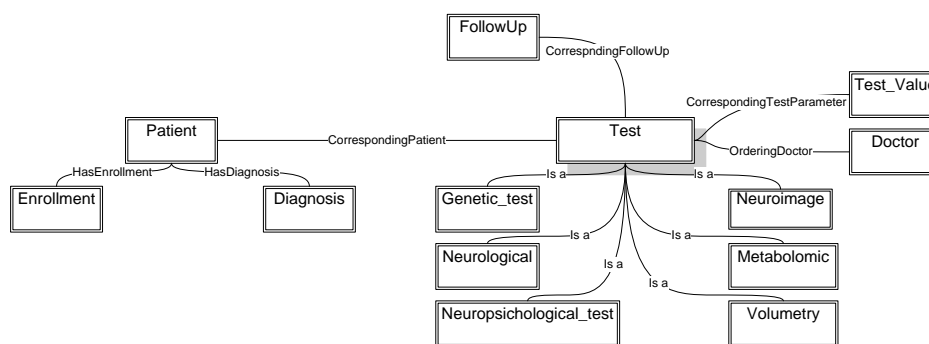


Figure 2. Vue d'ensemble de l'ontologie **Mind** (Sanchez *et al.*, 2011)

Les classes *Diagnosis* et *Enrollment* sont liées à la classe *Patient*, respectivement à travers les relations *hasDiagnosis* et *hasEnrollment*. Nous considérons que la classe

Diagnosis représente seulement le résultat du diagnostic. Les instances de la classe *TestValue* sont les données recueillies dans les interfaces graphiques. Par conséquent, *Test* et *TestValue* sont liés aux propriétés qui font référence à ces paramètres dans les interfaces graphiques. En d'autres termes, ce sont les résultats des différents tests effectués.

Une extension, pour ce qui est présenté dans la version publiée de l'ontologie **Mind**, est requise du fait que nous avons besoin d'intégrer d'autres concepts qui s'avèrent nécessaires pour un diagnostic plus fiable et pour prendre en considération la prise en charge de la MA. Les classes *Questioning* (Interrogatoire), *Symptoms* (Symptômes) et *Entourage* représentent les concepts généraux ajoutés à l'ontologie **Mind** pour ce qui concerne le diagnostic de la MA. Les classes *Therapies* (Thérapies), *Phase* et *Current_state* (Etat actuel) représentent les concepts généraux ajoutés pour ce qui concerne la prise en charge de la MA. *Questioning* est la superclasse des différentes données anamnestiques relatives à chaque patient enregistré (les données épidémiologiques, les antécédents familiaux de démence, l'âge d'apparition des premiers troubles cognitifs, etc.). Elle est liée à la classe *Patient* à travers la relation *correspondingQuestioning*. *Symptoms* est la superclasse des différents types symptomatiques que le patient peut avoir durant les différents stages de la MA (Dépressions, Apathie, Hallucinations, etc.). Elle est liée à la classe *Patient* à travers la relation *HasSymptoms*. *Entourage* est la classe qui représente les différents types des aidants du patient en question. Les connaissances relatives à ces personnes d'entourage aident à bien apprécier la situation de la personne malade. Cette classe est liée à la classe *Patient* à travers la relation *Caregivers(Aidants)*. *Therapies* est la superclasse de toutes les procédures thérapeutiques possibles pour une personne infectée par la MA. Elle liée à la classe *Patient* à travers la relation *HasTherapies*. La classe *Phase* représente le degré de sévérité de la MA du patient en question. Elle est liée à la classe *Patient* à travers la relation *In*. Enfin, la classe *Current_State* représente l'état actuel du patient et elle est liée à la classe *Patient* à travers la relation *HasCurrentState*. La figure 3 représente le schéma ontologique résultant de l'intégration des classes ci-dessus dans l'ontologie **Mind**.

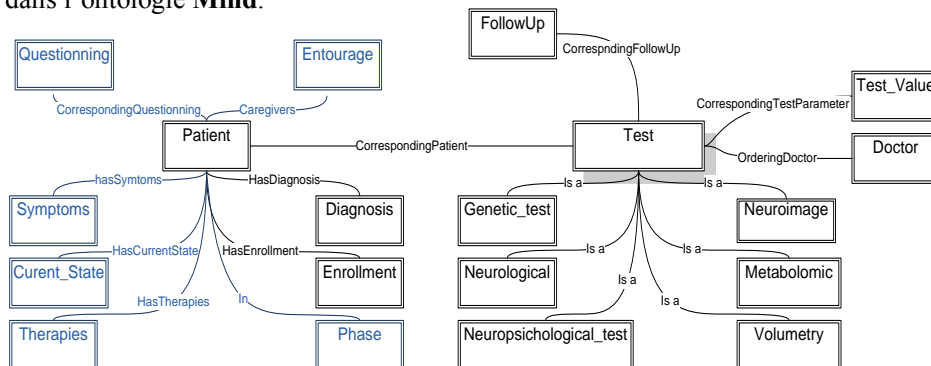


Figure 3. Schéma ontologique de la nouvelle ontologie noyau de domaine

5. Ontologie de domaine pour la MA

5.1 Ontologie de domaine précise (Etape 2)

Pour générer une ontologie de domaine pour la MA, nous avons suivi deux étapes. La première étape consiste à dégager les concepts de niveau hiérarchique plus bas déclinant les concepts généraux présentés dans l'ontologie **Mind**. Par exemple, nous avons intégré toutes les classes possibles relatives à la classe *Neuropsychological* (Neuropsychologique) qui représentent l'ensemble des tests neuropsychologiques effectués sur le patient en question. La deuxième étape consiste à dégager les concepts spécialisant les nouveaux concepts généraux ajoutés à l'ontologie initiale **Mind**. Par exemple, nous avons intégré toutes les classes possibles relatives à la classe générale *Therapies* qui représentent l'ensemble des traitements et des thérapies pharmacologiques et non-pharmacologiques (cf. figure4).

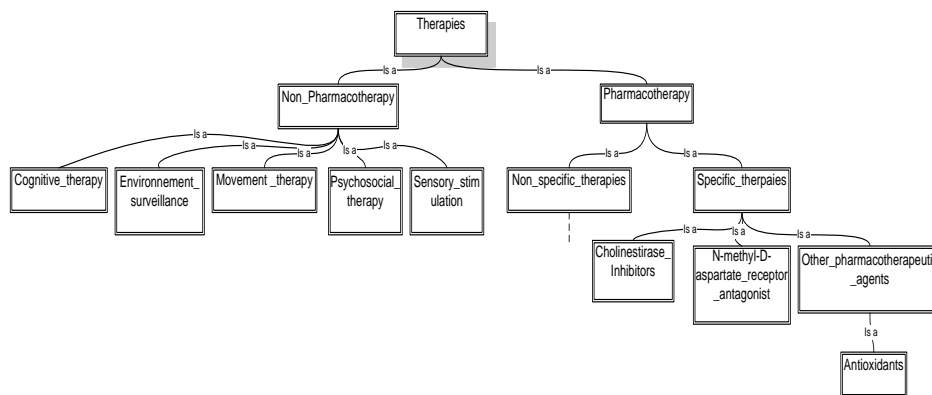


Figure 4. Schéma ontologique du concept « Therapies »

5.2 Fuzzification des termes (Etape 3)

Il existe dans la MA un nombre important de termes (désignant des concepts) qui constituent une source d'imprécision et d'incertitude, que ce soit au niveau de diagnostic ou au niveau de la prise en charge. En effet, rien n'est certain dans la MA. Pour remédier à ce problème, nous avons ajouté des classes floues et nous avons lié ces dernières aux concepts déjà définis en utilisant la relation floue de fuzzification. Une relation de fuzzification est une relation d'héritage entre un concept précis et des concepts flous. Cette relation de fuzzification permet de transformer au moins une propriété sous forme d'une variable linguistique en précisant sa fonction d'appartenance et ses attributs (Ghorbel *et al.*, 2011). Nous avons fuzzifié la classe *Patient* selon l'âge (Adulte, âgé, très âgé), la classe

Pharmacotherapy selon la quantité des doses prescrites (petite dose, moyenne dose, grande dose), la classe *Non-Pharmacotherapy* selon la durée du traitement non pharmacologique (Court terme, moyen terme, long terme), la classe *symptoms* selon l'intensivité des symptômes (Léger, modéré, sévère), la classe *Current-state* selon la sévérité de l'état du patient (Léger, modéré, sévère) et la classe *Diagnosis* selon le degré de certitude du diagnostic (Non pathologique, peut-être pathologique, pathologique). Concernant les relations, nous avons recensé une relation floue qui est *In*. Elle est pondérée selon le score MMSE (Mini Mental State Examination) du patient en question (In Phase1, In Phase2, In Phase3, In Phase4). Pour réaliser ces fuzzifications, nous avons utilisé les quatre fonctions d'appartenance présentées dans la figure 5 (Zadeh, 1975).

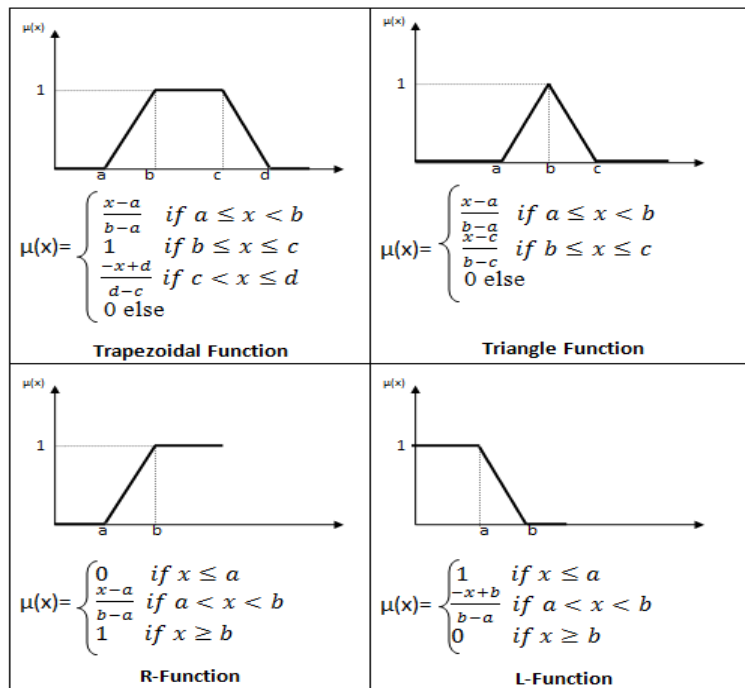


Figure 5. Les fonctions d'appartenance de la logique floue

Dans ce papier nous avons présenté, comme exemple, la fuzzification de la classe Patient et la fuzzification de la relation In. Toutes les valeurs, présentées dans ce qui suit, sont validées par nos experts.

Le concept *Patient* est identifié comme une classe floue selon la variable linguistique « Age ». Il est important de savoir le degré de vieillesse du malade pour un diagnostic et une prise en charge plus fiable. En effet, un patient adulte qui est dans la phase 3 de la MA est un cas très spécifique. Par conséquence, la classe

Patient est fuzzifiée en trois classes floues qui sont : *Elderly-Patient* avec une fonction d'appartenance monotone décroissante qui a comme paramètres $a=55$ et $b=60$, *Old-Patient* avec une fonction d'appartenance trapézoïdale qui a comme paramètres $a=55$ et $b=60$, $c=70$ et $d=75$ et *Very-Old-Patient* avec une fonction d'appartenance monotone croissante qui a comme paramètres $a=70$ et $b=75$. La figure 6 présente les fonctions d'appartenances des trois classes floues dégagées.

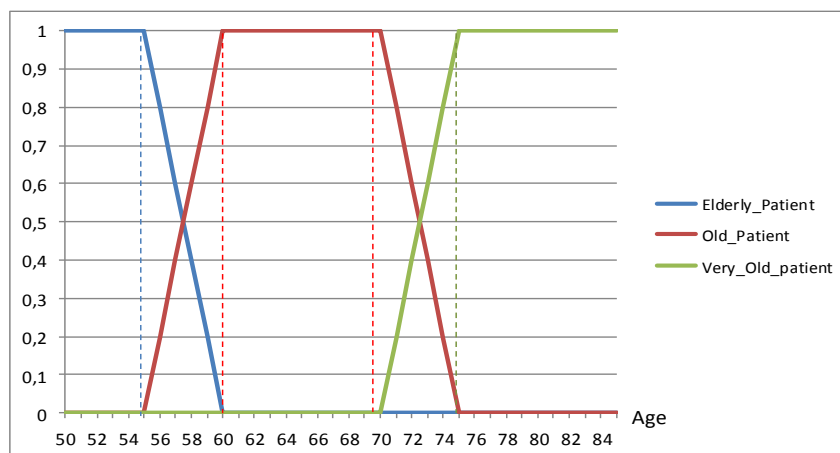


Figure 6. Les fonctions d'appartenances des classes floues du concept « Patient »

Un patient qui a un âge de 73 ans, par exemple, est considéré comme personne âgée (Old-Patient) avec un degré d'appartenance égal à 0,4. Il est aussi considéré comme personne très âgée (Very-Old-Patient) avec un degré d'appartenance égal à 0,6. Il n'est pas considéré comme un patient adulte (Elderly-Patient).

Une relation floue est définie comme un ensemble flou. Ses instances possèdent ainsi des degrés d'appartenance prenant leur valeur dans l'intervalle $[0,1]$. Ces degrés sont calculés conformément aux formules de calcul relatives à la fonction d'appartenance définie pour la relation (Ghorbel *et al.*, 2008 ; Ghorbel *et al.*, 2011). La relation *In* est identifiée comme une relation floue puisqu'il est difficile de déterminer la phase exacte de la MA concernant le patient en question. En effet, les tests effectués sur une personne malade peuvent donner des résultats approximatifs sur le degré de sévérité de la maladie. De plus, un patient peut être entrain de changer d'une phase à une autre, donc il peut appartenir à deux phases en même temps. En fait, il existe de nombreuses classifications des différents stades de la MA. Néanmoins, nous considérons qu'il existe quatre formes de MA en fonction du score MMSE du patient, sachant que le stade de démence varie selon le score MMSE. Ainsi, si on prend l'exemple d'un patient ayant un niveau scolaire supérieur, la démence est en stade léger si le score de MMSE est entre 20 et 26, en stade modéré si le score de MMSE est entre 10 et 19, en stade sévère si le score de

$MMSE \leq 9$ et en stade terminal si le score de MMSE est de valeur très basse. Par conséquence, la relation *In* est fuzzifiée en quatre relations floues où chaque relation a sa propre fonction d'appartenance. La figure 7 présente la fuzzification de la relation *In*.

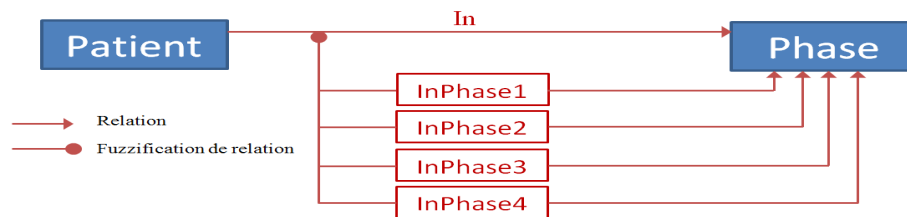


Figure 7. Fuzzification de la relation « In »

La relation *InPhase1* a une fonction d'appartenance monotone croissante qui a comme paramètres $a = 18$ et $b = 22$. La relation *InPhase2* a une fonction d'appartenance trapézoïdale qui a comme paramètres $a = 8$, $b = 12$, $c = 18$ et $d = 22$. La relation *InPhase3* a une fonction d'appartenance trapézoïdale qui a comme paramètres $a = 2$, $b = 4$, $c = 8$ et $d = 12$. Finalement, la relation *InPhase4* a une fonction d'appartenance monotone décroissante qui a comme paramètres $a = 2$ et $b = 4$. La figure 8 présente les différentes fonctions d'appartenance de ces relations.

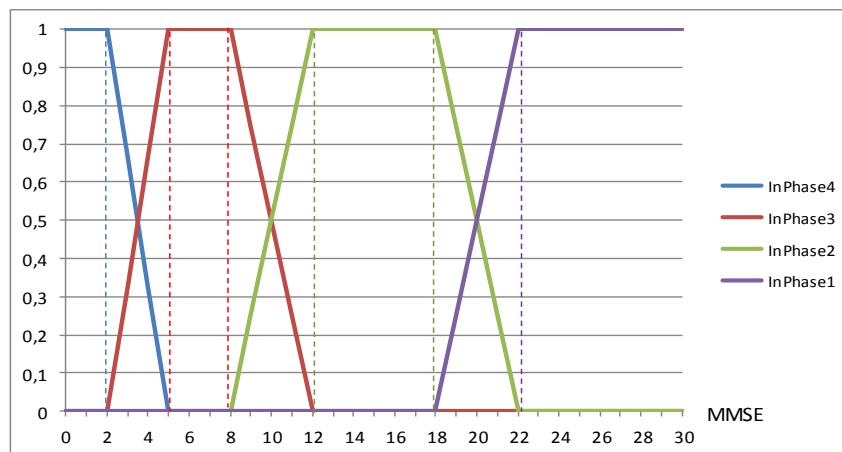


Figure 8. Les fonctions d'appartenance des relations floues

Dans une utilisation future de l'ontologie « AlzFuzzyOnto », chaque instance appartenant à l'une des classes va avoir un degré d'appartenance relatif à la fonction d'appartenance de l'entité à laquelle elle appartient. Toutes les relations « In » doivent être pondérer selon le score MMSE des instances du concept

« Patient » qui utilisent cette relation. Par exemple, l'instance « Patient001 » qui a comme score MMSE égale à 19 doit avoir : (1) une relation « In » avec l'instance « Phase1 » pondérée à $(19 - 18) / (22 - 18) = 0,25$, (2) une relation « In » avec l'instance « Phase2 » pondérée à $(-19 + 22) / (22 - 18) = 0,75$, (3) aucune relation avec l'instance « Phase3 » et l'instance « Phase4 ».

5.3 Opérationnalisation de l'ontologie « AlzFuzzyOnto »

L'ontologie « AlzFuzzyOnto » est implémentée en utilisant l'éditeur d'ontologie *PROTEGE* version 4.3. Les classes et les relations floues, y compris leurs fonctions d'appartenance, ont été intégrées à notre ontologie en utilisant le plugin *Fuzzy OWL2* élaboré par Fernando Bobillo et Umberto Straccia en 2011. En effet, *Fuzzy OWL Tab* dans *PROTEGE* permet de présenter les classes et les relations floues sous forme d'annotations. La figure 9 présente un exemple d'annotation floue « fuzzyLabel ». Le *Datatype* floue « OldAge » est le type de la propriété floue « Old_Age » de la classe floue « Old-Patient ». Elle est annotée avec une fonction d'appartenance trapézoïdale qui a comme paramètres $a = 55$, $b = 60$, $c = 70$ et $d = 75$. Ici, le type de la propriété « Old-Age » n'est pas « integer » mais « OldAge ». C'est-à-dire que chaque instance appartenant à la classe floue Old-Patient a un degré d'appartenance à cette classe en utilisant la fonction définie dans l'annotation.

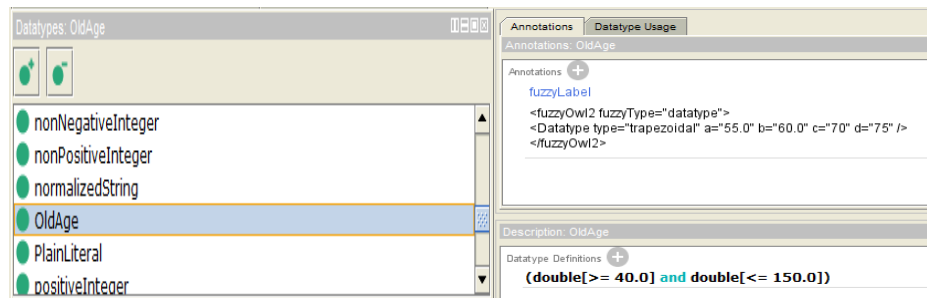


Figure 9. Annotation sur le Datatype « OldAge »

6. Utilité de la logique floue dans l'ontologie « AlzFuzzyOnto »

L'utilisation des technologies sémantiques, et tout particulièrement les ontologies, peut réduire les coûts de la santé: Parce que fournir aux médecins des outils, qui résument les données et les connaissances pertinentes dans une seule page pour une certaine décision, accélère le processus de prise de décision (Sanchez, 2014). À cet effet, notre ontologie floue est intégrée dans un system expert flou d'aide à la décision pour la MA (Zekri *et al.*, 2014). Sa capacité de représenter et stocker des connaissances floues sur la MA a permis de l'intégrer

avec un moteur d'inférence flou. Par conséquence, ce system devient capable d'inférer de nouvelles connaissances et d'aider à la décision en prenant en considération l'incertitude et l'imprécision présentes dans la MA. En effet, notre ontologie joue le rôle d'une base de fait flou dans ce system. Dans ce qui suit, nous allons donner un exemple qui montre l'utilité de la logique floue dans notre ontologie.

L'échelle de Hamilton est le test le plus utilisé pour évaluer l'intensité des symptômes dépressifs (Bentz *et al.*, 2008; Cabañero-Martínez *et al.*, 2007). Plus la note est élevée, plus la dépression est grave :

De 10 à 13: symptômes dépressifs légers

De 14 à 17: symptômes dépressifs modérés

Plus de 18: symptômes dépressifs sévères

Le problème c'est lorsque le résultat du test est égal à 10, 13, 14, 17 ou 18. Ce n'est pas logique que la différence d'un point va distinguer entre non dépression et dépression légère, dépression légère et dépression modérée, ou dépression modérée et dépression sévère. Donc, l'utilisation de la logique floue ici est indispensable. Par conséquence, on a besoin d'une ontologie floue qui supporte cette imprécision.

La classe « Depression Symtom » qui est une sous classe de la classe « Symptômes » est fuzzifier en trois classes floues selon la variable linguistique « Hamilton_Scale » : « Mild Depression Symtom », « Moderate Depression Symtom » et « Severe Depression Symtom ». La figure 10 représente les fonctions d'appartenance des trois classes floues y compris le cas lorsque le patient n'a pas de dépression.

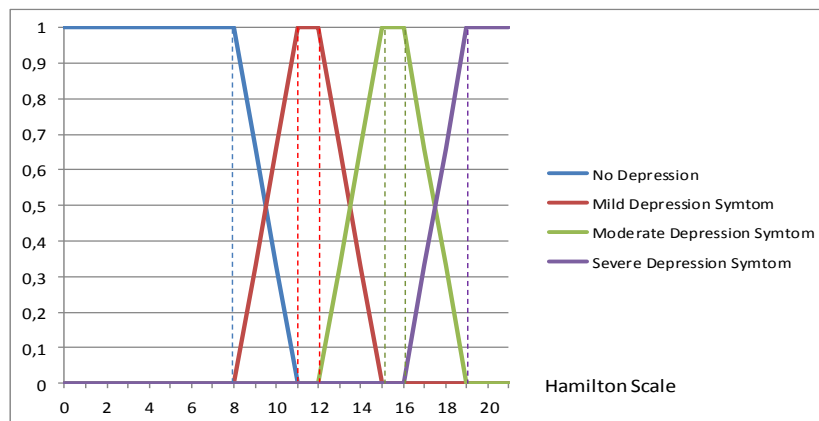


Figure 10. Les fonctions d'appartenance des classes floues du concept « Depression_Symtom »

7. Conclusion

Dans cet article, nous avons montré comment on peut définir une ontologie floue de domaine représentant des connaissances spécifiques à l'aide à la décision pour la maladie d'Alzheimer, en l'occurrence notre ontologie « AlzFuzzyOnto ». Notre contribution a consisté premièrement à étendre l'ontologie noyau de domaine **Mind** pour représenter plus de connaissances. En fait, nous avons pu ajouter à **Mind** un module capable d'assurer la représentation des connaissances relatives à la prise en charge de la MA. Notre deuxième contribution a consisté à montrer comment on peut recenser les points d'incertitude et d'imprécision présents dans certains concepts et certaines relations de l'ontologie pour générer des concepts et des relations flous permettant de représenter les informations et les données floues de la maladie. Ceci nous a permis d'obtenir une ontologie floue pour la MA, proche des scénarios du monde réel de la maladie et facilement intégrable avec des moteurs d'inférence floue.

L'expérimentation et la validation de cette ontologie constituent les premières perspectives de nos travaux sur la MA. Par conséquent, une instanciation de l'ontologie « AlzFuzzyOnto » est nécessaire. Nous comptons définir des individus pour cette ontologie en nous inspirant de cas réels de la MA. Aussi, nous procéderons à la construction (i) d'une base de règles floues contenant des connaissances expertes sur la MA et (ii) d'un moteur d'inférence floue.

Bibliographie

- Alzheimer's Association, (2012) Alzheimer's Disease Facts and Figures. Alzheimer's and Dementia: The Journal of the Alzheimer's Association. vol. 8, n° 2, p. 131-168.
- Bentz B.G., Hall J.R., (2008). Assessment of depression in a geriatric inpatient cohort: A comparison of the BDI and GDS. *International Journal of Clinical and Health Psychology*, vol. 8, p. 93-104.
- Bobillo F., Straccia U., (2011). Fuzzy ontology representation using OWL 2, *International Journal of Approximate Reasoning*, vol. 52, p. 1073–1094.
- Bodenreider O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res*, vol.32, p. 267–70.
- Cabañero-Martínez M.J., Cabrero-García J., Richart-Martínez M., Muñoz-Mendoza C.L., Reig-Ferrer A., (2007). Revisión estructurada de las escalas de depresión en personas mayores. *International Journal of Clinical and Health Psychology*, vol. 7, p. 823-846.
- California Workgroup on Guidelines for Alzheimer's Disease Management, *Guideline for Alzheimer's Disease Management*, Final Report, California Version 2008.
- Dramé K., Diallo G., Delva F., Dartigues J.F., Mouillet E., Salamon R., Mougín F., (2014). Reuse of terminological resources and text corpora for building a multilingual

- domain ontology: An application to Alzheimer's disease, *Journal of Biomedical Informatics* vol. 48, p. 171–182.
- Gangemi A., Borgo S., (2004). (Eds.), In Proceedings of Workshop on Core Ontologies in Ontology Engineering in the 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW).
- Ghorbel H., Bahri A., Bouaziz R., (2008). A Framework for Fuzzy Ontology Models, 2^{ème} Journées Francophones sur les ontologies.
- Ghorbel H., Maalej S., Bahri A., Bouaziz R., (2013). Un framework pour la génération semi-automatique d'ontologies floues, *Technique et Science Informatiques (TSI)*, vol. 32, n° 6, p. 671-698.
- Gruber T. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, In Formal Ontology in Conceptual Analysis and Knowledge Representation, Nicola Guarino and Roberto Poli Edition, Kluwer, Academic Publishers, T.R.
- Expertise collective (2007). *Maladie d'Alzheimer – Enjeux scientifiques, médicaux et sociétaux*, Paris, Editions Inserm.
- Malhotra A., Younesi E., Gundel M., Muller B., Heneka M.T., Hofmann-Apitius M., (2014). ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease, *Alzheimer's & Dementia*, vol. 10, p. 238–246.
- Monien B., Apostolova L., Bitan G., Early diagnostics and therapeutics for Alzheimer's disease. How early can we get there?, *Expert review of Neurotherapeutics*, vol. 6, 2009, p. 1293-1306.
- Pleckaityte M. (2010). Alzheimer's disease: a molecular mechanism, new hypotheses, and therapeutic strategies, *Medicina*, vol. 46 n° 1, p. 70-76.
- Refolo L.M., Snyder H., Liggins C., Ryan L., Silverberg N., Petanceska S., Carrillo M.C., (2012). Common alzheimer's disease research ontology: National Institute on Aging and Alzheimer's Association Collaborative Project, *Alzheimer's and Dementia*, vol. 8, n° 4, p. 372 – 375.
- Sanchez E. (2014). *Semantically Steered Clinical Decision Support Systems*, Thesis Submitted to the department of Computer Science and Artificial Intelligence, University of the Basque Country Donostia - San Sebastian.
- Sanchez E., Toro C., Carrasco E., Bonachela P., Parra C., Bueno G., Guijarro F., (2011). A Knowledge-based Clinical Decision Support System for the Diagnosis of Alzheimer Disease, *IEEE Healthcom*, vol. 13, p. 351-357.
- Toro C., Graña M., Posada J., Vaquero J., Szczerbicki E., (2009). Domain Modeling Based on Engineering Standards, p. 95–102.
- Zadeh L. (1975). The Concept of a Linguistic Variable and its Application to Approximate Reasoning, *International Journal of Information Science*, vol. 4, n° 4, p. 301-357.
- Zekri F., Ghorbel H., Bouaziz R., (2014). A Decision Support System Based on Fuzzy Specialized Rules for the Alzheimer Disease, *11th International Conference on Fuzzy Systems and Knowledge Discovery*, p. 490-496.

Combining Business Intelligence with Semantic Web: Overview and Challenges

Sébastien Laborie¹, Franck Ravat², Jiefu Song², Olivier Teste³

1. LIUPPA – T2I, Université de Pau et des Pays de l'Adour

2 Allée du Parc Montaury 64600 Anglet

sebastien.laborie@iutbayonne.univ-pau.fr

2. IRIT - Université Toulouse I Capitole

2 Rue du Doyen Gabriel Marty F-31042 Toulouse Cedex 09

{ravat|song}@irit.fr

3. IRIT - Université Toulouse II Jean Jaurès

1 Place Georges Brassens F-31703 Blagnac Cedex

teste@irit.fr

ABSTRACT. Under today's highly complex and dynamic business environment, external data (most often issued from web) need to be included in traditional On-Line Analytical Processing (OLAP) analysis so that decision-makers would be well-informed before making effective decision. Including external web data requires knowing the exact semantic meaning in order to use the right information at the right time. Semantic Web (SW) technologies allow semantically annotating data so that we can exchange several descriptions over web data, do reasoning over these descriptions and ensure interoperability between humans and systems. A combination of BI technologies with SW will both enhance BI analysis with web data and allow analyzing SW data through BI tools. In this paper, we firstly introduce basic concepts of the BI and SW domains. Then, we present recent research results using SW to enhance OLAP analysis. At last, we identify challenges requiring future research efforts to achieve a complete incorporation of BI with SW.

KEYWORDS: Semantic Web, Data Warehouse, Multidimensional Analysis

1. Introduction

The domain of Business Intelligence (BI) aims to provide a set of tools, methods and technologies for supporting and facilitating decision making. In the context of BI, a data warehouse is used to collect, organize and store subject-oriented, integrated, time variant and non-volatile data (Inmon, 1996 ; Kimball, 1996). Data from different sources (generally internal databases) are periodically added into data warehouse after being cleaned and transformed into a specific structure with the help of Extract-Transform-Load (ETL) process. Traditional BI tools, such as On-Line Analytical Processing (OLAP), have been successfully applied to large amount of

data coming from internal databases. However, the dynamic nature of today's business activities forces traditional BI to open its gate to external data in order to answer to more heterogeneous and open analysis scenario (Chen, Chiang, et Storey, 2012). As an increasing quantity of semantically annotated data is available over Internet¹, including Semantic Web (SW) information in traditional OLAP analysis process is a promising way to enhance traditional BI analyses (Trujillo et Maté 2012 ; Zorrilla *et al.*, 2012 ; Abelló *et al.*, 2013). For instance, a decision-maker may want a better overview of a product by populating a business report with web-published customers' opinions and markets' information (Berlanga *et al.*, 2014).

Even though BI and SW have been two different research directions over the last decades, recent research results show that the convergence of these two domains is inevitable and beneficial for both sides. BI offers powerful tools for analyzing large amount of web data, while SW data have an important density of valuable information that can be used for enriching business analysis (Thi et Nguyen, 2008 ; Kämpgen et Harth, 2011 ; Zorrilla *et al.*, 2012 ; Etcheverry et R. A. Vaisman, 2012 ; Abelló *et al.*, 2013 ; Ibragimov *et al.*, 2014 ; Aufaure et Chiky, 2014).

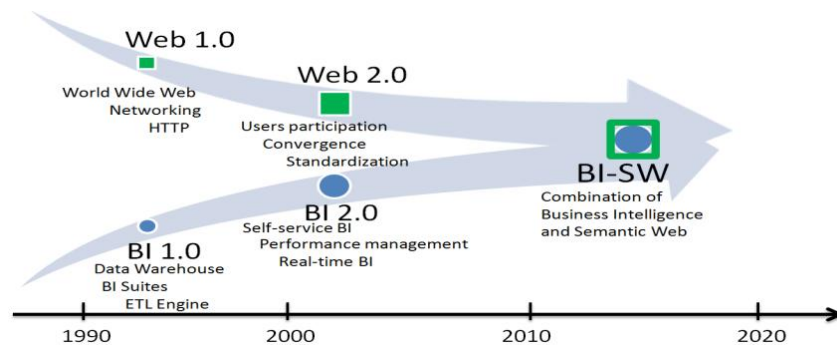


Figure 1. Evolutions in domains of BI and Web

Combining BI with SW, however, is not a trivial task due to the scalability, complexity and heterogeneity of SW data. It raises the following questions: How to integrate heterogeneous SW data in a BI system originally designed for factual data? How to carry out multidimensional analyses over large amount of SW data in the lack of relevant model? How to present analysis results containing both factual data and SW data? These questions are examples of issues waited to be resolved.

The aim of this paper is to present an up-to-date survey of research results and outline future research challenges in BI and SW domains. The rest of the paper is organized as follows. We (i) briefly present the concepts of BI and SW in the section 2; (ii) give an overview of recent research results combining the domain of BI with SW in the sections 3 and 4; (iii) discuss emerging trends and perspectives of future researches in the section 5.

¹ <http://linkeddata.org>

2. Concepts of Business Intelligence and Semantic Web

2.1 *Business Intelligence*

The term of *Business Intelligence* (BI) refers to a set of techniques used for collecting, extracting and analyzing business data to support decision-making process. Coming from heterogeneous and distributed operational sources, data used in decision-making process are stored in *Data Warehouse* after going through a process called *ETL* (standing for Extraction, Transformation and Loading).

Among different types of data warehouse, *On-Line Analytical Processing* (OLAP) data warehouse has been a specific research topic for over a decade. The concepts of OLAP were firstly proposed in (Codd, Codd, et Salley, 1993), they provide solutions for creating, managing, analyzing and reporting large amount of multidimensional data in an interactive way. Among all data models proposed for OLAP, the *Star Schema* (Kimball, 1996) is the most widely accepted model (Chaudhuri, Dayal, et Narasayya, 2011). At conceptual level, *Star Schema* presents data according to subjects of analysis (facts) and axes of analysis (dimensions). At logical level, *Star Schema* can be built on top of different types of databases: *Multidimensional OLAP* (MOLAP), *Relational OLAP* (ROLAP) and *Hybrid OLAP* (HOLAP). At physical level, *Star Schema* can be implemented in different ways, as long as the implementation conforms to the twelve evaluation rules defined in (Codd, Codd, et Salley, 1993), such as multidimensionality, transparency, accessibility, etc. Together with the multidimensional data model, a set of operators is indispensable for OLAP analysis. They permit to aggregate information (Drilldown, Rollup), filter analysis results (Slice, Dice) and change analysis axes (Pivot).

(Kimball, 1998) points out that the main advantages of OLAP model lie in its simplicity and understandability that permit users to interact with large amount of complex data in an efficient way. Nowadays, OLAP is a well-mastered technology when it comes to homogenous and structured data in classical data warehouse. However, as factual data provide only limited and partial views over open-world business scenarios (Zorrilla *et al.*, 2012), the data warehouse community looks for solutions for enriching data collection with external data.

2.2 *Semantic Web*

To accurately exploit web data, a system needs to be capable to read the exact semantic meaning of web-published information. An acknowledged way to publish machine-readable information is to use *Semantic web* (SW) technologies. The purpose of SW technologies is to fix a common vocabulary and a set of interpretation constraints (inferring rules) so as to semantically express metadata over web information and allow doing some reasoning on it. These technologies provide the capability of annotating web data with semantics, e.g., through RDF²

² <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>

and ontologies, hence generating a web of semantic linked data (e.g., Linked Open Data cloud³).

Tim Berners-Lee pointed out four principles that SW data should follow⁴: use *Uniform Resource Identifiers* (URIs) to identify object; use *Hypertext Transfer Protocol* (HTTP) to facilitate searching for objects by human-beings; use the *Resource Description Framework* (RDF)⁵ format as standard to provide descriptive information about an object; link URIs to others in order to connect individual data into a data web. Compared to traditional web technologies which focus mainly on data representation, SW puts a higher value on providing machine-readable information about web resources and relationships between resources.

More specifically, SW presents human knowledge through structured collections of information and sets of inference rules (Berners-Lee, Hendler, et Lassila, 2001). The basic data model is RDF permitting to express simple statements about resources, using named properties and values (cf. figure 2). Resources described by RDF are not necessarily retrievable on the web, they can be anything with a unique identity, from physical objects to abstract concepts (McBride, 2004). A *Triple Store* permits to store RDF data. The set of statements in a RDF Triple Store is composed of URIs, blank nodes and literals. A RDF triple refers to *subject*, *predicate* and *object*: a subject is a web resource identified by a URI or a blank node; an object can be a web resource or a literal that possesses a primitive value; a predicate is a binary relationship connecting a subject with an object. For instance, in the figure 2 we can find the predicate denoted by the label *Concerns* associating the resource *Sales* with another resource *ProductX*, and another predicate named *hasPrice* connecting the subject denoted *ProductX* to a textual literal “30” which is the product’s price.

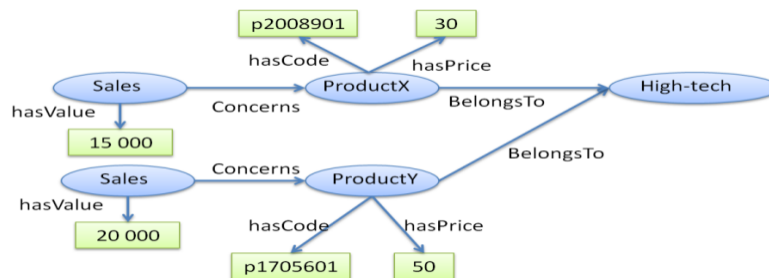


Figure 2. Example of data modeled in RDF format

There exist other SW formats with more powerful expressivity than RDF. Built on top of RDF, RDF Vocabulary Description Language (or RDF schema or RDFS⁶) is a language that defines the terms used in RDF graph. Equivalent to schema

³ <http://lod-cloud.net>

⁴ <http://www.w3.org/DesignIssues/LinkedData.html>

⁵ <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>

⁶ <http://www.w3.org/TR/rdf-schema/>

definition language in relational and object-oriented data model, RDFS is used to describe classes of resources. In other words, RDFS is a simple ontology definition language which allows expressing taxonomies. The concepts of RDFS are described in form of a set of predefined RDF resources with special meanings. However, the reasoning capacity of RDFS is very limited, only basic inferences about taxonomies are supported (Horrocks, Patel-Schneider, et van Harmelen, 2003). Facing to this issue, the Web Ontology Working Group of W3C develops more powerful ontology languages, such as OWL-Lite, OWL-DL, OWL-Full, which allows defining explicit, formal conceptualizations of domain models. In general, OWL enhances the expressivity of RDF and RDFS schema by adding Description Logic (DL). Hence, OWL is an ontology language with sufficient expressive power which can support efficient reasoning through well-defined syntax and semantics (Antoniou et van Harmelen, 2004).

By using the SW formats, web resources can be enriched with annotations and other markups capturing the semantic metadata of resources. However, not all current technologies are fully compatible with the semantic enrichment. For instance, traditional *Information Retrieval* (IR) technologies cannot directly exploit the annotated semantic meaning of web resources (Finin *et al.*, 2005). On the other hand, new research directions have been proposed to combine traditional research approaches with SW technologies, such as *Semantic Information Retrieval* (Fernández *et al.*, 2011), *Exploratory OLAP* (Abelló *et al.*, 2015) etc. In this paper, we only focus on the emerging research direction which aims at enhancing traditional BI with new SW technologies.

3. Overview of researches combining BI with SW

Nowadays, a large number of researches try to merge OLAP analysis with SW technologies both in data integration and data processing levels. This research direction permits to combine powerful tools and technologies in both domains. But it is not a trivial work mainly due to the reason that follows: OLAP requires a specialized data model to support multidimensional analysis over aggregated values of measurements at different granularity levels. However, SW does not dispose of appropriate model fully satisfying criteria about hierarchical levels proposed by (Codd, Codd, et Salley, 1993). Carrying out OLAP analysis directly over SW data is difficult and inefficient by the lack of suitable data model bridging the gap between SW and OLAP domains. Actually, OLAP is originally conceived for analysis over homogenous and stable warehoused data. With arrival of profusion of schema-less Web information, data become more and more heterogeneous and volatile. By mentioning the volatility of SW data we refer to the quick, unceasing and unpredictable changes in SW data sources. Traditional OLAP technologies are challenged while being applied to analyses over SW data.

Facing to these issues, lots of research efforts have been made to combining OLAP with SW. Two types of approaches can be identified (Figure 3). The first approach is OLAP-analyses oriented, which consists of extracting, transforming and then storing multidimensional SW information in traditional OLAP data warehouses

(§3.1), so that it can be analyzed through existing OLAP tools. The second approach is multidimensional modeling oriented, whose aim is to carry out OLAP analyses directly over RDF-like data modeled in an appropriate multidimensional format (§3.2). At the end of the section, we provide a conclusive table (cf. Table 1) that summarizes all mentioned work.

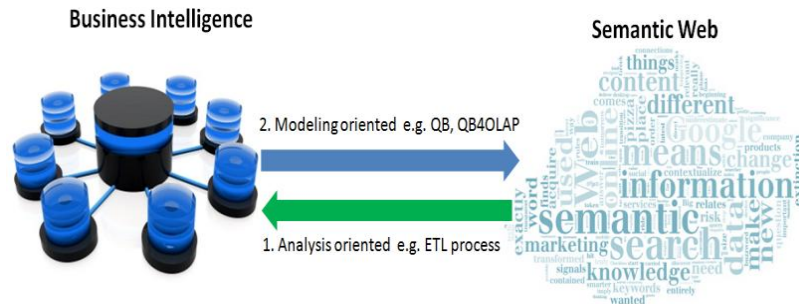


Figure 3. Main approaches to combining BI with SW

3.1 OLAP-analysis oriented approach

OLAP analyses are carried out through analysis operators, such as *roll-up*, *drill-down*, *rotate* and so on (Ravat *et al.*, 2008). Analysis results are usually presented in *Multidimensional Table* (MT) allowing visualizing several analysis axes around a subject. Based on a MT, decision-makers can further carry out OLAP operators to continue their analyses.

OLAP operators are only applicable to specialized data structures (Harinarayan, Rajaraman, et Ullman, 1996 ; Ravat *et al.*, 2008 ; Etcheverry et R. A. Vaisman, 2012), RDF descriptions, however, do not dispose component that can directly support OLAP analysis. For instance, in order to carry out drilldown and rollup operations, we need to represent data according to hierarchical levels within a dimension. However, even though RDF triple can be used to describe web resources and relationships between them (instance level), it does not allow revealing hierarchical relationships within a dimension structure (schema level). Facing to this issue, the OLAP analysis oriented approach consists of transforming SW data into OLAP cube via ETL processes. In this way, OLAP analysis can be carried out over extracted SW data through existing analysis tools. In the following, we will discuss about several works using this approach.

(Romero et Abelló 2007) propose a semi-automatic approach to define an OLAP data warehouse from a single domain ontology. The resulting data warehouse could potentially integrate heterogeneous web sources while following a traditional OLAP data model. This approach enables OLAP analysis to be carried out over extracted SW data. However, valuable information can be found in several domain ontologies in a real-world application. Since the approach proposed in (Romero et Abelló, 2007) is based on a single domain ontology, it does not provide solution for

reconciling overlapping concepts in different domain ontologies. (Nebot *et al.*, 2009) propose a framework to define semi-structured data warehouse from multiple domain ontologies. This data warehouse, called *Semantic Data Warehouse (SDW)*, uses ontology mappings in order to manage domain overlappings. Coherent instances from different domain ontologies are derived and then assembled to semi-automatically generate a targeted OLAP cube.

These works focus on extracting, transforming and loading SW data into OLAP cubes so that decision-makers can directly carry out OLAP analysis. The main advantage is the possibility of reusing existing OLAP tools while analyzing transformed SW data in OLAP cube. However, storing SW data into a relatively static local data warehouse goes against the highly dynamic nature of web-published information. Moreover, the ETL process is not yet totally automatic but quite time-consuming (Romero et Abelló, 2007 ; Nebot *et al.*, 2009 ; Pardillo et Mazon, 2011). From a user's perspective, i.e. requiring high data freshness but not necessarily continuous querying (Pedersen, Castellanos, et Dayal, 2015), semi-automatically or manually built local SW data warehouse can hardly react to changes in data sources in real-time. As a result, the consistency between warehoused data and data in online sources is hard to be maintained. The quality of decision would be low if decision-makers analyze obsolete data in an agile business context.

3.2 *Multidimensional modeling oriented approach*

To overcome the drawbacks of previous approaches, the other research axis consists of carrying out multidimensional analysis directly over SW data without ETL processes. Most of the current frameworks are based on the RDF Data Cube vocabulary (QB), a core vocabulary proposed by W3C aiming to publish statistical and multidimensional datasets in the RDF standard. Directly effectuating OLAP analysis over QB-based model seems to be more efficient because no more ETL process is required. But the principle question is that OLAP analysis requires a complex model of data cubes containing facts, dimensions, multiple hierarchies and levels (Ibragimov *et al.*, 2014). Even though QB allows representing hierarchical relationships between dimension instances via `skos:narrower`⁷, it does not provide mechanism to represent multiple levels on a dimension and the relationships between levels at schema level. (Etcheverry et R. A. Vaisman, 2012).

Facing to this issue, (Kämpgen, O'Riain, et Harth, 2012) define an extension of QB model in order to represent statistical data in a multidimensional model. They illustrate how to carry out OLAP analysis over data published in QB by using the SPARQL⁸ query language. However, their solution does not support dimensions with multiple hierarchies. Consequently, (Etcheverry et A. A. Vaisman, 2012) introduce a new multidimensional modeling language called Open Cube (OC), which supports multiple hierarchies in a dimension. Implementation of OLAP

⁷ <http://www.w3.org/2009/08/skos-reference/skos.html>

⁸ <http://www.w3.org/TR/sparql11-overview/>

operators through SPARQL queries are also presented in this work. However, OC is a specific modeling language, hence data already published in QB (which is standardized), cannot be reused by OC. To overcome this issue, (Etcheverry et R. A. Vaisman, 2012) introduce the QB4OLAP vocabulary. QB4OLAP extends and remains compatible with QB to support multidimensional modeling of SW data. In (Etcheverry, Vaisman, et Zimányi, 2014), an extension of QB4OLAP is proposed. It supports dimension with multiple hierarchies and it takes into account cardinalities between level members. Mechanisms to transform an existent relational data warehouse into QB4OLAP schema have also been presented in (Etcheverry, Vaisman, et Zimányi, 2014). The bi-directional compatibility between QB and QB4OLAP makes querying QB4OLAP with SPARQL possible, but issues about carrying out OLAP analysis in QB4OLAP model rather than simply querying still remains to be discussed. (Saad, Teste, et Trojahn, 2013) propose a conceptual multidimensional model based on QB which supports multi-facts, multi-dimensions and multi-hierarchies with different types (non-covering hierarchy). They also show how to implement OLAP operators via SPARQL queries with the proposed multidimensional model. To the best of our knowledge, (Saad, Teste, et Trojahn, 2013) were the first to address OLAP operators implementation through SPARQL queries in a complete multidimensional data model.

The multidimensional modeling oriented approach overcomes the problems of non-automaticity of ETL process: it provides compatible multidimensional modeling solutions for OLAP analyses over SW data. However, one fundamental principle of BI area, i.e., the materialization of data, is not fully taken into account by this approach. Most of the time, large datasets of SW data are queried on-the-fly, hence the efficiency of OLAP analysis using QB-like model is quite low (Kämpgen et Harth, 2013). Moreover, the quality of datasets varies from one to another; raw SW data without cleansing process may bring false information to decision-makers.

In the following table, we provide a summarized comparison of all listed works belonging to the two approaches.

Table 1. Summarized Comparison

OLAP analysis approach	Advantages : Reuse of existing OLAP technologies and tools				
	Disadvantages : ETL process non-automatic				
	Heterogeneous data sources	OLAP analysis	Multiple ontologies	Automatic cube generation	
(Romero et Abelló, 2007)	√	√			
(Nebot <i>et al.</i> , 2009)	√	√	√		
Multidimensional modeling approach	Advantages : Without need of ETL process				
	Disadvantages : low efficiency of analysis				
	Multiple levels	Multiple hierarchies	Reuse standard	Querying	OLAP operators
QB			N/A	√	
(Kämpgen, O’Riain, et	√		√	√	

Harth, 2012)					
(Etcheverry et A. A. Vaisman, 2012)	√	√		√	√
(Etcheverry et R. A. Vaisman, 2012)	√	√	√		
(Etcheverry, Vaisman, et Zimányi, 2014)	√	√	√	√	
(Saad, Teste, et Trojahn, 2013)	√	√	√	√	√

4. Contextualization of business analysis

Other than being used as data sources for analysis, SW data can also be exploited as complementary information to explain the context of business analysis. For instance, the web-published news talking about steady high temperature in a region could explain the increasing sales of air-conditioners. The combination of external SW data with factual data in an OLAP data warehouse provides decision-makers with multiple views over their business activities. Identifying relevant SW data to contextualize business analysis is a promising way to build decision support systems of the next generation, yet the contextualization of OLAP analysis is achieved mainly through text mining and information retrieval technologies (Perez *et al.*, 2008). As far as we know, no research has fully taken advantage of SW technologies to provide context for analysis. In this section, we briefly present existing techniques for OLAP analysis contextualization, wishing to provide inspiration for future research combining BI with SW.

Contextualization of business analysis can be achieved by retrieving relevant information stored in different systems. (Priebe, 2004) present a prototype permitting to associate relevant documents in content management system with predefined OLAP reports in OLAP system. Through the prototype he envisions different components of an enterprise portal that should share user's context in order to present separately stored but related information together. A formal approach permitting to communicate users' analysis context is presented in (Priebe, 2005). By using mechanisms of meta-searching over heterogeneous metadata, related factual and non factual data can be presented together so as to explain the context of business analysis. The meta-searching is based on metadata enriched with ontological concept mappings. The ontological concept mapping permits to associate the same concept in heterogeneous data sources to the same metadata. This provides a solution for handling the heterogeneity of data in different sources.

The approach proposed by (Priebe, 2005) allows a component of enterprise portal to communicate current user's task with other components, so that all components in a portal could display various information related to a given analysis context. The quality of contextualization mainly depends on information embedded in metadata. However, if decision-makers could freely express their analysis context, the contextualization process would be more flexible and more adaptable to users' needs. To this end, (Manuel Pérez-Martínez *et al.*, 2008) present an architecture of

data warehouse contextualized with documents. By integrating relevant document segments in OLAP cube, this contextualized data warehouse provides decision-makers with information ranked on the basis of relevance to current analysis context. While analyzing, decision-makers can visualize related document segments along with factual data in OLAP cube. The work of (Manuel Pérez-Martínez *et al.*, 2008) differs from (Priebe, 2005) mainly because (Manuel Pérez-Martínez *et al.*, 2008) permit decision-makers to express their own analysis context.

Another way to contextualize business analysis is to retrieve related information on Internet. (Roy *et al.*, 2005) present an approach to associate relevant unstructured data from web with factual data in data warehouse. Firstly, a set of keywords is obtained by exploiting SQL query results. Then, the set of keywords is augmented with more terms retrieved by following the foreign-keys pointers between tables in the data warehouse. At last, the augmented set of keywords is used to retrieve web information via a keyword-based search engine (e.g., Google), so that the analysis context can be explained by the returned search results. This approach is not based on additional semantic information other than factual data in the relational database. Of course, SW techniques (e.g., ontologies) would surely increase keyword retrieval quality. (Liu, Xin, et Alon Y, 2006) propose a mechanism to extract keywords from structured query itself without the need of query execution: instead of obtaining information from query's result, they exploit information embedded in the query. A query is transformed in a set of keywords by removing distractive and unrelated information. The extracted keywords are then used for keyword-based search in a search engine so as to provide analysis context. This is a more generic approach comparing to (Roy *et al.*, 2005), because all types of structured query (SQL query, XML query etc.) are supported by (Liu, Xin, et Alon Y, 2006). Furthermore, in this work we can find further discussion about the benefits of combining keyword extraction with domain knowledge. However, this discussion is very imprecise, a concrete integration strategy of keyword extraction with SW technologies is still missing in this work. What's more, all above-mentioned works are based on traditional IR technologies. We believe new IR research results would certainly improve the efficiency of contextualization process. For instance, *Semantic IR* can be used to exploit semantic meanings embedded in web resources (Fernández *et al.*, 2011). Thus, if the contextualization process has been built on *Semantic IR*, the returned results would be more accurate and more complete.

(Castellanos *et al.*, 2010) and (Castellanos *et al.*, 2012) propose a framework along with a prototype allowing identifying external events in streaming data that would potentially affect the business operations. Based on text-mining techniques, this framework permits to extract and correlate textual information from internal and external data sources. In this way, newly generated web information is constantly associated with related internal information, which provides decision-makers an up-to-date context for their decisions. The following table presents a synthetic view of aforementioned work.

Table 2. Summarized Comparison

	Heterogeneous data sources	Storage of retrieved context	Up-to-date information	OLAP analysis	Ontology based
(Priebe, 2004) and (Priebe, 2005)	√	√			√
(Manuel Pérez-Martínez <i>et al.</i> , 2008)	√	√		√	
(Roy <i>et al.</i> , 2005)	√		√		
(Liu, Xin, et Alon Y, 2006)	√		√		
(Castellanos <i>et al.</i> , 2010) and (Castellanos <i>et al.</i> , 2012)	√	√	√		

5. Future research direction

Various challenges need to be overcome before a complete and efficient combination of BI with SW. For instance, concerning SW data storage (Niinimäki et Niemi, 2009 ; Deliège et Pedersen, 2010 ; Nebot et Berlanga, 2012) and data aggregation reasoning (Calvanese *et al.*, 2008 ; Thorne et Calvanese, 2009). In this section, we mainly focus on two specific issues: data materialization and SW data integration, because few proposals related to these issues are made to fully take advantage of both BI and SW domains.

5.1. Data Materialization

One of the fundamental principles of data warehouse in the BI area is the materialization of data. Researches belong to the approach oriented OLAP analysis consist in a full materialization through ETL process at the price of losing the data freshness. On the other hand, multidimensional modeling oriented approach ignores data materialization: SW data are extracted and queried on-the-fly, which brings about problems in terms of querying efficiency and data quality. To overcome the above-mentioned problems, a promising future research direction consists in partially materializing SW in data warehouse. This partial materialization should be performed at two levels: raw data and aggregated data.

Raw data refer to initial web-published data that are not yet subjected to analysis. At raw data level, not all data but only some relatively stable SW data should be maintained in data warehouse. By mentioning stable data, we refer to read-only or read-mostly data with little change over time, such as country's name for geographical data. For insert heavy datasets, only data in very common analysis path should be materialized. Moreover, only relevant data in large online datasets should be materialized. Avoid warehousing irrelevant data requires a precise and efficient data acquisition process. Extensions of classical ETL technologies should be defined

to include new data acquisition rules. (Dayal *et al.*, 2009) point out that inspirations can be found from *Rule Learning* (Stephen, 1999) and *Hidden Markov Models* (Freitag et McCallum, 2000).

In the context of data warehouse, aggregated data refer to pre-summarized information that aims at accelerating analyses over regularly used data. Traditional OLAP tools already allow materializing aggregated data at different granularity levels. However, with the arrival of SW data in traditional OLAP data cube, the materialization of semantic graph data (e.g., RDF) does not always increase the efficiency of analysis if we follow classical aggregation rules. What's worse, analysis becomes sometimes less efficient in certain conditions with traditional aggregation functions (Kämpgen et Harth, 2013). New aggregation rules and functions need to be defined to support materialization of aggregated graph data in an efficient way. Inspiration can be found within *Query Shortcuts* technologies. More specifically, we can consider the materialized aggregated data as a set of shortcuts between the fact and certain disjunctive hierarchical levels in a graph model. Thus, based on the proposed algorithms in (Dritsou *et al.*, 2011), we can decide which shortcuts should be materialized in order to get the best trade-off between querying efficiency and optimal volume of data storage.

5.2. Automatic integration of SW data in OLAP cube

The common method to deal with unstructured (or less structured) data in OLAP data cube is to create data mappings through ontology. Most existing approaches assume that such ontology is easily built if not provided beforehand. In fact, in many cases finding an appropriate ontology for a specific domain is not a trivial work. On the other hand, building ontology from scratch is extremely complicated and thus not recommended. Therefore, automatically creating mappings between heterogeneous data with and without existing ontology is one of the future research challenges. Solutions for this issue can be found within the SW domain, especially the ones based on ontology alignment (Euzenat, 2013). For instance, a primitive data integration process can be manually defined with the help of semantic annotation and ontology mapping (Skoutas et Simitsis, 2007). This preliminary and manually-defined process could simply the automatic definition of future data integration process both in schema-level and instance-level (Rahm et Bernstein, 2001).

6. Conclusion

This paper provides an up-to-date overview of researches aiming to enhance OLAP analysis in the BI field with SW technologies. We can notice traditional OLAP can hardly deal with data coming from heterogeneous and external sources in open-world analysis scenarios. SW technologies come to rescue as they have been conceived to build semantic spaces over online information so that both humans and machines can get the correct semantic meaning of web published data. Enhancing OLAP analysis with SW technologies is a promising way to include external and heterogeneous information in traditional analysis process.

We discussed recent research results according to these approaches: (a) OLAP-analyses oriented approach which uses ETL process to integrate SW data in traditional OLAP data warehouses; (b) multidimensional modeling oriented approach which aims to define an appropriate multidimensional data model supporting direct OLAP analyses over RDF data collections. We concluded that SW technologies can indeed bring powerful tools to OLAP analysis, and OLAP can be used to efficiently analyze SW data. However, future research efforts are still needed to achieve a complete combination of OLAP with SW.

We envision a new data warehouse approach, which may be contextualized with SW data. This approach provides a promising solution for the restitution of both factual data and SW data during an analysis process. A number of researches have involved the contextualization of business analysis with external information by means of text mining or information retrieval. We believe that SW technologies will surely reinforce the ability of contextualization by providing semantically annotated information over web-published data.

Some directions for future research are outlined to make the best use of the two domains. We believe that fundamental principle of BI, such as data materialization, could improve efficiency and quality of analysis over SW data, while SW technologies, such as semantic annotations and ontology alignments, could provide theoretical and algorithmic basis for data warehouse evolution.

References

- Abelló, Alberto, Jérôme Darmont, Lorena Etcheverry, Matteo Golfarelli, Jose-Norberto Mazón, Felix Naumann, Torben Pedersen, Stefano Bach Rizzi, Juan Trujillo, Panos Vassiliadis, et Gottfried Vossen. 2013. «Fusion Cubes: Towards Self-Service Business Intelligence » *Int. J. Data Warehous. Min.* Vol. 9, n 2, p. 66-88.
- Abelló, Alberto, Oscar Romero, Torben Bach Pedersen, Rafael Berlanga, Victoria Nebot, Maria Jose Aramburu, et Alkis Simitsis. 2015. «Using Semantic Web Technologies for Exploratory OLAP: A Survey » *IEEE Trans. Knowl. Data Eng.* Vol. 27, n 2, p. 571-588.
- Antoniou, Grigoris, et Frank van Harmelen. 2004. «Web Ontology Language: OWL » Dans : *Handb. Ontol.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 67-92.
- Aufaure, Marie-Aude, et Raja Chiky. 2014. «From Business Intelligence to Semantic Data Stream Management » Dans : *Adv. Concept. Model.* Cham : Springer International Publishing, p. 85-93. ISBN : 978-3-319-12255-7, 978-3-319-12256-4.
- Berlanga, Rafael, María José Aramburu, Dolores M. Llidó, et Lisette García-Moya. 2014. «Towards a Semantic Data Infrastructure for Social Business Intelligence » Dans : *New Trends Databases Inf. Syst.* Cham : Springer International Publishing, p. 319-327.
- Berners-Lee, Tim, James Hendler, et Ora Lassila. 2001. «The semantic web » *Sci. Am.* Vol. 284, n 5, p. 28-37.
- Calvanese, Diego, Evgeny Kharlamov, Werner Nutt, et Camilo Thorne. 2008. «Aggregate queries over ontologies » Dans : *Proc. 2nd Int. Workshop Ontol. Inf. Syst. Semantic Web.* New York : ACM Press, p. 97-104. ISBN : 9781605582559.

- Castellanos, Malu, Chetan Gupta, Song Wang, Umeshwar Dayal, et Miguel Durazo. 2012. «A platform for situational awareness in operational BI ». *Decis. Support Syst.* Vol. 52, n^o 4, p. 869-883.
- Castellanos, Malu, Song Wang, Umeshwar Dayal, et Chetan Gupta. 2010. «SIE-OBI: a streaming information extraction platform for operational business intelligence ». Dans : *Proc. ACM SIGMOD Int. Conf. Manag. Data.* Indianapolis, Indiana, USA : ACM Press, p. 1105-1110. ISBN : 9781450300322.
- Chaudhuri, Surajit, Umeshwar Dayal, et Vivek Narasayya. 2011. «An overview of business intelligence technology ». *Commun. ACM.* Vol. 54, n^o 8, p. 88-98.
- Chen, Hsinchun, Roger H. L. Chiang, et Veda C. Storey. 2012. «Business intelligence and analytics: from big data to big impact ». *MIS Q.* Vol. 36, n^o 4, p. 1165-1188.
- Codd, E. F., S. B. Codd, et C. T. Salley. 1993. *Providing OLAP (On-line Analytical Processing) to User-analysts: An IT Mandate.* Codd & Associates,
- Dayal, Umeshwar, Malu Castellanos, Alkis Simitsis, et Kevin Wilkinson. 2009. «Data integration flows for business intelligence ». Dans : *Extending Database Technol. Adv. Database Technol.* New York, NY, USA : ACM Press, p. 1-11.
- Deliège, François, et Torben Bach Pedersen. 2010. «Position list word aligned hybrid: optimizing space and performance for compressed bitmaps ». Dans : *Proc. 13th Int. Conf. Extending Database Technol.* New York : ACM Press, p. 228-239.
- Dritsou, Vicky, Panos Constantopoulos, Antonios Deligiannakis, et Yannis Kotidis. 2011. «Optimizing Query Shortcuts in RDF Databases ». Dans : *Semantic Web Res. Appl.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 77-92.
- Etcheverry, Lorena, et Alejandro A. Vaisman. 2012. «Enhancing OLAP Analysis with Web Cubes ». *Semantic Web Res. Appl.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 469-483.
- Etcheverry, Lorena, Alejandro Vaisman, et Esteban Zimányi. 2014. «Modeling and Querying Data Warehouses on the Semantic Web Using QB4OLAP ». Dans : *Data Warehous. Knowl. Discov.* Cham : Springer, p. 45-56.
- Etcheverry, Lorena, et Ro A Vaisman. 2012. *QB4OLAP: A New Vocabulary for OLAP Cubes on the Semantic Web.*
- Euzenat, Jerome. 2013. *Ontology matching.* 2nd edition. New York : Springer,
- Fernández, Miriam, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, et Enrico Motta. 2011. «Semantically enhanced Information Retrieval: An ontology-based approach ». *Web Semant. Sci. Serv. Agents World Wide Web.* Vol. 9, n^o 4, p. 434-452.
- Finin, T., J. Mayfield, A. Joshi, R.S. Cost, et C. Fink. 2005. «Information Retrieval and the Semantic Web ». Dans : *Proc. 38th Hawaii Int. Conf. Syst. Sci.* IEEE, p. 1-10.
- Freitag, Dayne, et Andrew McCallum. 2000. «Information Extraction with HMM Structures Learned by Stochastic Optimization. ». Dans : *Innovative Applications of Artificial Intelligence Conferences.* Austin, Texas : AAAI Press / The MIT Press, p. 584-589.
- Harinarayan, Venky, Anand Rajaraman, et Jeffrey D. Ullman. 1996. «Implementing data cubes efficiently ». *ACM SIGMOD Rec.* Vol. 25, n^o 2, p. 205-216.

- Horrocks, Ian, Peter F. Patel-Schneider, et Frank van Harmelen. 2003. «From SHIQ and RDF to OWL: the making of a Web Ontology Language ». *Web Semant. Sci. Serv. Agents World Wide Web*. Vol. 1, n^o 1, p. 7-26.
- Ibragimov, Dilshod, Katja Hose, Torben Bach Pedersen, et Esteban Zimányi. 2014. «Towards Exploratory OLAP over Linked Open Data—A Case Study ». Dans : *BRITE*. HangZhou : p. 1-18.
- Inmon, William H. 1996. *Building the data warehouse*. 2nd ed. New York : Wiley Computer Pub, 401 p. ISBN : 0471141615.
- Kämpgen, Benedikt, et Andreas Harth. 2013. «No Size Fits All – Running the Star Schema Benchmark with SPARQL and RDF Aggregate Views ». Dans : *Semantic Web Semant. Big Data*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 290-304.
- Kämpgen, Benedikt, et Andreas Harth. 2011. «Transforming statistical linked data for use in OLAP systems ». ACM Press, p. 33-40. ISBN : 9781450306218.
- Kämpgen, Benedikt, Sean O’Riain, et Andreas Harth. 2012. «Interacting with statistical linked data via olap operations ». Dans : *Int. Workshop Linked APIs Semantic Web LAPIS 2012. 9th Extended Semantic Web Conference*. Heraklion, Greece : Citeseer, p. 36-49.
- Kimball, Ralph, éd. 1998. *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. New York : Wiley.
- Kimball, Ralph. 1996. *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. New York : John Wiley & Sons.
- Liu, Jing, Dong Xin, et Halevy Alon Y. 2006. «Answering Structured Queries on Unstructured Data ». Dans : *Web and Databases (WebDB)*. Chicago, Illinois : p. 20-25.
- Manuel Pérez-Martínez, Juan, Rafael Berlanga-Llavori, María José Aramburu-Cabo, et Torben Bach Pedersen. 2008. «Contextualizing data warehouses with documents ». *Decis. Support Syst.* Vol. 45, n^o 1, p. 77-94.
- McBride, Brian. 2004. «The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS ». Dans : *Handb. Ontol.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 51-65. ISBN : 978-3-662-11957-0.
- Nebot, Victoria, et Rafael Berlanga. 2012. «Building data warehouses with semantic web data ». *Decis. Support Syst.* Vol. 52, n^o 4, p. 853-868.
- Nebot, Victoria, Rafael Berlanga, Juan Manuel Pérez, María José Aramburu, et Torben Bach Pedersen. 2009. «Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses ». Dans : *J. Data Semant. XIII*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 1-36.
- Niinimäki, Marko, et Tapio Niemi. 2009. «An ETL Process for OLAP Using RDF/OWL Ontologies ». Dans : *J. Data Semant. XIII*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 97-119.
- Pardillo, Jesus, et Jose Norberto Mazon. 2011. «Using Ontologies for the Design of Data Warehouses ». *Int. J. Database Manag. Syst.* Vol. 3, n^o 2, p. 73-87.

- Pedersen, Torben Bach, Malu Castellanos, et Umesh Dayal. 2015. «Report on the Seventh International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE 2013) ». *ACM SIGMOD Rec.* Vol. 43, n^o 4, p. 55-58.
- Perez, J.M., R. Berlanga, M.J. Aramburu, et T.B. Pedersen. 2008. «Integrating Data Warehouses with Web Data: A Survey ». *IEEE Trans. Knowl. Data Eng.* Vol. 20, n^o 7, p. 940-955.
- Priebe, Torsten. 2005. «Building Integrative Enterprise Knowledge Portals with Semantic Web Technologies. ». Dans : *Intell. Learn. Infrastruct. Knowl. Intensive Organ. Semantic Web Perspect.* IGI Global, p. 146-188. ISBN : 9781591405030, 9781591405054.
- Priebe, Torsten. 2004. «INWISS–Integrative Enterprise Knowledge Portal ». Dans : *Demonstr. 3rd Int. Semantic Web Conf. ISWC 2004.* Hiroshima, Japan : p. 33-36.
- Rahm, Erhard, et Philip A. Bernstein. 2001. «A survey of approaches to automatic schema matching ». *VLDB J.* Vol. 10, n^o 4, p. 334-350.
- Ravat, Franck, Olivier Teste, Ronan Tournier, et Gilles Zurfluh. 2008. «Algebraic and Graphic Languages for OLAP Manipulations ». *Int. J. Data Warehous. Min.* Vol. 4, n^o 1, p. 17-46.
- Romero, Oscar, et Alberto Abelló. 2007. «Automating multidimensional design from ontologies ». Dans : *Int. Workshop Data Warehous. OLAP.* ACM Press, p. 1-8.
- Roy, Prasan, Mukesh Mohania, Bhuvan Bamba, et Shree Raman. 2005. «Towards automatic association of relevant unstructured content with structured query results ». Dans : *Int. Conf. Inf. Knowl. Manag.* ACM Press, p. 405-412. ISBN : 1595931406.
- Saad, Rafik, Olivier Teste, et Cásia Trojahn. 2013. «OLAP Manipulations on RDF Data following a Constellation Model ». Dans : *Int. Semantic Web Conf. ISWC2013.* Sydney :
- Skoutas, Dimitrios, et Alkis Simitis. 2007. «Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data ». *Int. J. Semantic Web Inf. Syst.* Vol. 3, n^o 4, p. 1-24.
- Stephen, Soderland. 1999. «Learning Information Extraction Rules for Semi-Structured and Free Text ». *Mach. Learn. - Spec. Issue Nat. Lang. Learn.* Vol. 34, n^o 1-3, p. 233 - 272.
- Thi, A, et Binh Thanh Nguyen. 2008. «A Semantic approach towards CWM-based ETL processes ». *Proc. -Semant.* Vol. 8, p. 58-66.
- Thorne, Camilo, et Diego Calvanese. 2009. «Controlled Aggregate Tree Shaped Questions over Ontologies ». Dans : *Flex. Query Answering Syst.* Springer Berlin, p. 394-405.
- Trujillo, Juan, et Alejandro Maté. 2012. «Business Intelligence 2.0: A General Overview ». Dans : *Bus. Intell.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 98-116.
- Zorrilla, Jose-Norberto, Óscar, Irene, Florian Daniel, et Juan Trujillo, éd. 2012. *Business Intelligence Applications and the Web: Models, Systems and Technologies.* IGI Global, ISBN : 9781613500385, 9781613500392.

Leveraging Knowledge from the Linked Open Data Cloud in the task of Reverse Geo-tagging

Victor Charpenay^{1,2}, Előd Egyed-Zsigmond¹, Harald Kosch²

1. *Université de Lyon*

*INSA Lyon, LIRIS, 20 Avenue Albert Einstein
69621 Villeurbanne Cedex, France*

victor.charpenay@insalien.org, elod.egyed-zsigmond@insa-lyon.fr

2. *Universität Passau*

*Innstraße 41
94032 Passau, Germany*

harald.kosch@uni-passau.de

ABSTRACT. Currently, Image Reverse Geo-tagging use probabilistic algorithms that rely on the keywords describing a picture to guess the localization of the depicted scene. However, such algorithms still perform poorly and show clear limitations. Notably, the location estimation only occurs at the landmark level; regions or countries are reduced to their centroid.

In this paper, we address this particular issue by exploring a semantic approach, which identifies geographical entities among the keywords to localize the picture (being a landmark or a country). We leverage the Linked Open Data cloud to find possible entities. The benefits of our approach, as opposed to numerical approaches, include an in-depth study of the “geo-relevance” of an image.

RÉSUMÉ. Actuellement, la géo-localisation d'une image consiste à appliquer des algorithmes probabilistes sur les mots-clés la décrivant pour estimer la position de la scène qu'elle représente. Cependant, de tels algorithmes montrent des limites clairement identifiables. En particulier, l'estimation se fait toujours à l'échelle d'un point, les régions et pays étant réduits à leur barycentre. Dans cet article, nous nous concentrons sur ce problème en explorant une méthode sémantique qui identifie des entités géographique (issues du Linked Open Data) pour localiser une photo (qu'il s'agisse d'un point sur une carte ou un pays). L'avantage d'une telle approche vis-à-vis des méthodes numériques est notamment la possibilité d'étudier la pertinence géographique d'une image.

KEYWORDS: information retrieval, Semantic Web, Linked Data, multimedia.

MOTS-CLÉS : recherche d'informations, Web sémantique, métadonnées, multimedia.

1. Introduction

The number of photos available on the Web is constantly growing, due to our storage, sharing and editing activity on platforms like Flickr¹. Gathering these photos has been made easy by Flickr, whose database of several millions of photos is publicly accessible via an API. However, processing them is a harder task. It has motivated more than two decades of research in the field of Multimedia Information Retrieval (MIR). Reverse Geo-tagging is one of the application fields of MIR.

Reverse geo-tagging consists in estimating the location of a picture without contextual information (such as surrounding text or a caption). For instance, it might be easy to recognize that an image was taken in the Sahara if it depicts red sand dunes and a camel.

The openness of Flickr API has facilitated the creation of large data sets, shared among researchers. MediaEval² is an annual benchmarking initiative where research teams compete against each other on the same data. Since 2010, MediaEval contains a placing task, addressing the problem of reverse geo-tagging. Since 2013, a placeability prediction sub-task was added after performances of the state-of-the-art repeatedly proved to be poor compared to human performances. It is often claimed that many photos are actually not placeable. However, there is another key problem, which is the following: in our Sahara picture, despite its discriminative attributes, it does not illustrate a precise location on Earth with fixed latitude/longitude but a region, which covers no less than eleven countries. A reverse geo-tagging application that only outputs coordinates is then likely to make an error of hundreds of kilometers (given that the Sahara is 1,800 km wide and 4,800 km long).

That is why we explore the use of semantics in this paper, in contrast to existing methods that are mostly numerical and probabilistic, to localize a collection of pictures. With the help of a knowledge base, we hope to be able to give an extended output in the form of a geographical object containing geo-coordinates along with other information, so as to increase reverse geo-tagging expressiveness and investigate further the notion of “placeability” –what we thereafter also call geo-relevance.

Next section summarizes the state-of-the-art in reverse geo-tagging and points out the main limitations that motivated us to take a new semantic-oriented direction. We called the resulting method “geo-semanticification”. Section 3 details our implementation of a geo-semanticifier and section 4 gives the results of some experiments and observations around geo-semanticification. The paper ends with an open conclusion about the added-value of semantics in reverse geo-tagging.

1. <https://www.flickr.com/>

2. <http://multimediaeval.org/>

2. Related Work

Our goal is to merge two very distinct research topics, namely reverse geo-tagging, as part of information retrieval, and knowledge representation and semantics. We review first the state-of-the-art of reverse geo-tagging. As it leads to the conclusion that the most effective method to localize image is to take advantage of the keyword distribution of a collection of human annotated images, we review next existing methods to capture knowledge from those keywords. In the context of semantics, such keyword structure is called “folksonomy”, as it corresponds to an attempt to make a crowd agree on a common classification (or taxonomy) of a resource set in an unstructured fashion.

2.1. Reverse Geo-tagging

Research about geo-localizing images has been made possible by the trend of tagging photos with geo-coordinates, in the first place on Flickr. It first triggered a renewal in MIR, many research projects took the direction of Geo-based Multimedia Retrieval. Among them, one can find a precursory work of reverse geo-tagging (Crandall *et al.*, 2009). From their 35 million images data set crawled from Flickr, they were able to detect popular landmarks by clustering photos. The results of the geo-localization task were promising: for a given city, about 65% of the test photos were given the appropriate landmark among the city’s ten principal landmarks (however, in contrast, reverse geo-tagging aims at localizing arbitrary images all around the globe without restriction).

2.1.1. Content-based methods

As geo-based multimedia retrieval gained interest, a recurring problem was observed: even though the number of geo-tagged photos is constantly growing, most of the photos still lack such data; would it be possible to automatically recover it? This problem was first addressed by researchers working in the field of computer vision, who proposed content-based methods. True reverse geo-tagging started with a project called IM2GPS (Hays, Efros, 2008). The authors exploited Scene Recognition descriptors (color histograms, GIST, tiny images...) to perform location estimation. For a training set of 6 million photos, they reached an accuracy of 25% in a range of about 750 km, which they compare to the size of a small country.

Parallel to IM2GPS, another content-based location estimator was developed, re-using a technique from classical information retrieval: Latent Semantic Analysis (LSA) (Cristani *et al.*, 2008). The article mentions that 40% of their test images was located correctly. However, such results are hardly comparable to others since their data set only contains around 3,000 photos of a finite number of places in south-eastern France. Other works centered on geo-localization with prior assumptions include recognizing famous landmarks on Earth (Li *et al.*, 2009) or a city among a list (Fang *et al.*, 2013).

However, location estimation from the content only quickly proved limited. Crandall *et al.* stated that “at a landmark scale (100 m),” visual content is “very effective in estimating location” while “at a metropolitan scale (100 km), text tags are (...) highly effective for estimating location, but the image features are no longer useful” (Crandall *et al.*, 2009). Overall, if one uses keywords (or tags) associated to an image instead of its content, results are significantly improved. On Flickr, most of the photos are tagged (especially because photographers’s visibility is increased this way). Therefore, tag-based methods currently dominate research around reverse geo-tagging.

2.1.2. Tag-based methods

The first tag-based proposal was made by (Serdyukov *et al.*, 2009). Since then, their method was re-used and extended several times. The underlying framework of their approach is called Language Model. They splitted the world in small regions and set up a probability distribution—the “language” that is spoken—in each region. Then, for an input image, its tag set is compared to that of every region. By integrating geographical specificities in the process, the authors achieved a precision of 29.6% with 100×100 km regions and a precision of 47% for regions three times bigger. The last result is almost two times better than what IM2GPS showed, with region dimensions that are two times smaller.

During the five years following (Serdyukov *et al.*, 2009), the language model has been added pre-processing steps (Joshi *et al.*, 2010) or post-processing refinement (Van Laere *et al.*, 2011; Trevisiol *et al.*, 2013). In an application paper, Joshi *et al.* propose a way to filter out all tags that may be irrelevant to set up the language model (Joshi *et al.*, 2010). To do so, conditional entropy from Information Theory is used: tags that are spread around the globe are not likely to give information about where the photo was taken (that is, their entropy is high with regards to the regions) and thus could be skipped.

Along with trying to improve the language model offline, it is also possible to refine the query online, as post-processing. After a region is identified as the most probable location according to the tag set of the input image, one can either find the closest image in terms of tags instead of computing a medoid (as in (Van Laere *et al.*, 2011)) or re-apply a more precise language model in the selected region (in a divide-and-conquer fashion, as in (Trevisiol *et al.*, 2013)). The latter achieves best results with an accuracy of 50% within 100 km. It has been hardly outperformed until now (without external resources) and can be seen as current state-of-the-art. It results from a participation in the MediaEval Placing Task 2012.

As a language model discards the fact that our resources are photos, it is also possible to apply it to tweets, which is getting more and more popular (Cheng *et al.*, 2010; Hauff, Houben, 2012; Roller *et al.*, 2012), or to short news (O’Hare, Murdock, 2012).

2.1.3. *MediaEval 2013*

MediaEval, as a benchmarking initiative, gives a good overview of the state-of-the-art in reverse geo-tagging and helps identify emerging trends in the field. First, the two best performing participants used a language model (Popescu, 2013; Cao, 2013). The third best participant used a slightly different approach but still based on tag probability distribution (Davies *et al.*, 2013). They all proposed diverse ad-hoc improvements to the process but the most interesting one is maybe the definition of the “geographicity” of a tag, so that photos whose tags have low geographicity are considered as unplaceable (Popescu, 2013). They could then be substituted with similar but placeable photos in the training set, if available. The “geographicity” of a tag equals its probability to appear around its most probable regions from the language model.

The results given by (Popescu, 2013) seem impressive: within 1 km, the estimation accuracy is 43% and within 100 km, it reaches 72%. These are the best results until now³. However, there are several observations showing that the proposed method has strong limitations. The most important one is the following: the introduction of a “geographicity” was not the main reason why it performed efficiently. It is mostly due to a larger training set. Starting from a size of 8.5 million images, they increased it to 90 million. Ten times more images were required for an improvement of about 17% at 100 km. Further increase would mean scalability issues. Furthermore, as the size of the training set gets larger, if the test set remains unchanged, there is an overfitting effect; the probability that very similar pictures exist between training set and test set rises, which weakens the role the language model.

What is more, the “geographicity” filtering, although it contributes to improve the overall accuracy, does not fit tags designating countries. It detects mostly landmarks and cities (an empirical radius of 15 km was used to characterize the proximity with a region). Previous attempts to define “geographicity” already existed in the literature. For instance, (Trevisiol *et al.*, 2013) defined a “geo-relevance” function, based on tag frequency and haversine distance, that suffers the same problem. Its parameters have to be hand-tuned so that it eliminates the right tags.

If we put apart results from (Popescu, 2013) for the reasons mentioned above, other methods perform too poorly to be used at a large scale. Only half of the photos could be localized correctly. The challenging issue is about finding a geo-relevance mechanism that could help determine a priori if a photo can be considered as placeable or not. As we saw, language model in itself cannot discriminate tags, hence the search for an alternative function. Our contribution to the field of reverse geo-tagging is the exploration of a new kind of geo-relevance function, based on semantic information instead of numerical information.

3. Results of MediaEval 2014 are not known at the time of writing.

2.2. *Semantics in Folksonomies*

Flickr and its millions of tagged photos is not only of interest in multimedia information retrieval (then seen as a database of enriched visual content) but also in the study of collaborative systems on the Internet (then seen as a social platform where users share knowledge). It makes available a large “folksonomy”.

Folksonomies, expressing unstructured and consensual knowledge, may be opposed to structured taxonomies and ontologies. The latter are models considered as exhaustive that define relations between given concepts. They may cover a limited domain or intend to be high-level. A taxonomy only defines hierarchical relations and thus is less expressive than an ontology. From this perspective, knowledge can be seen as trees or graphs.

A very popular framework that embodies the idea of graph knowledge is W3C’s Resource Description Framework (RDF), which is at the basis of the Semantic Web⁴. It has given birth to a huge graph of interconnected knowledge bases: the Linked Open Data (LOD) cloud⁵. Among others, the LOD cloud includes knowledge derived from the whole of Wikipedia in a semantic repository called DBpedia (Lehmann *et al.*, 2014).

If the tags attached to an image were associated to a concept defined in the LOD cloud, one could determine if a tag is geographically relevant or not, thanks to an analysis of its semantic neighborhood.

Nonetheless, Flickr and other folksonomies on the Web are not related to the Semantic Web (yet) and are not integrated into the LOD cloud. The need to perform automatically or semi-automatically a semantic association between tags and concepts emerged. It was addressed in the last few years by several research teams, reviewed in (García-Silva *et al.*, 2012). Most methods first disambiguate a tag thanks to lexical resources (e.g. WordNet⁶) or ad-hoc measures. Then, they use a semantic search engine (e.g. Swoogle⁷) or other semantic repositories to retrieve matching Semantic Web entities. The most interesting methods achieve a precision of about 90% in the test cases they published (Angeletou *et al.*, 2008; Tesconi *et al.*, 2008). However, their tests were not performed with tag sets as generic as needed in our current focus, as explained in section 4.

To finish, it is worth mentioning DBpedia Spotlight⁸ as another tool to perform semantic association. Nonetheless, this tool was thought to be used for full text instead of tags only and, as such, it uses methods that slightly differ from those reviewed by García-Silva *et al.*

4. <http://www.w3.org/standards/techs/rdf>

5. <http://linkeddata.org/>

6. <http://wordnet.princeton.edu/>

7. <http://swoogle.umbc.edu/>

8. <http://spotlight.dbpedia.org/>

3. Implementation

With the idea of introducing semantics in reverse geo-tagging in mind, we previously identified that tags in a folksonomy could be “augmented” with a semantic context, referred to as semantic association. We believe that such a process could help define a more expressive geo-relevance function, which is a major issue in reverse geo-tagging.

We designed a system we called geo-semanticization that contains the two following steps:

- Semantic association (as defined by García-Silva *et al.*)
- Geographic knowledge extraction (further discussed in section 3.2)

For the semantic association, we implemented several methods described in (García-Silva *et al.*, 2012) so as to find the best performing one in our context. Next, we evaluated knowledge bases available in the LOD cloud that provide geographic information and selected the most suitable one for our application.

3.1. Semantic Association

The review from García-Silva *et al.* lists eight possible semantic association candidates. Though, they do not all fit our specific criteria, which are the following: we are looking for a method that (1) is fully automatic, given that reverse geo-tagging deals with millions of pictures, (2) does not require any prior restriction about the folksonomy (any topic, any lexical form...), (3) provides a Semantic Web entity as an output, with existing semantic neighborhood. Among those described, two methods involve the tagger or require human intervention, one method only handles verbs and two methods do not rely on fully-defined ontologies. Thus, we only re-implemented the three remaining methods presented by (Angeletou *et al.*, 2008), (García-Silva *et al.*, 2009) and (Tesconi *et al.*, 2008) and compared them.

Those three methods show similar procedures, as identified by (García-Silva *et al.*, 2012). Given an input tag, they first include a context identification phase, where words semantically related to the tag are gathered. Then, a second phase consists in a context-based disambiguation (only needed if the tag is found several meanings). At last, during a semantic identification phase, the meaning found previously is matched with an entity of the Semantic Web. As the input tag is associated to a resource (e.g. a photo) and a tagger, its context is either the remaining tags of the resource or all other tags used by the tagger on other resources.

Because each method has its specificities, we proceeded to some adaptations. First, as Flickr tags are especially noisy, tag cleaning is required. We defined a common tag filtering strategy that aims at combining their respective proposals. The following tags are removed as pre-processing: (1) words known as photography-specific concepts, like `nikon`, `blackwhite` or `iphonography` (list of 26 words); (2) tags contain-

ing digits, since they are most likely camera names (e.g. d700, d80) or technical details (e.g. 50mm); (3) Flickr machine tags⁹.

Moreover, as presented next, we used DBpedia as a knowledge base. Unfortunately, as certain resources used by the authors were not publicly available, we could not perfectly re-implement their work. Notably, (García-Silva *et al.*, 2009) and (Tesconi *et al.*, 2008) use an index of Wikipedia pages that takes into account disambiguation pages and redirect pages (TAGora Sense Repository and Tagpedia, respectively). Although both claim to make it available, none is still accessible on the Web and because we did not have enough resource to re-process all Wikipedia pages, we designed a simplified index. Instead of using the whole text of an article, our index detects disambiguation pages from the single title, thanks to Wikipedia guidelines for titles¹⁰ (for instance, possibly ambiguous titles should contain additional information in parenthesis or after a comma). Of course, our index could not equal performances of original ones and introduces a certain amount of false positives. Redirect pages could not be processed at all.

So as to be able to compare results of the different methods, we discard Semantic Web entities that are not from DBpedia. As a consequence, for Angeletou *et al.*'s method, we did not use Watson semantic search engine¹¹ as they do, we query a DBpedia semantic repository directly on entities' label.

We also implemented two simple baselines as further comparison point. They provide naive disambiguation processes. The first one queries the Wikipedia index and randomly selects one of the results as the most probable meaning for the query tag. The purpose of this baseline was to ensure (or to put in doubt) that a further selection strategy is needed. The second baseline takes into account the context of a tag, that is, its co-occurring tags on one resource (a Flickr image in our case). For each result of the index query, it computes a weight using the TF-IDF formula. This baseline was used to discuss the impact of a context in the results.

3.2. Geographic Knowledge Extraction

Semantic association is not strongly tied to a specific semantic repository and theoretically applies to any LOD knowledge. However, we have to ensure that the concepts we retrieve contain geographic information, if any. For instance, if it designates a city, the most important information is its coordinates or related geographic entities (e.g. landmarks in the city or the country to which it belongs). The relevant geographic knowledge is extracted by performing SPARQL queries on the knowledge base. There exists 31 geographic knowledge bases in the LOD cloud (out of 295) and 19.43% of the total number of atomic statements deal with geography¹². We can men-

9. <https://www.flickr.com/groups/api/discuss/72157594497877875>

10. http://en.wikipedia.org/wiki/Wikipedia:Article_titles#Disambiguation

11. <http://watson.kmi.open.ac.uk/WatsonWUI/>

12. <http://lod-cloud.net/state/>

tion GeoNames¹³, Linked GeoData¹⁴ or Yahoo! GeoPlanet¹⁵. However none of them is exhaustive: GeoNames does not provide a hierarchy between geographical entities, Linked GeoData collection is incomplete for some regions and Yahoo! GeoPlanet does not contain GPS coordinates.

That is why we concentrated on generic knowledge bases, DBpedia being the biggest one. Such repositories, although their structure is sometimes heterogeneous, also provide geographical data. Table 1 illustrates that places (all geo-localized concepts) represent almost 20% of the whole DBpedia data. Places have geo-coordinates, they are typed according to a place ontology¹⁶ (that distinguishes populated places from natural regions, for instance) and may contain additional information like area or related landmarks. Moreover, DBpedia states that it contains GPS coordinates for 1,094,000 concepts, which is more than the number of places. Indeed, some entities are not directly landmarks but are strongly related to a place. For example, one could think of a festival (an event, according to DBpedia classes) or a sculpture that is exposed in a square (which is classified as a work in DBpedia).

Table 1. The most frequent high-level classes in DBpedia and their frequency (source: <http://wiki.dbpedia.org/Ontology39>)

Class	Instances
Resource (overall)	3,220,000
Place	639,000
Person	832,000
Work	372,000
Species	226,000
Organization	209,000

All this made us base our geo-semantifier on DBpedia only. All concepts retrieved from the semantic association are defined within DBpedia. To extract geographical knowledge from a given concept, we simply check its type. If it is of type “Place” (defined in DBpedia’s ontology), we further query its GPS properties, expressed as properties from the WGS 84 ontology¹⁷.

4. Results

Previous section presented our two-steps geo-semantification process. In the remainder of this paper, we illustrate its benefits over numerical methods and discuss its shortcomings.

13. <http://www.geonames.org/>

14. <http://linkedgeodata.org/About>

15. <https://developer.yahoo.com/geo/geoplanet/>

16. <http://mappings.dbpedia.org/server/ontology/classes/>

17. http://www.w3.org/2003/01/geo/wgs84_pos

To perform several tests, we randomly selected 1,000 images from a database of 14 million elements, whose acquisition is described in (Mousselly-Sergieh *et al.*, 2014). All images have unique photographers. In the next section, we selected one tag from each photo (as the evaluation is per tag instead of per image). We took care that the 1,000 tags we then obtain are unique.

4.1. Semantic Association Benchmarking

As mentioned earlier, the survey about semantic association methods highlighted three suitable ones for our geo-semanticifier. We carried out a quick benchmarking to determine which performs best. It is worth noting that we used a simplified Wikipedia index for two of the three methods, which may have worsen some results.

We found that the method proposed by (García-Silva *et al.*, 2009) performed poorly. It only reaches an accuracy of 26.0%, i.e. only one fourth of the tags were associated the right Semantic Web entity, if at least one was available (all tags were manually reviewed). As no result was published by the original authors, we decided to put this method aside and concentrate on the two others: (Angeletou *et al.*, 2008; Tesconi *et al.*, 2008). We found an accuracy of 62.2% and 57.7% respectively. At first sight, the method from Angeletou *et al.* performs better. However, it has a poorer coverage, i.e. less tags could be given a meaning since no concept was found during the disambiguation process. Their respective coverage is 66.0% and 79.0%. With such results, we decide to conduct experiments with both methods.

One could note that the accuracy results we give are far below those claimed by original authors (that both reach at least 90%). We argue that it is mostly due to very different tag sets. In the case of (Angeletou *et al.*, 2008), they find a coverage of 26.6%, which is very low. As their selection method is not clear, they may have applied much stricter criteria than we did to decide whether a tag can be “semanticized” or not, leading to uncomparable results. About (Tesconi *et al.*, 2008), there are several observations that make us think that their test set was “easier” than ours. First, while we have 1,000 taggers for 1,000 tags, they only selected 9 users who contributed a total of 3,500 tags. Moreover, the average number of tags per resource is 3.38 in their case and 5.8 in ours. It let us think that the users they selected had well-behaved tagging habits, thus obtaining a clean test set, contrasting with Flickr recurring noise.

4.2. Geo-semanticification

4.2.1. Location Estimation

We first evaluate the performances of the overall geo-semanticification process. Here, we found that Tesconi *et al.*’s approach clearly outperforms Angeletou *et al.*’s one, as seen in Figure 1a. The reasons of this phenomenon are discussed later, using Figure 1b. Given such results, we used the method from Teconi *et al.* as the default semantic association method in our implementation for comparison purpose with other

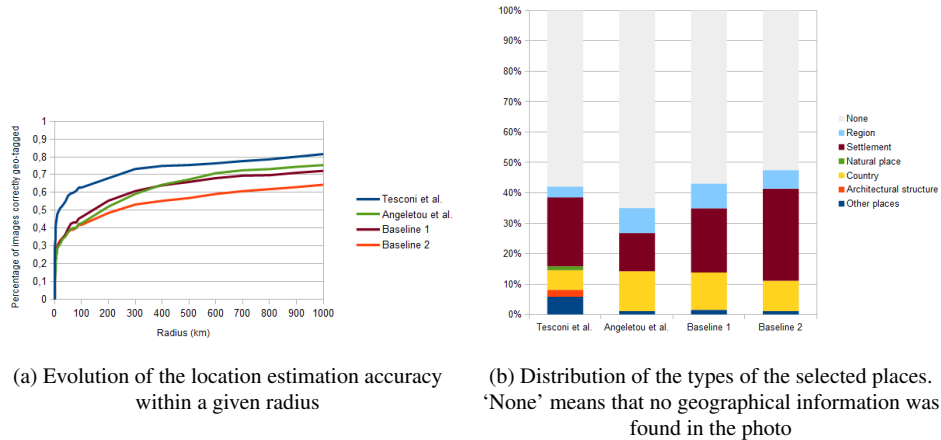


Figure 1. Quantitative evaluation of our geo-semantifier. Both diagrams show results for the two semantic association methods benchmarked in previous section, along with results for the two baselines described in section 3.1

reverse geo-tagging methods. We kept Angeletou *et al.*'s method and our baselines to analyze newly introduced measures.

If we compare the best results achieved by our geo-semantifier (i.e. the blue curve in Figure 1a) with (Popescu, 2013), the latter still performs better: within 1 km, with Tesconi *et al.*'s method, we get an accuracy of 20.1% (while Popescu get 43%) and within 100 km, we obtain 62.9% (72% for Popescu). However, our method outperforms other results of the state-of-the-art, which reach around 50% within 100 km, as in (Trevisiol *et al.*, 2013). Moreover, if we look further the results, we observe that the gap between our method and Popescu's shrinks as error radius grows. Within 1,000 km, accuracies are almost equivalent (81.6% and 83%, respectively). If we refer to a scale introduced by Hays and Efros in IM2GPS that states that cities have an average radius of 25 km, regions of 200 km and countries of 750km, it confirms that the strength of geo-semantification lies in managing localization at the scale of a country as well as at a landmark scale. We already identified previously that Popescu's geographicity measure can hardly detect countries.

What is more, geo-semantification provides information that no numerical reverse geo-tagger could give. For instance, if we take advantage of the area data that comes along with place concepts in DBpedia, we are able to define a "geo-semantic accuracy". It gives a single value that computes the ratio of test images that were localized within a radius defined by their associated DBpedia concept: given a location estimation x of a photo located at x_p , given the area A of the place concept used to estimate the location, the estimation is considered as geo-semantically accurate if $dist(x, x_p) \leq \sqrt{\frac{A}{\pi}}$ (that is, we compare the euclidian distance between the true lo-

cation and its estimation with the radius of the shape of A , approximated as a circle). Errors due to the spherical shape of the Earth are ignored.

The geo-semantic accuracy for our implementation is given in table 2. We can see that, regardless of the scale (landmark or country), it always shows a poor accuracy. Moreover, compared to our two baselines, the performance gain of the non-trivial semantic association is poor (less than 10%). This observation confirms that the main improvements are to be found in the semantic association step and does not depend much on DBpedia. During our experiments, we figured out that one of the advantages of the method proposed by Tesconi *et al.* is that it is partially based on the “popularity” of a DBpedia concept (measured as the size of the corresponding Wikipedia page). For instance, there exists several places called Berlin in the world. However, most of the people ignore the existence of the 20 towns in the United States bearing the same name as the German capital. It is obvious that the popularity of a concept is highly correlated with the most probable meaning of a tag in a folksonomy; this aspect may be further investigated to improve semantic association.

Table 2. Number of location estimations within a radius defined by their associated place concept, used as an error interval. We called this measure geo-semantic accuracy

Method	Geo-sem. accuracy (%)
Tesconi <i>et al.</i>	66.3
Angeletou <i>et al.</i>	60.1
Baseline 1	54.1
Baseline 2	47.8

Along with the geo-semantic accuracy, it is possible to further analyze the results thanks to the DBpedia place ontology. As the latter formalizes several place classes, it is possible to visualize which kind of places help most localize a picture, as shown in Figure 1b. Since, the place types available in the DBpedia ontology are extremely diverse, we only show high-level classes (“Settlement” denotes things like cities, towns or villages). Here, we can put forward that the method of Angeletou *et al.* will more likely select countries and regions than settlements for localization, which leads to a lower overall accuracy, as stated before. The diagram also explains the appearance of the different curves in Figure 1a. For instance, the distance from which Angeletou *et al.* gets better than baseline 1 is roughly 300 km, which corresponds to an intermediary size between a region and a country. Before that point, the baseline detects more precise concepts while afterwards, it accumulates semantic association error (the detected places are wrong). Similar observation could be made to compare the two baselines with each other. Moreover, it is interesting to see that the method of Tesconi *et al.* favors a larger variety of places. This feature could have never been observed with classic reverse geo-tagging systems (we cannot tell if the selected places are correct though).

Figure 1b shows another important information: more than half of the photos cannot be localized in any case. It brings us back to the question of geo-relevance of a photo and placeability prediction, as formulated in MediaEval 2013.

4.2.2. *Placeability Classification*

Our assumption is that a photo can be considered as not geo-relevant if it was not tagged with geo-located concepts. To verify it, we made a simple photo classification: if a photo contains geo-located tags, it has the class “placeable”, otherwise it has the class “not placeable”. Figure 2 shows a sample for each category.

Each sample illustrates the typical kind of image for each class. Placeable photos likely depict precise landmarks that are representative of the place where they were taken. Here, we have six photos representing buildings, whose name was input as a tag by the photographer, like the Bonn Minster (Germany), the oldest church of Malaysia, in Malacca or the city of Marseille (France), embodied by the CMA-CGM tower (the second highest French tower outside Paris). The two first photos (up left) do not represent landmarks. Though, as they symbolize events at known places, they are indirectly placeable. The first one is centered on a poster advertizing (in French) for an exhibition at the city hall of Paris. The second one is an image of a concert that took place at the Metropool concert hall, in Hengelo (Netherlands). The two next photos in the first row are exceptions, though. It is hard to guess that they were taken respectively in San Gabriel mountains (USA) and during the carnival of Barcelona (Spain).

In contrast, not placeable photos, in the sample, are all close-ups or medium shots. They focus on an object (a flower, paintings on a wall, a car, books. . .) or people (here having unclear gestures since the frame is too narrow). They could have been taken anywhere –in the western world– and are geographically irrelevant.

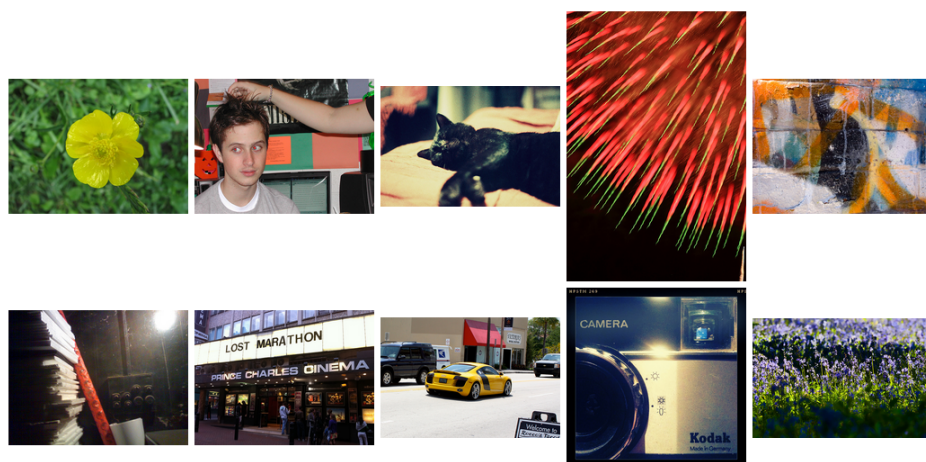
The samples we presented here mostly echo our assumption. However, if one browses further all classified photos, one can notice that the class of placeable photos contains a certain number of false positives. Our geo-semantifier can localize some images even though their location is not clear or even impossible to guess for humans. It means that the photographers tagged them with places even if it was not the main purpose of the picture. That is, the presence of a geo-localized concept in the tags associated to the image is not a sufficient criterion to classify images as placeable. As a future work, we propose to investigate other possible criteria to improve the classification, such as distance measurement in the knowledge graph or machine learning on topological features, derived from the semantic neighborhood of the concepts that tag the photos.

5. Conclusion

In this paper, as we reviewed and pushed forward the use of semantics and the Linked Open Data in the task of reverse geo-tagging, we revealed several elements contributing to a future work.



(a) Placeable photos



(b) Not placeable photos

Figure 2. Samples of placeable and not placeable photos according to our geo-semanticifier. The majority of photos in 2a are buildings, i.e. landmarks, whose name appears in their tags. In 2b, all the photos are close-ups or medium shots focusing on a given object

First, after we selected the best performing approach for semantic association in our study (from Tesconi *et al.*), geo-semanticification was able to outperform the state-of-the-art, reaching a location estimation accuracy of about 63%. Along with competing with classical reverse geo-tagging, geo-semanticification offers new possibilities in terms of geographical relevance. Among others, we defined a geo-semantic accuracy thanks to the knowledge obtained during the process of geo-localizing the image.

Moreover, we began to explore the relationship between image tags and the geo-relevance of an image. Although our classifier was trivial, we were already able to

conclude that if photos that do not contain geographic tags are likely *not* geo-relevant, photos with such tags are not always of interest for localization; Flickr users often tag photos with a city or a country even though it is not related to its content. It made us think of a machine learning problem to build more evolved classifiers.

To finish, the location estimation accuracy we reached was lower than expected, partly due to the fact that the methods we surveyed had never been tested in a real-world application. We are convinced that it is worth investigating further tagging habits in folksonomies, especially the correlation between the probability of a given sense to match a tag and its popularity on Wikipedia.

Acknowledgements

This work was conducted within the frame of a double-degree between the French engineering school INSA and the university of Passau. It was grant-aided by the French region Rhône-Alpes and the Franco-German university (UFA-DFH).

References

- Angeletou S., Sabou M., Motta E. (2008, November). *Semantically enriching folksonomies with FLOR*. Retrieved from [http://oro.open.ac.uk/23497/1/10.1.1.142.7787\[1\].pdf](http://oro.open.ac.uk/23497/1/10.1.1.142.7787[1].pdf)
- Cao J. (2013). Photo Set Refinement and Tag Segmentation in Georeferencing Flickr Photos. In *Working notes proceedings of the mediaeval 2013 multimedia benchmark workshop*, pp. 2–3. CEUR-WS.org.
- Cheng Z., Caverlee J., Lee K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th acm international conference on information and knowledge management - cism '10*, p. 759. New York, New York, USA, ACM Press.
- Crandall D. J., Backstrom L., Huttenlocher D., Kleinberg J. (2009). Mapping the world's photos. *Proceedings of the 18th international conference on World wide web - WWW '09*, pp. 761.
- Cristani M., Perina a., Castellani U., Murino V. (2008, June). Geo-located image analysis using latent representations. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- Davies J., Preston J., Hare J., Jain N., Samangoeei S., Dupplaw D. (2013). Identifying the Geographic Location of an Image with a Multimodal Probability Density Function. In *Working notes proceedings of the mediaeval 2013 multimedia benchmark workshop*, pp. 19–20. CEUR-WS.org.
- Fang Q., Sang J., Xu C. (2013, October). GIANT: geo-informative attributes for location recognition and exploration. In *Proceedings of the 21st acm international conference on multimedia - mm '13*, pp. 13–22. New York, New York, USA, ACM Press.
- García-Silva A., Corcho O., Alani H., Gómez-Pérez A. (2012). Review of the state of the art: Discovering and associating semantics to tags in folksonomies. *The Knowledge Engineering Review*, pp. 1–24.

- García-Silva A., Szomszor M., Alani H., Corcho O. (2009, March). Preliminary Results in Tag Disambiguation using DBpedia. In *Proceedings of the fifth international conference on knowledge capture*. Facultad de Informática (UPM). Retrieved from http://oa.upm.es/6377/1/Preliminary_Results_in_Tag_D.pdf
- Hauff C., Houben G.-J. (2012, August). Placing images on the world map: a microblog-based enrichment approach. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval - sigir '12*, p. 691. New York, New York, USA, ACM Press.
- Hays J., Efros A. A. (2008, June). IM2GPS: estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition*, Vol. 05, pp. 1–8. IEEE.
- Joshi D., Gallagher A., Yu J., Luo J. (2010, July). Inferring photographic location using geotagged web images. *Multimedia Tools and Applications*, Vol. 56, No. 1, pp. 131–153.
- Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P. N. *et al.* (2014). *DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia*. Semantic Web Journal.
- Li Y., Crandall D. J., Huttenlocher D. P. (2009, September). Landmark classification in large-scale image collections. In *2009 IEEE 12th international conference on computer vision*, pp. 1957–1964. IEEE.
- Mousselly-Sergieh H., Döller M., Egyed-Zsigmond E., Watzinger D., Huber B., Kosch H. (2014, March). *World-Wide Scale Geotagged Image Dataset for Automatic Image Annotation and Reverse Geotagging*. In *Multimedia systems conference 2014, mmsys '14*. Retrieved from <http://liris.cnrs.fr/publis/?id=6425>
- O'Hare N., Murdock V. (2012, April). *Modeling locations with social media*. *Information Retrieval*, Vol. 16, No. 1, pp. 30–62.
- Popescu A. (2013). *CEA LIST's Participation at MediaEval 2013 Placing Task*. In *Working notes proceedings of the mediaeval 2013 multimedia benchmark workshop*, pp. 1–2. CEUR-WS.org.
- Roller S., Speriosu M., Rallapalli S., Wing B., Baldrige J. (2012, July). *Supervised text-based geolocation using language models on an adaptive grid*. , pp. 1500–1510.
- Serdyukov P., Murdock V., Zwol R. van. (2009, July). *Placing flickr photos on a map*. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval - sigir '09*, p. 484. New York, New York, USA, ACM Press.
- Tesconi M., Ronzano F., Marchetti A., Minutoli S. (2008). *Semantify del.icio.us: automatically turn your tags into senses*. In *Social data on the web, workshop at the 7th international semantic web conference*.
- Trivisoli M., Jégou H., Delhumeau J., Gravier G. (2013, April). *Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach*. In *Proceedings of the 3rd acm conference on international conference on multimedia retrieval - icmr '13*, p. 1. New York, New York, USA, ACM Press.
- Van Laere O., Schockaert S., Dhoedt B. (2011). *Finding locations of flickr resources using language models and similarity search*. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval - ICMR '11*, pp. 1–8.

Session Modèles et méthodes (2)

Proposition d'un modèle d'innovation par l'usage des SI

Abdelkader ACHI¹, Camille SALINESI²

1. CRI, université Paris 1 Panthéon-Sorbonne
90 rue de Tolbiac, 75013 Paris, France
abdelkader.achi@malix.univ-paris1.fr

2. CRI, université Paris 1 Panthéon-Sorbonne
90 rue de Tolbiac, 75013 Paris, France
Camille.salinesi@univ-paris1.fr

RÉSUMÉ. L'usage intensif et l'évolution des Technologies de l'Information et des Communications (TIC) sont à l'origine de l'innovation de nouveaux produits, services et modèles d'affaires dans l'écosystème. Ainsi, les Systèmes d'Information (SI) et les technologiques émergentes, porteurs d'innovation, positionnent la direction des systèmes d'information (DSI) comme un acteur incontournable dans la démarche d'innovation. L'objectif du présent article est de proposer un modèle d'innovation par l'usage des SI afin d'aider les entreprises à identifier les leviers favorables pour qu'une organisation innove en s'appuyant sur ses SI.

La contribution est un modèle conceptuel élaboré à partir des variables présentant les déterminants de l'innovation par l'usage des SI, et qui a été élaboré au moyen d'une revue de la littérature appuyée par la description de dispositifs mis en œuvre au sein des entreprises.

ABSTRACT. The intensive use and development of Information and Communication Technologies (ICT) are behind the innovation of new products, services and business models in the ecosystem. Thus, the information (SI) systems and emerging technology, innovation carriers, position management information systems (DSI) as a key player in the innovation process. The objective of this article is to propose a model of innovation through the use of SI to help companies identify levers to favor an organization innovates relying on its SI.

The contribution is a conceptual model developed from the variables of the determinants of innovation through the use of SI, which was developed through a literature review supported by the description of the systems implemented within businesses.

MOTS-CLÉS : Innovation par les Systèmes d'Information, modèle d'innovation, leviers d'innovation

KEYWORDS: Innovation by Information Systems, innovation model, innovation levers

1. Introduction

La transformation radicale du monde contemporain ouvre un nouvel âge d'or à l'innovation technologique (Burrus, 2013), et les entreprises se voient envisager une approche à deux vitesses de leur Systèmes d'Information (SI)¹ : d'un côté une approche conventionnelle de l'informatique, en tant que fournisseur traditionnel d'automatisation, et de l'autre un monde numérique, non-linéaire, qui se caractérise par l'incertitude et le besoin d'innover, collaborer et partager des nouveaux savoirs (Gartner, 2014). Cette révolution est liée d'une part à l'usage intensif des Technologies de l'Information et des Communications (TIC), et d'autre part à l'accélération de la diffusion de l'information, la mise en réseau des entreprises et l'émergence de nouveaux espaces de création, de partage des connaissances et d'innovation tels que le crowdsourcing, les communautés, les médias sociaux, etc. En effet, l'innovation naît des fortes interactions entre toutes les composantes de l'écosystème. Ainsi, les récents progrès dans les TIC sont à l'origine des nouveaux modèles d'affaires, transforment les chaînes de valeur, modifient la structure des marchés, et bien entendu sont la source de nouveaux marchés (Delmond et al., 2013; Aubert et al., 2011). L'innovation peut alors prendre des formes diversifiées qu'on lui connaît : produit, procédé, organisationnelle, technologique, commerciale, et modèle d'affaires (OCDE, 2005 ; Aubert et al., 2011), dans lesquelles la connaissance et les TIC jouent un rôle central et complémentaire. De ce fait, les SI porteurs d'innovations, positionnent la DSI en tant qu'acteur clé de l'innovation en entreprise. Par ailleurs, en dehors de cas isolés de succès (Google, Apple, Facebook, Amazon), nous ne disposons pas encore de modèle de bonnes pratiques ni de référentiel pour mener cette démarche d'innovation en s'appuyant sur les SI. Les entreprises cherchent à actionner les bons leviers pour relever les défis d'innovation au sein de l'écosystème mais s'interrogent encore sur la maturité d'une démarche d'innovation ouverte compte tenu des risques.

Les SI porteurs d'innovations, positionnent la direction des systèmes d'information (DSI) en acteur clé dans cette démarche. Les entreprises engagées dans cette voie, et qui souhaitent tirer le meilleur parti de leurs écosystèmes, peuvent alors s'appuyer sur leur SI pour innover et soutenir la performance organisationnelle. Partir du postulat du leadership de la DSI dans les enjeux d'innovation, soulève deux questions : Quelle démarche la DSI doit-elle adopter pour être promoteur de l'innovation par l'usage des SI ? Quelles sont les bonnes pratiques et les risques associés ? Alors que la littérature présente de nombreux travaux relevant de l'ingénierie des SI, et malgré l'importance et l'actualité du sujet, on trouve encore relativement peu de publications sur le rôle des SI en matière d'innovation.

1. Les SI sont définis par Reix (2004) comme « un ensemble organisé de ressources : matériel, logiciel, personnel, données, procédures ... permettant d'acquérir, de traiter, stocker communiquer des informations dans les organisations » (p.75). Une autre définition proposée par Reix et Rowe (2002) : « un système d'information est un ensemble d'acteurs sociaux qui mémorisent et transforment des représentations via des technologies de l'information et des modes opératoires » (p.11).

Afin de répondre à notre question de recherche, nos travaux partent de l'hypothèse fondamentale que l'innovation est fortement liée à l'usage des SI au sein de l'écosystème. Nous proposons de développer notre modèle en deux temps. La première phase réside en une revue de la littérature. Cette revue de la littérature, nous a permis de recenser les variables sous-jacentes à nos questions, ainsi que les dépendances entre ces variables. Des témoignages, cas et descriptions de dispositifs sont aussi collectés au cours de cette analyse. La seconde phase est une phase de validation du modèle ainsi formé à l'issue de la première phase. Le présent article présente les résultats préliminaires de la revue de la littérature. La seconde phase, qui est en cours de réalisation et fera l'objet d'une publication future, résidera en une enquête auprès d'entreprises.

L'objectif de cet article est de proposer un modèle qui formalise l'articulation entre différentes dimensions des SI afin d'innover. Outre qu'il nous permettra d'avoir une meilleure compréhension des leviers de l'innovation par l'usage des SI, un tel modèle peut être la base d'un outil de mesure de la maturité du processus d'innovation ou de développement stratégique dans les entreprises.

La première partie de cet article présente un focus sur l'innovation par l'usage des SI. La deuxième partie détaille les différentes dimensions du modèle d'innovation par l'usage des SI en s'appuyant sur l'analyse des dispositifs mis en œuvre par les entreprises pour répondre aux nouveaux défis d'innovation dans l'écosystème avant de mettre en évidence les travaux connexes.

2. Proposition

2.1. L'innovation par l'usage des SI

Du fait qu'elle repose de plus en plus sur une logique de réseaux et d'interactivité, l'innovation est basée principalement sur les TIC. Ces dernières permettent en effet à l'entreprise de travailler de concert avec toutes les parties prenantes internes et externes pour bénéficier de leurs nouvelles idées et de leur expertise (Chesbrough 2003, 2006a). Grâce à l'évolution des TIC, les organisations peuvent utiliser et intégrer de façon durable, continue et systématique des connaissances à la fois internes et externes pour innover. En effet, les TIC sont une source majeure d'avantages compétitifs (Raymond, 2003 ; Liang et al. 2010). Elles permettent de reformuler les règles de la concurrence, de modifier l'essence même des organisations et favorisent l'innovation en accélérant la diffusion de l'information, le partage des connaissances et la mise en réseau entre les entreprises. Les SI, qui couvrent un champ plus spécifique que les TIC (Kéfi, Kalika 2004) permettent pour leur part de tisser des liens plus étroits entre les entreprises et leurs clients, réduisant ainsi les limites géographiques. Il est évident qu'outre la composante technologique des SI, il y a des aspects liés aux interactions entre les différentes parties prenantes de l'organisation et ses objectifs stratégiques, élément capital dans le processus d'innovation ouverte contemporaine au sein de l'écosystème. En outre, le rôle des SI a évolué. Il est passé d'une fonction

traditionnelle de support des activités à celle d'instrument de base de la stratégie (Reix, 2004). Ainsi, au-delà du rôle "historique" des SI en tant que fournisseur de technologies de l'information par la mise à disposition des outils au service des acteurs, les SI se sont imposés comme source de création de la valeur et d'innovation au sein de l'entreprise.

Les entreprises utilisent ainsi de plus en plus les SI pour lancer de nouveaux produits ou services (Laudon & Laudon, 2010). En conséquence, le métier de la DSI passe progressivement d'un rôle de gestionnaire de techniques à un rôle d'innovateur dans les architectures qui seront le levier de la transformation des organisations (Rochet, 2006). L'objectif d'automatisation des processus et contrôle des coûts a cédé la place à un recentrage sur les métiers afin de créer la valeur par les SI. Cette tendance a été bien caractérisée dans les études menées par le CIGREF (CIGREF-McKinesy, 2008 ; CIGREF-Capgemini Consulting 2009) qui confirment le rôle du leadership que la DSI peut avoir dans l'innovation. En effet, lorsque la fonction SI a atteint un degré de maturité qui lui permette de dépasser le rôle de fonction support et de se positionner en partenaire des métiers, 37% des entreprises intègrent le potentiel des SI dans le processus d'innovation.

Les pratiques des DSI ont évolué, gagnant ainsi la confiance des parties prenantes. Cela leur a permis de développer et d'améliorer les produits et services de l'entreprise dans une approche d'accompagnement des acteurs métiers. Dès lors que la DSI maîtrise de manière irréprochable les fondements techniques, elle a en effet la possibilité de mettre en œuvre les leviers d'action situés à l'interface avec les métiers ; notamment en matière d'innovation par l'usage des SI. Ce phénomène est d'autant plus fréquent que la DSI par sa fonction de veille, est à même de réaliser des combinaisons innovantes : technologies nouvelles / nouvelle organisation des métiers. En outre, les transformations du monde contemporain ouvrent un nouvel âge d'or à l'innovation technologique (Burrus, 2013), et les entreprises se voient envisager une approche à deux vitesses de leur SI : d'un côté une approche conventionnelle de l'informatique, en tant que fournisseur traditionnel d'automatisation, et de l'autre un monde numérique, non-linéaire, qui se caractérise par l'incertitude et le besoin d'innover, collaborer et partager des nouveaux savoirs (Gartner, 2014). De ce fait, les SI porteurs d'innovations, positionnent la DSI en tant qu'acteur clé de l'innovation en entreprise. Enfin, les SI contribuent au processus d'innovation par la fourniture de différentes catégories d'outils logiciels, notamment les systèmes d'aide à la décision, les systèmes de knowledge management et les systèmes collaboratifs (Frenez-Walch et Romon 2006 ; Standage 2005 ; Parent et Chanal, 2009). Ces outils facilitent ainsi la gestion des connaissances, les échanges entre les parties prenantes et la prise de décision. Les outils de conception numérique et de prototypage rapide permettent également des expérimentations, Boland et al (2007) ont ainsi montré l'apport des maquettes numériques qui favorisant l'échange de connaissances des acteurs de métiers différents.

Les SI sont ainsi au cœur des dispositifs mis en œuvre par les entreprises pour répondre aux enjeux business et innover, l'encadré 1 illustre deux innovations de business modèle en s'appuyant sur les SI.

Encadré 1

« Zara : un « business model », renforcé par les systèmes d'information » (Collin et al., 2009)
Le groupe familial Inditex, devenue un des leaders de l'habillement notamment au travers de sa marque Zara, elle a construit un business model unique avec une idée simple : pouvoir répondre en quelques jours aux nouvelles tendances de la mode. Quand les concurrents prennent onze mois pour créer un nouveau produit, Zara prend deux semaines. L'entreprise suit en instantané les ventes de ses produits, les remontées d'information et livre en quelques jours à ses points de vente dans le monde les produits les plus recherchés par les clients. Les systèmes d'information ont ainsi permis de renforcer la configuration originale de la chaîne de valeur de Zara sur deux points essentiels que sont la connaissance immédiate du marché et l'efficacité de son système de supply chain.

« Dell, changement de la chaîne de valeur et la disparition des distributeurs » (Loilier et Tellier, 2013)

La multinationale Dell a mis en place un modèle économique en rupture avec le modèle dominant de l'industrie du matériel informatique : suppression du maillon des distributeurs, customisation par les acheteurs, implantation de la production à proximité des zones de ventes. Dell a innové un business model en s'appuyant sur les SI reposant principalement sur une offre nouvelle, changement de la chaîne et la distribution en direct.

Les démarches d'innovation ont connu une évolution à la hauteur des enjeux économiques. En effet, les nouvelles approches de l'innovation de la société contemporaine sont le fruit de combinaisons des facteurs de l'information liées aux SI et aux interactions de tous les acteurs de l'écosystème. Cette réflexion sur l'innovation par l'usage des SI dans l'écosystème, nous amène à nous intéresser à l'articulation des différentes dimensions liées aux SI et le contexte des organisations afin d'innover.

2.2. Modèle proposé

Au-delà de cas isolés de succès (Google, Apple, Facebook, Amazon), nous ne disposons pas encore de modèle de bonnes pratiques ni de référentiel pour mener cette démarche d'innovation en s'appuyant sur les SI. L'appréciation de la contribution des SI au business et l'innovation ne peut se faire sans tenir compte de toutes les dimensions des SI et celles relatives à son contexte organisationnel d'implémentation. En effet, la DSI réalise des activités de veille (Tambe et al. 2012), de gestion de la connaissance (Austin et al., 2012) en s'appuyant sur le capital humain (OCDE, 2001a, b & c), la culture de l'organisation et son degré d'ouverture (Brown & Eisenhardt, 1997) afin de soutenir le modèle d'affaires de l'entreprise et contribuer à son développement.

Placées en amont, ces différentes dimensions constituent la base de notre modèle d'innovation par l'usage des SI. L'arbitrage entre ces dimensions est déterminant afin de tirer le meilleur profit des SI et d'en faire un moteur au service de l'innovation et la croissance organisationnelle. En effet, cette innovation peut prendre différentes formes qu'on lui connaît : produit, procédé, organisationnelle, technologique, commerciale, et modèle d'affaires (OCDE, 2005 ; Aubert et al., 2011). (Figure 1)

- Organisation du processus d'innovation « modèles d'innovation » : cette dimension matérialise l'organisation du processus et son degré d'ouverture afin de faciliter la synergie entre toutes les composantes de l'écosystème.
- Veille technologique et métier : il s'agit des activités d'écoute de son environnement en vue de mieux s'adapter aux changements.
- Gestion des connaissances : processus de création et partage des connaissances afin d'innover.
- Capital humain : cette dimension fait référence aux connaissances, compétences, expériences, ainsi qu'aux motivations et attitudes des salariés d'une organisation.
- Culture d'innovation : cette dimension regroupe l'identité, les normes, les usages, les habitudes et le système d'interdits et d'obligations partagés par une collectivité ou une population donnée.
- Innovation (produit, procédé, organisationnelle, technologique, commerciale, et modèle d'affaires)

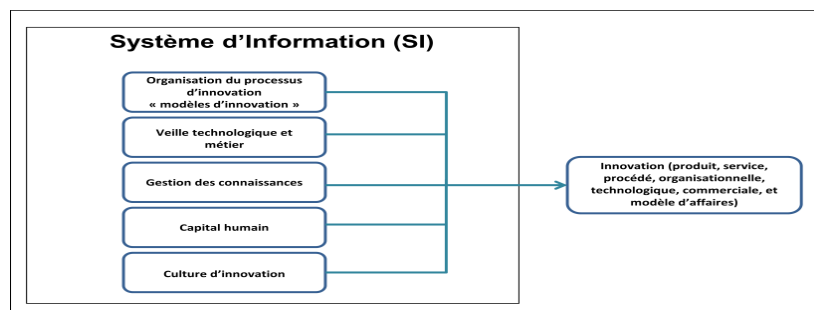


Figure 1: Modèle conceptuel de l'innovation par l'usage des SI

2.3. Les dimensions du modèle d'innovation par l'usage des SI

2.3.1. Dimension organisationnel du processus « modèles d'innovation »

L'innovation a été initialement modélisée par Joseph Schumpeter (1883-1950) comme une « boîte noire ». Dans cette perspective une innovation est définie comme une invention qui a trouvé son marché. Le concept de « boîte noire » désigne un processus linéaire reliant d'une part l'invention, et d'autre part l'innovation. Le processus d'innovation a fait l'objet de deux approches théoriques opposées, le « technology-push » (poussé par l'offre) de Schumpeter (1912, 1942) et la « demand pull » (tiré par la demande) de Schmookler (1966).

Dans la perspective du « technology-push », le processus d'innovation suit une séquence linéaire du « laboratoire jusqu'au marché » sans boucles de rétroaction, ni

retour sur les étapes précédentes du processus. Dans cette perspective les avancées de la science et de la technologie sont le principal stimulus du processus d'innovation, à travers les nouvelles occasions d'inventer qu'elles procurent. Ainsi, l'innovation est conçue dans une logique top-down et linéaire. Les découvertes scientifiques sont diffusées de façon mécanique et sont supposées changer l'ensemble de la société.

Dans l'approche « demand pull » de Schmookler (1966), le processus d'innovation est une résultante de l'attente des usagers, clients, ou autres parties prenantes. Cette approche met en avant l'idée que l'activité inventive est avant tout sensible à satisfaire les demandes et besoins. Cette vision affirme ainsi une sorte de souveraineté du consommateur/utilisateur/usager sans pour autant remettre en cause le caractère linéaire du processus.

A ces deux approches s'ajoute celle de « lead-users » de Hippel (1986). Cette dénomination désigne le rôle des utilisateurs de technologie comme acteurs majeurs de l'innovation. Cette approche, aussi communément qualifiée de « l'innovation par l'usage », fait l'objet d'un consensus nouveau d'après lequel les idées créatives à l'origine des innovations peuvent provenir de tout acteur au sein de l'entreprise comme à l'extérieur, y compris de la part des consommateurs. L'avantage supposé de cette approche est que le caractère risqué voire aléatoire, de l'acceptation de nouveaux produits par le marché l'innovation pourrait diminuer, du fait que l'innovation provient de ceux là même qui constituent le marché. Ce phénomène permettrait ainsi de lancer un nombre plus important de projets, tout en sachant que seuls quelques-uns seront des succès. Cette conception tiendrait compte de l'acceptabilité sociale des innovations du fait que l'innovation est devenue un processus interactif. Il s'agit en effet de prendre en compte les effets de feed-back entre la philosophie de l'offre, poussée par la technologie, et la logique de la demande, propulsée par les consommateurs, ce qui favorise la recherche de praticiens de l'innovation chez tous les acteurs, et la mise en place de démarche nouvelles pour les faire coopérer.

Le modèle « fermé » dans ces différentes versions a été à l'origine de la majorité de produits et services créés durant une grande partie du XXe siècle (Rothwell, 1994), il repose sur l'hypothèse selon laquelle, une organisation dispose en interne des connaissances ou des ressources nécessaires au développement de l'innovation. En revanche, cette démarche n'est plus omniprésente aujourd'hui, et pour de nombreuses organisations le crédo n'est plus celui d'une injection massive de science produite par la recherche qui soutiendrait la production de masse et les gains de productivité, mais plutôt une transformation des métiers et la création de produits et marchés nouveaux au moyen des capitaux disponibles (Rochet, 2006). Cela a ouvert la voie à de nouveaux modèles tels que le « modèle en réseaux » (Powell, Koput, et Smith-Doerr, 1996) ou les modèles « ouverts » d'innovation (Chesbrough, 2003). D'après, Riccio et Bonnet (2012), le management de technologies organisationnelles (articulation entre technologiques et organisation) opère sur la combinaison originale des innovations et contribue de plus en plus à l'ouverture des modèles d'innovation. Ce phénomène d'ouverture est accentué du fait que la technologie soutenant les innovations devient de plus en plus complexe et les

sources de connaissances et d'expertises peuvent être dispersées dans l'écosystème entre de multiples acteurs (entreprises, fournisseurs, partenaires, usagers, etc.). Cette conception est d'autant plus crédible que le développement rapide des TIC a accéléré les capacités d'ingénierie en intensifiant la circulation et le partage de connaissances sources d'innovation. Ainsi, les entreprises ne comptent plus exclusivement sur leurs propres ressources pour répondre aux exigences du monde contemporain des affaires. Les modèles d'innovation ouverte, fruit d'interactions de l'ensemble des acteurs de l'écosystème se sont de fait imposés, en parallèle des modèles traditionnels (linéaire, interactif, etc.). L'heure est donc, à l'innovation collaborative, aux partenariats et à l'ouverture (Chesbrough, 2003). Ce qui a donné lieu à un éventail très large de pratiques basées sur la mise en connexion de l'entreprise avec des détenteurs de ressources à l'extérieur des entreprises : connaissances, compétences, savoir-faire, expériences, expertise, etc.

La transition du modèle ancien « fermé » vers le modèle « ouvert » s'est accompagnée par le passage d'une connaissance contrôlée à l'intérieur de l'entreprise à une connaissance distribuée et détenue par de multiples acteurs de l'écosystème. Cette évolution a été poussée par les avancées technologiques en particulier Internet. Cela a donné lieu à un éventail très large de pratiques basées sur la mise en connexion de l'entreprise avec des détenteurs de ressources à l'extérieur de ces entreprises : connaissances, compétences, savoir-faire, expériences, expertise, etc. D'où l'intérêt de s'interroger sur le ou les modes d'organisation du processus d'innovation à adopter afin de répondre aux enjeux business et profiter de toutes les ressources offertes par l'écosystème en s'appuyant sur les outils offertes par les SI.

2.3.2. Dimension veille technologie et métier

Dans un contexte concurrentiel et en forte évolution technologique, la capacité d'une entreprise à disposer d'informations sur l'évolution de son environnement afin d'adapter sa stratégie, son organisation et préparer le futur devient vital (Duncan, 1972; Subramanian, Fernandes & Harper, 1993; Borgatti et Cross, 2003). Cependant, cela ne se limite pas à la mise en place d'un système de collecte d'informations, mais l'entreprise doit disposer d'un véritable système de veille. Ainsi, la veille se définit comme un processus volontariste et organisé par lequel une entreprise écoute son environnement en vue de mieux s'adapter aux changements de celui-ci (Koenig, 1996, Choo, 1999, 2002). De ce fait, l'écoute permanente de l'environnement permet une lecture rationnelle et anticipative des événements liés aux activités de l'entreprise. En effet, cette pratique réponds aux différentes attentes d'information des décideurs pour leur permettre de s'adapter aux évolutions de l'environnement externe de leur entreprise et réduire l'incertitude (Thiétart, 1990; Lesca 1997), les pratiques de veille soutenues par les SI représentent ainsi une source d'innovation (Tambe et al. 2012). L'association entre veille-innovation/veille-crédation, conduit souvent les organisations à mettre en place une activité d'intelligence économique et plus particulièrement un processus de veille technologique, « veille technologique et créative » (Azemard, 2001), ou « veille artistique » (Mahé, 2004). En effet, la veille technologique permet aux entreprises de se tenir informées des toutes dernières évolutions technologiques et de rester ainsi

aux aguets des nouvelles technologies pour répondre aux besoins de l'innovation (OCDE, 1993). D'après Lesca (1994), « Une entreprise qui veut être durablement compétitive devrait être constamment aux aguets afin d'être en mesure de tirer profit des changements qui, la plupart du temps, s'imposent à elle ».

La veille est un outil d'aide à la décision (Réix, 2000), qui a pour objectif de repérer les acteurs importants, nouveaux ou potentiels, les alliances entre acteurs, les développements en cours, les nouveaux produits, les tendances émergentes en termes de technologies et de projets de recherche du marché en question. Par ailleurs, la veille consiste à favoriser la créativité (Marteau & Lesca, 1986) et le renouvellement rapide (Bayar & Montagnier 1996; Veltz, 2000; Guilhot B. & Alaoui, 2005). De ce fait, sa principale mission consiste à soutenir l'innovation (Julien et al., 1999; Burkhart, 2001, Chambaud 2011). D'autant plus que dans un contexte en pleine transformation, la veille constitue une activité de support qui contribue à la création de valeur (Lesca et Mancret, 2003), son objectif vise à réduire la « myopie managériale » au regard de l'environnement par l'exploration ou l'observation, afin d'identifier des événements pouvant éveiller un intérêt pour développer de nouveaux produits ou services (Albright, 2004). Cette démarche par laquelle l'entreprise scrute l'environnement afin de détecter opportunité ou menace, permet d'envisager et de stimuler des idées nouvelles. Ainsi, cette activité d'observation et de surveillance permanente de l'environnement peut alimenter en amont les démarches d'anticipation et de créativité (Lesca et Caron-Fasan, 2006). Plus encore, Gorla (2009) conclut que, la veille créative doit se positionner au début et en parallèle du processus d'innovation. D'autant plus que pour faire du SI une source d'innovation, il est indispensable de déployer un processus de veille technologique et métier de l'entreprise en synergie avec les équipes de recherche et développement. (Bourdon et Lehmann-Ortega, 2007; Kohli, 2007).

L'optimisation des capacités d'innovation n'est possible que par un management de l'innovation favorisant l'accélération de la veille et la capitalisation des connaissances, produit de l'intelligence collective (Groff, 2009).

2.3.3. Dimension gestion de la connaissance

Les entreprises qui ont réussi des initiatives en gestion des connaissances affichent une meilleure performance d'innovation (Mitchell et Kulik, 2000). La gestion des connaissances reflète la stratégie consistant à structurer le capital de connaissances d'une organisation, en lien avec ses ambitions stratégiques et ses besoins d'innovation, supportée par une infrastructure technologique et organisationnelle, organisée autour du processus de gestion de connaissances, avec l'humain comme premier lieu d'interaction et de création de connaissance (Jacob et Parlat, 2000). Schumpeter considère les innovations comme des « combinaisons nouvelles » réalisées par l'entrepreneur à partir des nouvelles technologies. Il est clair que l'innovation est le résultat des interactions entre la connaissance, les technologies et les acteurs de l'écosystème.

Dès lors que les SI deviennent un maillon essentiel de la chaîne de valeur, ils peuvent se positionner au cœur d'une stratégie d'innovation, car ils soutiennent un

échange permanent de connaissances entre l'organisation et son environnement. Les SI permettent ainsi de coordonner les outils et les produits de l'intelligence collective. La gestion de l'innovation et la gestion des connaissances sont d'ailleurs considérées comme deux « pratiques » non exclusives ; elles se complètent et peuvent même parfois se confondre (Drucker, 1993 ; OCDE, 2005).

Différents travaux de recherche considèrent que la capacité d'une organisation à innover dépend essentiellement du processus de production et de la circulation de la connaissance (Drucker 1993 ; Nonaka, 1994 ; Nonaka et Takeuchi, 1995 ; Von Krogh et al., 2000). L'innovation est donc le résultat de la transformation et la combinaison de connaissances de nature différente (explicites/tacites) (Leonard et Straus, 1997). En effet, le processus de création de la connaissance a pour objectif fondamental de répondre aux enjeux d'innovation. La production des connaissances est le résultat de deux activités essentielles, la première repose sur la création interne de connaissances, dont le moteur est la créativité de l'entreprise. La seconde est liée à l'apprentissage dit « organisationnel » (Argyris et Schön, 1978) qui cristallise la connaissance au niveau collectif de l'organisation, et qui consiste à intégrer et remodeler constamment et collectivement de nouveaux savoirs et savoir-faire. En effet, le processus de créativité est basé sur des techniques dites de créativité afin de créer des connaissances. Il est souvent organisé autour de groupes de travail dont l'objectif est de fournir de nouvelles idées, qui sont par la suite évaluées, et intégrées dans une activité de conception innovante. Ainsi, la capacité créatrice dépend de la créativité de différents acteurs de l'entreprise qui s'est vu élargir dans les nouveaux dispositifs d'innovation ouverte à l'ensemble des acteurs de l'écosystème. Selon Austin et al (2012), les outils fournis par les SI permettent de favoriser le processus créatif par la variété des idées et l'élargissement de la population des participants. Ils facilitent la combinaison de connaissances hétérogènes et revisiter ou voir les connaissances existantes dans l'entreprise avec un nouvel œil.

La perspective interactionniste appréhende l'innovation comme le résultat d'interactions complexes entre les individus et les structures organisationnelles, l'usage des SI contribue à accentuer ces interactions, d'autant plus que la diffusion et l'accès rapide aux nouvelles connaissances sont des facteurs déterminants de l'innovation (Lambin, 2011). Les TIC permettent ainsi d'accélérer le processus d'innovation par la capitalisation des savoirs (Pavlou et El Sawy, 2006). Elles facilitent aussi l'identification des besoins émergents du marché (Tambe et al., 2012) et soutient l'organisation pour apporter les réponses les plus appropriées aux évolutions de l'environnement (Song et al., 2005). Ainsi Kmieciak et al. (2012) démontrent le rôle des TIC pour une meilleure compréhension du marché en facilitant les échanges avec les clients, ce qui favorise l'innovation collaborative par la participation des clients à la conception des produits répondant à leurs attentes. L'encadré 2 met en lumière les pratiques en matière de veille et partage des connaissances afin d'innover tout en s'appuyant sur les SI.

Encadré 2

« Groupe La Poste : Veille & Partage des connaissances » (Les matinales, 2012)

La direction d'innovation a lancé une plateforme « BlueKiwi » dédiée au partage des connaissances en termes de veille spécifiquement sur les activités internet impactant les métiers du Groupe. Cette communauté d'innovateurs compte 400 personnes, le partage du contenu, et l'animation visent à valoriser ce contenu afin d'identifier des partenaires au sein du Groupe.

« InnoCentive : Crowdsourcing de résolution de problèmes » (Afuah et Tucci, 2012)

InnoCentive est une plateforme d'intermédiation qui permet aux entreprises de soumettre des problèmes pour lesquels n'ont pas de solution auprès de la foule des chercheurs, experts et particuliers. Le principe est simple, l'entreprise envoie de manière confidentielle un problème sur le site web d'InnoCentive. L'individu qui apporte la meilleure solution gagne une prime. L'entreprise s'appuie ainsi sur des ressources et des connaissances externes afin d'innover.

La connaissance peut être perçue comme un flux, en perpétuelle construction tout en mettant le capital humain au cœur des différents dispositifs. D'où notre questionnement sur quel rôle peut jouer le capital humain dans le processus d'innovation par les SI.

2.3.4. Dimension capital humain

Le capital humain joue un rôle fondamental dans l'acquisition d'un avantage concurrentiel et dans la création de valeur (Hamel and Prahalad, 1994 ; Edvinsson et Malone, 1997; Lev, 2001). « Le capital humain fait référence aux connaissances, compétences, expériences, ainsi qu'aux motivations et attitudes des salariés d'une organisation » (Davenport, 1999). Les travaux effectués dans le cadre de projets européens ont mis en lumière le rôle du capital humain dans le processus d'innovation et de croissance (OCDE, 2001a, b & c). En effet, il représente le vecteur principal de l'innovation, par le biais d'une part de sa mobilité, de sa volonté d'interaction et d'autre part de son accumulation. Cette accumulation de capital humain permet d'agréger les connaissances de chaque individu et de les mettre en commun, la coopération et l'échange étant des moteurs reconnus de l'innovation (Ciccone & De la Fuente, 2002). Ainsi, la valeur ajoutée du capital humain étant son niveau de connaissance, les individus les mieux qualifiés seraient les plus à même de constituer une force motrice de l'innovation (Brenner, 2003), c'est ce que qualifie Florida (2003) de capital humain « créatif » ou de « talent ».

L'adoption des TIC par les organisations nécessite l'existence préalable d'un capital humain capable de s'adapter avec les spécificités des TIC. En effet, l'innovation en entreprise est le fruit du capital humain. Pour susciter la créativité, les dirigeants doivent impliquer leurs collaborateurs et développer un style de management adapté à l'innovation (Reix, 2002). L'entreprise doit tirer profit de l'intelligence collective car l'innovation se situe à tous les niveaux et services de l'entreprise et ne se limite pas au département recherche et développement. En outre, les entreprises qui souhaitent atteindre un haut niveau de performance doivent combiner un positionnement stratégique optimal et un capital humain de meilleure qualité que celui de leurs concurrents (Chamak et Fromage, 2006). Elles doivent identifier et développer les compétences humaines et créer un contexte qui favorise l'expression de ces compétences, leur capitalisation et leur diffusion. L'aptitude à innover et engendrer la performance dépend largement des capacités de son capital humain (Prager, 2005). « L'exemple d'eBay illustre bien cette dépendance: il faut non seulement des scientifiques et des ingénieurs pour développer la plateforme, mais aussi des connaissances commerciales étendues et un esprit d'entreprise dont font montre les millions de petits négociants qui utilisent la plateforme. Et enfin, il faut des centaines de millions de clients capables de placer des enchères sur le site » (Prager, 2005).

Le capital humain est une ressource vitale à toutes les étapes du processus de l'innovation, il joue un rôle primordial tant lors des phases de recherche et de conception du service, qui doivent être animées par la vision des dirigeants, qu'au niveau du développement technique ou commercial, avec l'acquisition éventuelle de compétences spécifiques (DGCIS, 2012). En outre, un projet d'innovation n'a pas

d'impact seulement sur ceux qui en sont acteurs. Dans certain cas, un plan de conduite du changement devra être pensé et mis en œuvre afin d'accompagner l'ensemble des acteurs de l'organisation. La mobilisation du capital humain constitue ainsi, un élément crucial permettant à l'entreprise d'accroître sa capacité de transformation et d'innovation. « Les hommes et les femmes sont – par leur connaissance intime des clients et de l'environnement, par leurs savoir-faire en permanence affûtés, par leur imagination et par les initiatives qu'ils sont autorisés à prendre pour satisfaire de façon originale le client – la principale source de différenciation et de performance de l'entreprise » (Badot, 1999).

Cette réflexion nous amène au questionnement sur quelle culture peut favoriser la créativité et l'innovation.

2.3.5. Dimension culture d'innovation

D'après l'American Heritage Dictionary, la culture de l'entreprise se définit comme : « l'ensemble des usages, des arts, des croyances, des institutions et de toutes les productions du travail et du génie humain, partagés et transmis par une collectivité ou une population donnée ». En effet, lorsqu'on évoque la culture, on cherche à mettre l'accent sur ce qu'il y a de commun aux membres d'une entreprise (Thévenet, 1993). Mais qu'est-ce qui explique que la culture d'une organisation est un levier important dans le développement de sa capacité à innover ? Quel est rôle peut jouer la culture d'innovation ?

Trompenaars (1994), définit la culture comme : « la manière dont un groupe de personnes résout des problèmes ». Cette définition est plus proche du domaine de l'innovation des entreprises, dont la survie dépend de l'efficacité avec laquelle les entreprises résolvent les problèmes auxquels elles sont confrontées et s'adaptent aux changements. En effet, l'innovation par son importance stratégique constitue un « axe » majeur dans la culture de l'organisation. Ainsi, pour faire émerger une culture d'innovation, Sainsaulieu (1994) a évoqué la question de la cohésion sociale produite par la transmission de normes, valeurs et représentations suffisamment communes pour faire émerger une « sociabilité collective » entre les membres de l'entreprise. En effet, la culture d'entreprise est perçue comme le résultat d'un processus cumulatif et spécifique au sein duquel les interactions sociales jouent un rôle déterminant pour assurer l'innovation. Elle se construit progressivement au travers des interactions entre actions d'innovation et pratiques de routines quotidiennes, d'où la complexité de la relation entre innovation et culture d'entreprise ainsi que les mécanismes d'émergence d'une culture d'innovation. D'après Alter (2002), la culture de l'entreprise est basée sur les occasions d'innover dans la technique, l'organisation, le produit et la gestion qu'induit la vie des entreprises face aux pressions du marché et les évolutions technologiques. L'entreprise qui a ainsi une culture insistant sur l'innovation tend à développer plus d'innovations, la culture dans ce cas là joue un rôle important dans la créativité et le développement de l'innovation. Ainsi, lorsqu'une culture d'ouverture à l'innovation caractérise la culture d'entreprise cela lui permet une ouverture aux nouvelles idées, et l'acceptation du changement. Les croyances et les valeurs partagées par son

capital humain influencent positivement son degré d'ouverture afin d'agir et gagner du temps en adoptant constamment des comportements innovants.

Dans ce sens, Brown & Eisenhardt (1997) ont étudié la relation entre la culture et l'innovation, ils distinguent ainsi deux niveaux d'innovation : l'innovation radicale ou de rupture, et l'innovation incrémentale. L'innovation radicale modifie de manière profonde les références utilisés et nécessite le recours à des compétences nouvelles, et des technologies. Dans ce cas, la culture d'innovation contribue à motiver les individus et attirer les meilleurs talents en leur donnant l'opportunité de se réaliser. Cette culture stimule donc la capacité innovatrice. Quant à l'innovation incrémentale, elle permet une amélioration continue des références existantes et n'exige pas de savoir-faire nouveaux. Il s'agit plutôt d'une adaptation et elle est perçue comme une acceptation/ouverture à de nouvelles idées.

Par ailleurs, les projets d'innovation sont perçus comme étant risqués dans la mesure où leur valeur future est incertaine (Erickson et Jacobson, 1992). L'innovation est une activité, qui par nature, fait encourir des risques, d'où la formule de Flaubert (1966) « Innovation : toujours dangereuse ». Ainsi, pour innover et obtenir des résultats, il faut accepter de prendre des risques. Cette capacité à prendre des risques est un élément déterminant pour favoriser l'innovation. De ce fait, sans prise de risque, les idées créatrices ne pourront pas être testées et aboutir à une ou des innovations. En conséquence, de nouvelles pratiques organisationnelles s'imposent : permettre l'échec, ne pas tout planifier, ajuster ses innovations en fonction des opportunités (flexibilité). Les pratiques des entreprises matière de gestion du capital humain et les modèles culturels à adopter afin de répondre aux enjeux de l'innovation par l'usage des SI évoluent ce qui amène à une transformation pour s'adapter, l'encadré 3 nous donne un aperçu de cette démarche au sein du groupe pages jaunes.

Encadré 3

« Pages Jaunes (Transformation, Capital humain) : Génération des idées » (Les matinales, 2012)
Dans le cadre du programme de transformation, le groupe Pages Jaunes développe l'innovation participative par la mise en place d'une plate forme ouverte afin de recenser les idées internes et externes et innover. Les propositions des acteurs internes et externes sont recueillies par un réseau d'experts, qui en sélectionne les meilleures pour une mise en œuvre. Le suivi des contributions est disponible à tout moment par une simple connexion à la plateforme. Cette incitative a permis de mobiliser l'ensemble du capital humain en dépassant les obstacles organisationnels et hiérarchiques et d'accéder à des connaissances en provenance des structures internes et externes. L'absence d'anonymat sur cette plateforme a permis de donner une reconnaissance à des individus peu visible. Aussi, l'accumulation des points permet de leur faire gagner des cadeaux.

Enfin, l'étude de la littérature sur les liens entre le processus d'innovation par l'usage des SI d'une part, et les différentes types d'innovations produites d'autre part, fait émerger cinq dimensions contributrices aux différentes formes d'innovations (La figure 1). En effet, les différentes publications et études soulignent certes l'existence d'un lien entre les différentes dimensions et l'innovation mais ne permettent pas d'en appréhender l'importance de chaque dimension ni l'arbitrage judicieux entre elles ou la démarche à mettre en place au sein de l'entreprise afin d'accélérer cette contribution des SI à l'innovation. D'où

notre intérêt à mener une étude quantitative auprès des entreprises afin d'expérimenter notre modèle et s'assurer davantage sur la nature de la relation entre les SI et l'innovation, et les mécanismes permettant d'accélérer telle contribution.

3. Travaux connexes

Ce travail de recherche a pour but de proposer un modèle de maturité structuré par domaine de processus (13 domaines de processus) sous forme de bonnes pratiques (60 bonnes pratiques) issues de la revue de la littérature et implémentées aux sein des entreprises (Figure 2). En effet, il s'agit d'une nouvelle approche évolutive en fonction de la conception de l'innovation. Ce framework guide les entreprises afin de diagnostiquer la maturité de la démarche d'innovation en s'appuyant sur les SI et définir une démarche d'amélioration continue du fait qu'il sert comme outil d'évaluation de la maturité. L'originalité principale réside dans l'usage des connaissances extraites de la recherche scientifique et des bonnes pratiques mises en œuvre au sein des entreprises et reconnues par la communauté des praticiens. Ce modèle qui est similaire au CMMI, aide les entreprises en leur fournissant des bonnes pratiques, un outil et des facteurs de succès à respecter lors de la mise en œuvre. Le but est de mener une démarche progressive adaptée au contexte de chaque organisation afin d'innover de manière continue et durable. Ainsi, lors de l'étude exploratoire, nous enrichissons notre modèle qui sera revu en phase d'expérimentation prévue en étude quantitative auprès d'un large échantillon d'entreprises de différents secteurs d'activités ce qui nous permettra de valider le modèle de maturité, l'outil d'évaluation ainsi que les facteurs de succès à respecter lors de l'implémentation au sein des entreprises.

CATEGORIES	DOMAINES DE PROCESSUS								
PILOTAGE	SI Stratégie d'innovation		GSI Gouvernance des SI						
INGENIERIE D'INNOVATION	IGE Inspiration, Génération d'idées et Exploration des concepts			TDE Test, Démonstration & Evaluation		GPD Gestion de projet de développement		LM Lancement sur le marché	
SUPPORT	GCH Gestion du capital humain	PI Protection de l'innovation	VTM Veille Techno & métier	GC Gestion des Connaissances	FI Financement de l'innovation	MP Mesures de performance	CI Culture d'innovation		

Figure 2: Modèle de maturité de l'innovation par l'usage des SI

4. Conclusion

Cette recherche avait pour objectif, d'une part, d'explorer le processus d'innovation en s'appuyant sur les SI et, d'autre part, de conceptualiser les différentes dimensions de ce processus afin de proposer un modèle de maturité d'innovation par l'usage des SI. Ce travail s'appuyait également sur une analyse des

dispositifs mis en œuvre par quelques entreprises. Afin d'expérimenter ce modèle, nous sommes en phase de réalisation d'interviews auprès d'un échantillon d'entreprises pour valider, enrichir et/ ou revoir certains éléments du modèle et d'en faire un modèle de maturité, présenté selon une structure similaire au CMMi tout en montrant sa validité générale.

Bibliographie

- Afuah A., Tucci C., (2012). Crowdsourcing As a Solution to Distant Search. *Academy of Management Review*, Vol. 37 Issue 3, p. 355-375.
- Agence de développement du Val de Marne (2012), Coopération et innovation, *Cycle 10, Les matinales*.
- Albright K.S., (2004). Environmental scanning : radar for success. *The Information Management Journal*, Mai-Juin, 2004.
- Alter N., (2002). *Les logiques de l'innovation: approche pluridisciplinaire*, éditions la Découverte. Paris.
- Argyris C., Schön D., (1978). *Organisational Learning: a Theory of Action Perspective*, Addison- Wesley, Reading.
- Aubert B., Cohendet P., Montreuil B., Le Roux R., Peccatte C., Rouges J-F., (2011). *Comprendre l'innovation à l'aide des technologies de l'information et des communications*. Rapport de recherche, Montréal.
- Austin RD., Devin L., Sullivan EE., (2012). Accidental innovation: Supporting valuable unpredictability in the creative process. *Organization Science*, 23(5), p.1505-1522.
- Azemard Gh., (2001). L'offre de jeux vidéo : évolutions du marché. *Médiamorphoses*, INA, Bry-sur-Marne.
- Badot O., (1999). Analyse critique de la théorie du « design organisationnel » : Le cas des sociétés de commercialisation de services de radiotéléphonie mobile de type GSM en France. *Revue de gestion des ressources humaines*, n° 29-30, p.49-61.
- Bayar V., Montagnier P., (1996). « Le boom des technologies de l'information ». *L'Observateur de l'OCDE*, n° 198, 02-03.
- Béatrice C., Foratier Y., Potié P., (2009). *Valoriser la DSI au sein de l'entreprise : Faire du marketing de la DSI un levier de la performance*, Dunod, Paris
- Boland R.J.Jr., Lyytinen K., YOO Y., (2007). Wakes of innovation in project networks: The case of digital 3-D representations in architecture, engineering, and construction. *Organization Science*. 18(4), p.631-647.
- Borgatti S.P., Cross R., (2003). « A relational view of information seeking and learning in social networks », *Management Science*, vol. 49, no 4, p. 432-445.
- Bourdon I., Lehmann-Ortega L., (2007). Systèmes d'information et innovation stratégique : une étude de cas. *Systèmes d'Information et Management*, Vol. 12, No 1.
- Brenner T., (2003). *Policy Measures to Create Localised Industrial Clusters*, in D. Fornahl and T. Brenner (eds.) : The Influence of Co-operations, Networks, and Institutions on Regional Innovation Systems, Cheltenham.
- Brown S.L., Eisenhardt K.M., (1997). The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations. *Administrative Science Quarterly*, 42, p.1-34.
- Burkhart., (2001). Competitive Intelligence and the product life cycle. *Competitive Intelligence review*, vol.12, n° 3, p.35-43.
- Burrus D., (2013). *Today's CIO : The Chief Innovation Officer*, http://www.huffingtonpost.com/daniel-burrus/todays-cio-the-chief-inno_b_3989501.html
- Chamak A., Fromage C., (2006). *Le capital humain*, Editions Liaisons juin 2006, p.48-49.

- Chambaud S., (2011). Veille et innovation. *Documentaliste - Sciences de l'information*, 2011, vol. 48, n°1, p.22.
- Chesbrough H., (2006). "Open Business Models: How to Thrive in the New Innovation Landscape". *Harvard Business School Press*, Boston.
- Chesbrough H.W., (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Cambridge, MA: Harvard Business School Publishing.
- Choo C.W., (2002). « Information management for the intelligent organization: the art of scanning the environment ». *American Society for information science and technology by information today*, Medford.
- Choo C.W., (1999). « The art of scanning the environment ». *ASIS Bulletin*, n°3, *special issue on information seeking*, p13-19.
- Ciccone A., De la Fuente A., (2002). *Human capital in a global and knowledge-based economy*. UFAE and IAE Working Papers 562.03, Unitat de Fonaments de l'Anàlisi Econòmica (UAB) and Institut d'Anàlisi Econòmica (CSIC).
- CIGREF-CAPGEMINI CONSULTING, (2009). *L'information : prochain défi pour les entreprises, Pratiques de création de valeur par les SI et leur usage, cartographie*. Etude, 52 pages.
- CIGREF-MCKINESY COMPAGNY, (2008). *Dynamique de création de valeur par les Systèmes d'information : Une responsabilité partagée au sein des Directions des Grandes Entreprises*. Etude, 35 pages.
- Davenport T.O., (1999). *Human capital: What it is and Why People Invest it*, Jossey-Bass, San Francisco, CA
- Delmond M-H., Keravel A., Coelho F., MAHLER, (2013). *Business models, coproduction de valeur et systèmes d'information*, Springer-Verlag France, p.5-27.
- Dgeis (2012). *Osez l'innovation*, <http://www.entreprises.gouv.fr/files/files/guides/osez-l-innovation.pdf>
- Drucker P.F., (1993). *Au-delà du capitalisme : La métamorphose de cette fin de siècle*, Paris : Dunod.
- Duncan R.B., (1972). « Characteristics of organizational environments and perceived environmental uncertainties ». *Administrative Science Quarterly*, vol. 17, n° 3, 1972
- Erickson Gary M., Jacobson R., (1992). Gaining competitive advantage through discretionary expenditures: returns to R & D and advertising. *Management Science*, 1992, vol. 38, September, p.1264-1279.
- Fernez-Walch S., Romon F., (2008). *Dictionnaire du Management de l'innovation*, Editions Vuibert.
- Flaubert G., (1966). *Dictionnaire des idées reçues*, Nizet A.G. V° Innovation, Paris.
- Florida R., (2003). *The Rise of the Creative Class – Basic books*, New York.
- Gartner. (2014). *DSI doivent se préparer à la gestion de l'information à deux vitesses comme troisième ère de l'aube d'informatique*, diffusé par Dan Worth 15 janvier 2014 <http://www.v3.co.uk/v3-uk/news/2322642/cios-must-prepare-for-two-tier-it-management-as-third-era-of-computing-dawns>
- Goria S., (2009). «Entre la veille stratégique et l'innovation, la démarche de veille créative : ce que la veille créative emprunte aux wargames sur plateau », *Séminaire international Veille Stratégique Scientifique et Technologique*, VSST'2009, Nancy.
- Groff A., (2009). *Manger l'innovation, 100 questions pour comprendre et agir*, Afnor Editions
- Guilhot B., Alaoui A., (2005). *La compétitivité internationale : stratégie pour les entreprises françaises, édition l'Harmattan*, 304 p.
- Hamel G., Prahalad C.K., (1994). "Competing for the future". *Harvard Business Review*, vol. 72, n° 4, 1994, p. 122.

- Jacob R., Pariat L., (2000, janv.). *Gérer les connaissances. Un défi de la nouvelle compétitivité du 21^{ème} siècle: Information, connaissance, interaction, gestion des ressources humaines*. Rapport de recherche commanditée. Québec, CEFRIO.
- Julien P.A., Raymond L., Jacob R., and Ramangalahy C., (1999). *Types of technological scanning in manufacturing SMEs : an empirical analysis of patterns and determinants*. *Entrepreneurship & Regional Development*, vol 11, p 282-300.
- Kéfi K. (2004). *Evaluation des systèmes d'information : une perspective organisationnelle*, *Economica*.
- Kmieciak R., Michna A., et Meczynska A., (2012). "Innovativeness, empowerment and IT capability: evidence from SMEs". *Industrial Management & Data Systems*, 112(5), p.707-728.
- Koenig G., (1996), *Management stratégique : paradoxes, interactions et apprentissage*, Paris : Nathan, p.544.
- Kohli R., (2007). Innovating to Create IT-Based New Business Opportunities at United Parcel Service. *MIS Quarterly Executive*, Vol. 6, N°4, December 2007, p. 199-210.
- L. Edvinsson M.S., Malone, (1997). *Intellectual capital: realizing your company's true value by finding its hidden roots*, HarperCollins, New York.
- LAMBIN J.J., (2011). *Quel avenir pour le capitalisme ?*, Dunod, Paris.
- Laudon K., LAUDON J., (2010). Chapitre 3 – Les systèmes d'information et la stratégie des organisations. *Management des systèmes d'information*. 11e éd.–Pearson Education
- Leonard D., STRAUS S., (1997). Putting Your Company's Whole Brain to Work. *Harvard Business Review*, mai-juin 1997, p. 69-79.
- Lesca H., Mancret G., (2003). *Innovation: résultat de la perception de signaux faibles et de l'intelligence collective*.
- Lesca H., (1997). *Veille stratégique, concepts et démarche de mise en place dans l'entreprise, Guide pour la pratique de l'information scientifique et technique*, Ministère de l'éducation Nationale de la Recherche et de la technologie, p.27.
- Lesca N., Caron-Fasan M.L., (2006). *Veille anticipative : une autre approche de l'intelligence économique*. Ed Hermès Lavoisier, 280 pages.
- Lev B., (2001). *Intangibles: Management, Measurement, and Reporting*, Brookings Institution Press, Washington D.C.
- Liang T.P., You J.J., Liu C.C., (2010). A Resource-based Perspective on Information Technology and Firm Performance: a Meta Analysis. *Industrial Management & Data Systems*, Vol. 110, n°8, p. 1138-1158.
- Loilier, T. et A. Tellier (2013), *Gestion de l'innovation : Comprendre le processus d'innovation pour le piloter*, 2ème édition, Editions EMS.
- Mahé E., (2004). *Pour une esthétique informationnelle : La création artistique comme anticipation des usages sociaux des technologies de l'information et de la communication?* Université Rennes 2 Haute Bretagne. Thèse en sciences de l'Information et de la Communication, p.212.
- Marteau G., Lesca H., (1986). Pourquoi surveiller l'environnement. *Direction et Gestion*, vol. 2, p.12-22.
- Mitchell C., KULIK T., (2000). Knowledge Management: Becoming an E-Learning Organization. *The Conference Board*, New York.
- Nonaka I., Takeuchi H., (1995). *The Knowledge-Creating Company: How the Japanese Companies Create the Dynamic of Innovation*, Oxford University Press, New York.
- OCDE, (2001). *Investment in Human Capital through Post-compulsory Education and Training – Working Paper*.
- OCDE, (2001). Knowledge and Skills for Life, First Results from PISA 2000, *Programme for International Student Assessment*. Paris.
- OCDE, (2001). The new economy: beyond the hype. *The OECD growth project*. Paris.

- OCDE, (2003). *Les TIC et la croissance économique*, p.9-13,
<http://www.oecd.org/dac/ictcd/docs/otherdocs/OtherOECD ICT eco growth fr.pdf>
- OCDE, (2005). *Manuel d'Oslo* – 3ème édition.
- Parent R., Chanal V., (2009). Quels business models pour les plateformes Web 2.0: les apports de la théorie des marchés bi-faces. *Actes de la Conférence Internationale de Management Stratégique*, Grenoble.
- Pavlou P., EL SAWY O., (2006). From IT Leveraging Competence to Competitive Advantage in Turbulent Environments: the Case of New Product Development. *Information Systems Research*, 17(3), 198-227.
- Powell W.W., Koput K.W., Smith-Doerr L., (1996). *Interorganizational Collaboration and the Locus of Innovation*, Administrative Science Quarterly, 41(1): p.116-145.
- Prager. J.C., (2005). *Diagnostic de l'Innovation en Alpes-Méditerranée*. Groupe de travail Innovation,
- Raymond L., (2003). *Les systèmes d'information, dans GREPME, Les PME : Bilan et perspectives*. 2e édition, Cap-Rouge, Québec, Les presses Inter Universitaires, Paris, France, Economica, chap. 9.
- Reix R., (2002). Changements organisationnels et technologies de l'information. *Conférence invitée à l'Université Saint-Joseph*, Beyrouth, Liban (28/10/2002).
- Reix R., (2004). *Systèmes d'information et management des organisations*, Vuibert, Paris.
- Reix R., (2000). *Systèmes d'informations et management des organisations*, Editions Vuibert.
- Riccio P.M., Bonnet D., (2012). *TIC et innovation organisationnelle*, Presses des Mines.
- Rochet c., (2006). Cahier N°3 *Spécial Innovation*, Cahier de recherche, CIGREF.
- Rothwell R., (1994). Towards the fifth-generation Innovation Process. *International Marketing Review*, vol. 11, n°1, p.7-31.
- Sainsaulieu R., *méthodes pour une sociologie de l'entreprise*, Presses de la fondation nationale des sciences politiques, Lyon, 1994.
- Schmookler J., (1966). *Invention et Economic Growth*, Harvard University Press, Cambridge.
- Schumpeter J. A., (1934). *The Theory of Economic Development*, Harvard University Press, Cambridge.
- Standage T., (2005). *The Future of Technology*, Profile Books Ltd, London.
- Subramanian R., Fernandes N., Harper E., An empirical examination of the relationship between strategy and scanning. *The Mid-Atlantic Journal of Business*, vol. 29. n° 3, p. 315-331, 1993.
- Tambe P., Hitt L., Brynjolfsson E., (2012). The Extroverted Firm: How External Information Practices Affect Innovation and Productivity. *Management Science*, Vol. 58, n°5, p. 843-859.
- Thevenet M., *Culture d'entreprise*, 1993, PUF, Paris.
- Thiéart R.A., (1990). *La stratégie d'entreprise*. Mac Graw Hill, Paris.
- Trompenaars E., (1994). *L'entreprise multiculturelle*, Maxima. Laurent du Meshil, p.30.
- Veltz P., *Mondialisation, villes et territoires : l'économie d'archipel*, PUF, Paris, 2000.
- Von Hippel E., (1986). Lead-users : Sources of novel product concept. *Management Science*, 32, 7, july, 791-805.
- Von Krogh G., Ichijo K., Nonaka I., (2000)., *Enabling Knowledge Creation*, Oxford University Press, New York.

Extensions du diagramme d'activité pour contrôler l'accès au SI

Salim Chehida¹, Akram Idani², Yves Ledru³,
Mustapha Kamel Rahmouni⁴

1. Université d'Oran Es-sénia
31000 Oran, Algérie
Salim.Chehida@imag.fr
2. Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
Akram.Idani@imag.fr
3. Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
Yves.Ledru@imag.fr
4. Université d'Oran Es-sénia
31000 Oran, Algérie
kamel_rahmouni@yahoo.fr

RÉSUMÉ. L'ouverture des organisations et de leurs Systèmes d'Information (SI) pose le problème de leur sécurité. La définition d'une politique de contrôle d'accès est l'une des principales étapes dans la conception d'un SI. Ce travail propose une approche pour la spécification d'une politique de sécurité, basée sur le modèle RBAC, au niveau des workflows. Cette approche consiste à propager les permissions définies sur un diagramme de classes modélisé avec le profil SecureUML, vers des contraintes associées aux activités décrivant un processus métier. Les diagrammes d'activité sont définis à deux niveaux : un niveau abstrait qui ne détaille pas ces permissions et un niveau concret où des contraintes sont associées à certaines actions ou à l'ensemble du diagramme.

ABSTRACT. The evolution of organisations and their information systems towards more openness raises the challenge of their security. The definition of an access control policy is a major activity in the design of an Information System. This paper proposes an approach for the specification of security policies, based on the RBAC model, at the workflow level. This approach propagates permissions defined on a class diagram, using the SecureUML profile, towards constraints linked to the business process activities. Activity diagrams are defined at two levels: an abstract level which does not detail these permissions and a concrete level where constraints are associated to specific actions or to the whole diagram.

MOTS-CLÉS : RBAC, Workflow, Processus métier, SecureUML, Diagramme d'activité d'UML 2

KEYWORDS: RBAC, Workflow, Business process, SecureUML, UML2 Activity Diagram

1. Introduction

Avec la croissance fulgurante que connaissent le monde des télécommunications et l'ouverture des Systèmes d'Information (SI), la spécification de ces SI ne peut plus se contenter de la seule modélisation fonctionnelle mais doit prendre en compte les besoins de sécurité. La définition des cas d'utilisation et des activités associées doit s'enrichir en intégrant le contrôle d'accès dans la description de ces processus. Adopté comme une norme ANSI / INCITS (ANSI, 2004), Role-Based Access Control (RBAC) (Ferraiolo *et al.*, 2003) est le modèle de contrôle d'accès le plus répandu dans les systèmes informatiques. En RBAC, l'accès à un objet est accordé à un utilisateur en fonction du rôle qui lui est associé. Cependant, si RBAC est bien adapté à l'expression d'une politique de contrôle d'accès dans une vue statique comme le diagramme de classes, il est plus difficile de formaliser une politique de contrôle d'accès dans une vue dynamique du système (Basin *et al.*, 2006). Ce travail s'intéresse à cette problématique ; nous proposons une nouvelle approche qui permet de représenter une politique de sécurité, basée sur le modèle RBAC, dans une vue dynamique en exprimant les règles de contrôle d'accès sur les "workflows"¹ d'un SI.

SecureUML est un profil UML qui permet de spécifier une politique de contrôle d'accès RBAC sur les diagrammes d'UML. Ce profil étend le modèle RBAC en exprimant des contraintes contextuelles : les contraintes d'autorisation. Ces contraintes portent sur l'état spécifié par le diagramme de classes et conditionnent l'évaluation des permissions. Nous proposons de compléter cette vue statique par des diagrammes d'activité d'UML 2 (UML2, 2011), qui sont l'un des modèles préconisés pour la modélisation des workflows. Nous étendons ces diagrammes pour qu'ils représentent le déroulement d'un processus métier en tenant compte des différentes permissions et contraintes d'autorisation d'une spécification SecureUML.

Notre approche identifie les exigences de sécurité dès les premières étapes du cycle de vie de développement. Elle intègre trois vues de modélisation. La première vue est fonctionnelle ; elle est représentée par le diagramme de cas d'utilisation qui montre des acteurs interagissant avec les grandes fonctions d'un système. La deuxième est statique ou structurelle et décrit les données d'un système sous forme de classes et d'associations, et les permissions qui leur sont associées. Elle est décrite en SecureUML. La troisième est dynamique et permet d'établir un pont entre les deux premières visions. Pour construire ce pont, nous avons utilisé le diagramme d'activité à deux niveaux différents. Le premier est abstrait et permet de décrire les cas d'uti-

1. Flux de travail

lisation par une coordination d’actions de haut niveau exécutées par les acteurs du système. Le deuxième est concret et consiste à exprimer les actions abstraites par des actions de bas niveau qui représentent les opérations des classes. Ce diagramme concret tient compte des permissions et contraintes d’autorisation exprimées en SecureUML qui seront associées comme pré-conditions aux actions concrètes concernées.

Le présent travail vise l’expression d’une politique de contrôle d’accès au niveau des activités d’un processus métier et poursuit les objectifs suivants :

- Propager une politique d’accès statique au niveau des activités.
- Vérifier les permissions dans l’exécution des activités métiers.
- Localiser les actions critiques.
- Définir une représentation facile à intégrer et à transformer dans des plateformes logicielles.

Dans la deuxième section, nous présentons les travaux qui traitent du contrôle d’accès au niveau des workflows. La troisième section présente SecureUML, le point de départ de notre approche ainsi que l’exemple de planification de réunions, qui illustre cet article. La section 4 discute l’expression des permissions et contraintes d’autorisation sous forme de préconditions sur les activités d’un processus métier. La section 5 définit le métamodèle qui permet d’étendre le diagramme d’activité pour le contrôle d’accès. Enfin, la dernière section conclut notre travail et présente quelques perspectives.

2. Contrôle d’accès au niveau de Workflow

Dans (WFMC, 1999), un workflow est défini comme étant “*l’automatisation totale ou partielle d’un processus d’entreprise, au cours duquel on échange d’un participant à un autre, des documents, des informations ou des tâches pour action, et ce selon un ensemble de règles procédurales*”. Ces règles sont nécessaires car elles conditionnent le bon fonctionnement du workflow. Dans le présent travail nous nous intéressons en particulier aux règles de sécurité issues de politiques de contrôle d’accès. Contrôler l’accès au niveau des workflows, selon (Kandala, Sandhu, 2002) consiste à assigner aux utilisateurs, conformément aux règles de l’organisation, des permissions pour effectuer certaines tâches au sein de l’organisation en fonction de leurs qualifications et responsabilités. Plusieurs travaux comme (Ahn *et al.*, 2000), (Bertino *et al.*, 1999) et (Wainer *et al.*, 2003), ont proposé des solutions basées sur le modèle RBAC. Une partie de ces travaux ont étendu le modèle RBAC pour contrôler l’accès au niveau des workflows. Ces extensions incluent par exemple : la spécification des politiques d’autorisation dynamiques (Ma *et al.*, 2011) et la délégation dynamique de tâches (Gaaloul, 2010) et (Wainer *et al.*, 2007). Dans notre travail nous ne cherchons pas à étendre RBAC, mais plutôt à combiner des formalismes approuvés et bien connus tels que SecureUML pour la conception de politiques RBAC et les diagrammes d’activités d’UML. L’objectif en est de disposer d’un ensemble de modèles accessibles aux divers acteurs d’un développement de logiciels.

Trois langages graphiques sont largement utilisés pour la modélisation des workflows : BPMN (BPMN2, 2011), les Réseaux de Pétri (Murata, 1989) et le diagramme d'activité d'UML 2 (UML2, 2011). (Geambasu, 2012) compare BPMN et le diagramme d'activité d'UML2 et conclut qu'ils sont équivalents pour représenter des processus de façon compréhensible et pour décrire des processus métiers. Notre choix dans cet article porte sur les diagrammes d'activité d'UML2 car ils incluent un support riche pour représenter les workflows : les actions, les flux de contrôle et de données ainsi que le choix, le parallélisme, les séquences et les événements. Parmi les travaux qui ont utilisé le diagramme d'activité d'UML 2 pour contrôler l'accès au niveau des workflows, (Jurjens, 2010) a proposé l'utilisation d'un package stéréotypé par « rbac » avec les trois tags `protected`, `role`, et `right` qui définissent respectivement l'ensemble des actions protégées d'un diagramme d'activité, les utilisateurs assignés aux rôles exécutant l'activité et l'affectation des actions protégées aux rôles qui peuvent les utiliser. (Strembeck, Mendling, 2011) ont défini un méta-modèle pour l'extension du diagramme d'activité vers le modèle RBAC.

3. SecureUML

UML est une notation graphique standard de l'OMG (Object Management Group) utilisée pour l'analyse des besoins et la conception d'un système. Elle présente l'avantage de pouvoir se décliner sous forme de « profils » permettant l'extension de ces diagrammes pour spécifier un aspect particulier d'un système. Plusieurs travaux ont proposé des profils utiles pour la spécification des politiques de contrôle d'accès basées sur le modèle RBAC. Parmi ces profils nous pouvons citer AuthUML (Alghathbar, 2012) qui propose des extensions du diagramme de cas d'utilisation et UMLsec (Jurjens, 2010) qui présente un profil pour étendre le diagramme d'activité.

Dans notre approche, nous utilisons le profil SecureUML (Basin *et al.*, 2006) (Basin *et al.*, 2009) qui permet de représenter une politique de sécurité basée sur le modèle RBAC dans une vue statique. Ce diagramme utilise des permissions, représentées par des classes associatives, pour exprimer les règles de contrôle d'accès. Une permission est liée à une classe stéréotypée par «Role» qui représente les utilisateurs affectés au rôle, et une autre classe stéréotypée par «Entity» qui représente la classe cible de la permission. La classe associative stéréotypée par «Permission» signifie que les utilisateurs assignés à «Role» sont autorisés, par la permission en relation, à accéder aux éléments de «Entity». Il est possible de soumettre cette permission à des conditions contextuelles, appelées conditions d'autorisation. Une permission SecureUML est définie par un ensemble d'attributs. Chaque attribut permet d'assigner une action à la permission et est défini par trois propriétés : des stéréotypes pour définir le type de la ressource à protéger, le nom de l'attribut qui détermine la ressource protégée, et le type de l'attribut pour spécifier l'action autorisée par la permission.

3.1. Exemple illustratif : organisation de réunions

Afin d'illustrer notre travail, nous allons utiliser l'exemple de l'organisation de réunions, proposé initialement par (Feather *et al.*, 1997). Ce système s'adresse à deux types d'utilisateurs : les "initiateurs" planifient des réunions et les "participants" répondent aux invitations des initiateurs.

Dans un premier temps, nous représentons les différents besoins fonctionnels du système en utilisant le diagramme de cas d'utilisation de la figure 1. Ce dernier exprime les différentes façons dont les acteurs peuvent utiliser le système. L'acteur Initiator peut créer des réunions, inviter des participants et suivre leurs réponses. L'acteur Participant répond aux invitations et suit les confirmations des réponses.

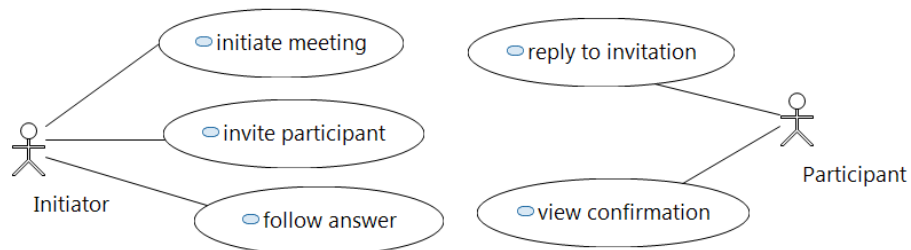


Figure 1. Le diagramme des cas d'utilisation du système d'organisation de réunions

Le système d'information permet d'enregistrer les données des personnes (participants et initiateurs), des invitations, des réunions et des propositions de changement, ainsi que les liens entre ces données. Afin d'assurer la confidentialité et l'intégrité des réunions, le système d'organisation des réunions applique une politique de contrôle d'accès qui définit une seule cible de sécurité: la classe Meeting. Elle se limite aux données les plus sensibles qui sont constituées par les données des réunions telles que la date, l'heure et le lieu.

La figure 2 présente la spécification SecureUML de la politique de sécurité, elle comprend le diagramme de classes et ajoute trois permissions. La première "Create-Meeting" spécifie que seul un initiateur peut créer des réunions. La deuxième "InitiatorMeeting" exprime qu'une réunion ne peut être lue ou modifiée que par un initiateur, pour autant qu'il soit le créateur de cette réunion. Cette restriction est exprimée par la contrainte d'autorisation associée à la permission. La dernière permission "ParticipantMeeting" exprime que les participants peuvent également lire les informations des réunions, pour autant qu'ils fassent partie des personnes invitées à la réunion. Les deux permissions "InitiatorMeeting" et "ParticipantMeeting" sont attachées à une contrainte d'autorisation exprimée en OCL (OCL2, 2012). Nous utilisons le mot clé "Caller" du type String dans les expressions OCL pour faire référence au nom de l'utilisateur. Souvent ces contraintes font le lien entre les informations issues du modèle de sécurité (comme l'utilisateur et ses rôles) et l'état du modèle fonctionnel.

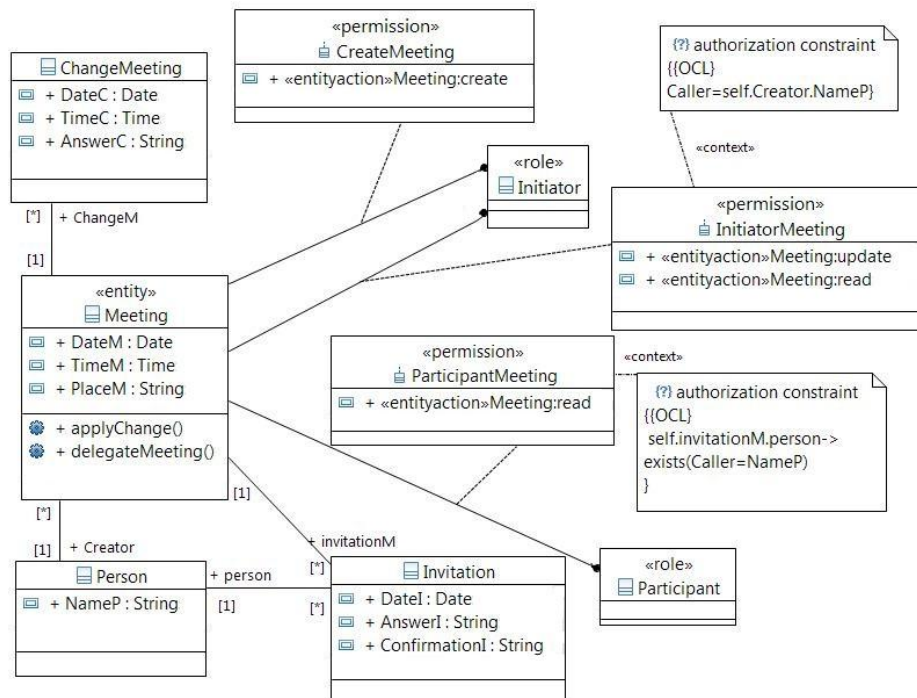


Figure 2. Le diagramme de classes SecureUML du système d'organisation de réunions

3.2. Contrôle d'accès aux opérations

Les actions de type «EntityAction» représentent des opérations abstraites qui donnent lieu à des appels d'opérations concrètes réalisables sur des classes fonctionnelles. Dans le cadre de notre exemple, les actions autorisées au travers des permissions se traduisent comme suit :

- Modification de la classe Meeting (EntityAction Update) : permet d'appeler les setters d'attributs (DateM, PlaceM et TimeM), ainsi que les setters des extrémités d'associations : Creator, invitationM, et ChangeM; et permet d'invoquer les opérations applyChange et delegateMeeting qui sont des opérations de modification.
- Creation de la classe Meeting (EntityAction Create) : permet l'appel au constructeur d'instances de la classe Meeting.
- Lecture de la classe Meeting (EntityAction Read) : permet l'appel à toutes les opérations de lecture d'attributs et d'extrémités d'association de la classe Meeting.

Nous considérons comme critique toute opération d'une classe protégée qui correspond à une action d'une permission. La contrainte d'autorisation associée à la permission exprime une condition nécessaire pour la réalisation des opérations critiques autorisées.

4. Contrôle d'accès aux activités

Cette section montre comment exprimer les permissions sous forme de préconditions associées aux opérations critiques et comment les propager sur les activités d'un processus métier. Cela se fait en trois phases : la première consiste à décrire chaque cas d'utilisation par une activité abstraite, la deuxième permet de spécifier une activité abstraite par une activité concrète et la dernière complète ce diagramme d'activité pour contrôler l'accès à ses actions.

4.1. Activité abstraite

Le diagramme d'activité d'UML 2 (UML2, 2011) fournit un langage de modélisation des workflows qui est utilisé pour décrire le déroulement d'un cas d'utilisation. Ce modèle est composé des nœuds d'activité, des nœuds d'action, des nœuds d'objet et des nœuds de contrôle. Ces nœuds sont reliés par deux types d'arcs pour représenter les flux de contrôle et de données. L'élément principal d'un diagramme d'activité est l'activité. Son comportement est défini par une coordination d'actions. Plusieurs auteurs, comme (Roques, 2006), ont spécifié un cas d'utilisation par un ensemble de séquences d'actions qui représentent des tâches exécutées par des acteurs et qui produisent un résultat observable. Les tâches sont des opérations abstraites qui permettent d'une part de bien visualiser le comportement d'un cas d'utilisation et d'autre part de communiquer facilement et précisément avec les acteurs du SI. Une activité abstraite permet de représenter un cas d'utilisation par un ensemble de tâches en coordination, représentées par des nœuds d'action, qui seront exécutées par les acteurs du cas d'utilisation. Une activité abstraite peut montrer plusieurs scénarios d'exécution. La figure 3 montre un diagramme d'activité qui décrit les deux cas d'utilisation *reply to invitation* et *follow answer* de la figure 1 par deux activités abstraites. Les activités sont placées dans des partitions séparées qui précisent les acteurs responsables de ces tâches.

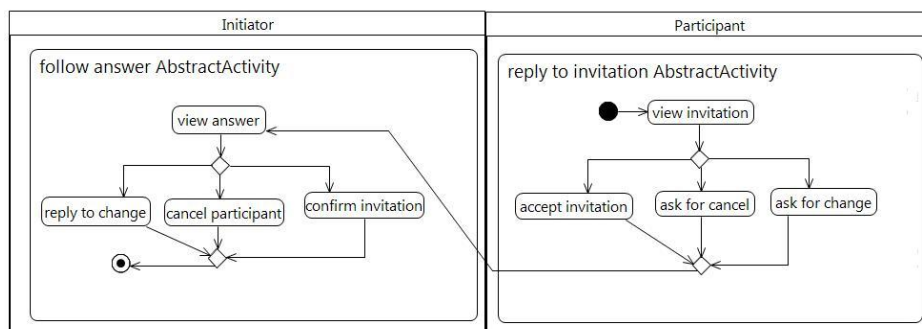


Figure 3. Activités abstraites des cas d'utilisation "reply to invitation" et "follow answer"

4.2. *Activité concrète*

Les activités concrètes sont construites à partir des activités abstraites. Les activités abstraites décrivent le système comme une boîte noire sans détailler les objets qui le composent. Une fois qu'on dispose d'un diagramme de classes, ces diagrammes abstraits peuvent être raffinés en précisant les objets et les opérations qui les réalisent. Une activité concrète décrit une activité abstraite en spécifiant le comportement de ses différentes tâches par une coordination d'un ensemble d'actions concrètes qui font référence à des opérations sur les objets des classes. L'exécution de chaque tâche fait appel à une ou plusieurs opérations concrètes. Une opération peut être, par exemple, une affectation de valeurs à des attributs, un accès à la valeur d'une propriété structurelle (attribut ou terminaison d'association), la création d'un nouvel objet ou lien, . . . Dans la figure 4, chacune des tâches *view answer* et *reply to change* de l'activité abstraite *follow answer* (la partie gauche de la figure 3) est décomposée en une coordination d'actions concrètes. Les tâches *confirm invitation* et *cancel participant* font chacune appel à une seule action concrète.

Les actions concrètes appellent une opération d'une instance de la classe. Ces instances sont définies comme des paramètres de l'activité concrète. Ceux-ci fournissent des entrées nécessaires à l'exécution des actions. Les instances M, I, C des entités (Meeting, Invitation et ChangeMeeting), dans la figure 4, sont indispensables pour l'exécution de l'activité *Follow answer*. Une activité concrète regroupe une famille de scénarios qui seront exécutés par un utilisateur représentant l'acteur du cas d'utilisation décrit par l'activité.

4.3. *Pré-condition de contrôle d'accès*

La séparation des préoccupations encourage à spécifier indépendamment les aspects fonctionnels et sécuritaires d'un système d'information. Cependant, leur interaction doit être prise en considération au niveau des diagrammes d'activités concrètes. Cette section explique comment exprimer une politique de sécurité spécifiée par SecureUML sur les activités concrètes fonctionnelles. Nous allons montrer comment dans la réalisation d'un processus métier, les permissions du diagramme SecureUML seront prises en compte.

Dans le diagramme d'activité de la figure 4, les permissions et leurs éventuelles contraintes d'autorisation sont exprimées par des préconditions. Ces préconditions sont soit associées à l'ensemble du diagramme, soit attachées à des actions concrètes.

La précondition associée à l'ensemble du diagramme, aussi appelée précondition d'activité, est stéréotypée par «PreCondition». Dans la figure 4, celle-ci impose que l'activité soit exécutée par un utilisateur associé au rôle "Initiator". Elle s'exprime en OCL par *Initiator.assignedUser -> includes (User)*. Cette précondition porte sur toutes les actions de l'activité. Ici, elle traduit le fait que le cas d'utilisation "follow answer" est exécuté par l'acteur "Initiator" dans le diagramme des cas d'utilisation

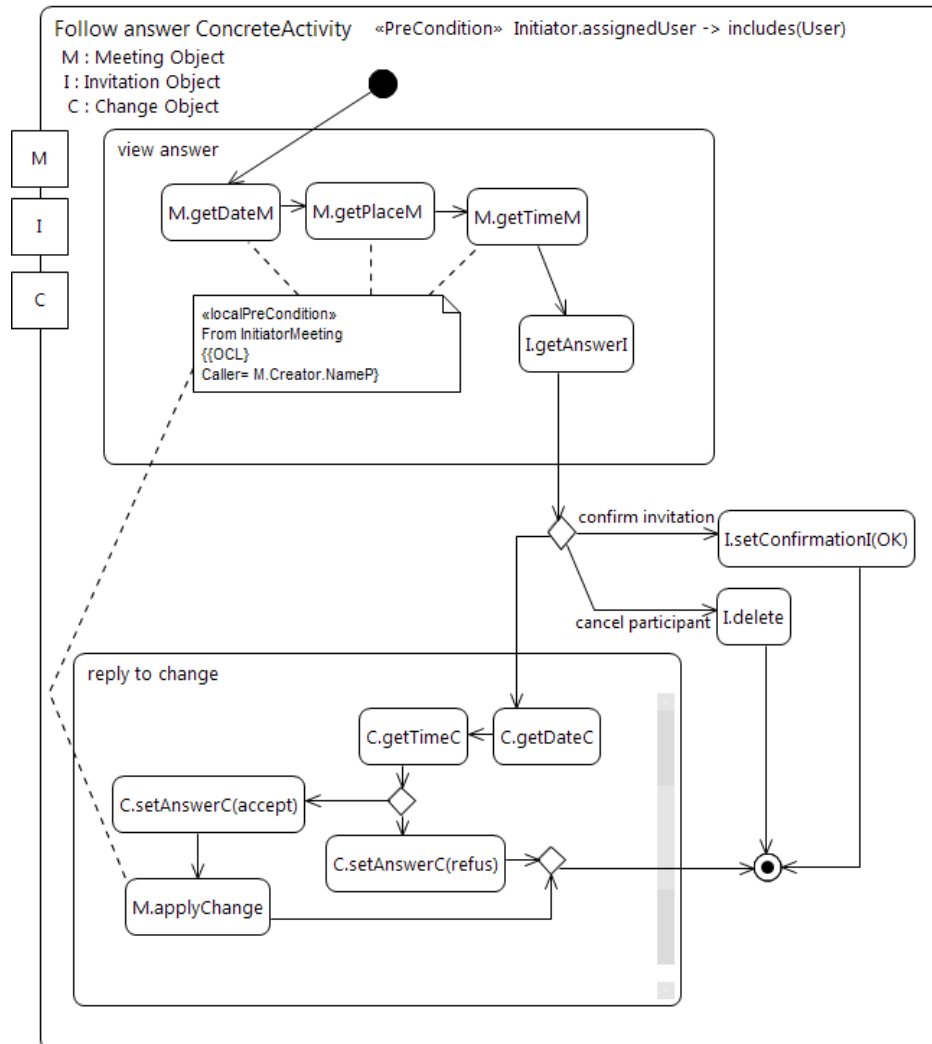


Figure 4. Contrôle d'accès à l'activité concrète " Follow answer "

(figure 1). Nous considérons donc les acteurs associés aux cas d'utilisation comme des rôles.

La précondition d'activité exprime le fait que toutes les actions doivent être exécutées par un utilisateur dans un rôle donné. Cependant, certaines actions impliquent des opérations critiques dont les permissions sont associées à des contraintes d'autorisation. Il faut tenir compte de ces contraintes d'autorisation dans la définition de l'activité. Pour ce faire, des conditions supplémentaires sont exprimées sous la forme de contraintes stéréotypées comme «localPreCondition» et reliées aux actions concrètes(UML2, 2011). Dans la figure 4, cette précondition locale est associée aux

quatre actions *M.getDateM*, *M.getPlaceM*, *M.getTimeM* et *M.applyChange*. Elle garantit que l'utilisateur qui exécute ces actions est le créateur de la réunion: *Caller = M.Creator.NameP*, ce qui correspond à la contrainte d'autorisation de la permission "InitiatorMeeting" dans la figure 2.

Les préconditions d'activité et les préconditions locales définissent dans quelles conditions les actions concrètes doivent être réalisées. On peut également les voir comme des gardes qui sont évaluées lors de l'exécution du diagramme et qui garantissent que la politique de contrôle d'accès est bien respectée. Pour la figure 4, on peut se contenter d'évaluer la précondition d'activité en entrée de l'activité, et la précondition locale avant l'exécution de *M.getDateM* si les conditions suivantes sont respectées:

1. L'utilisateur n'utilise pas d'autre rôle que celui ou ceux prescrits par la précondition d'activité lors de l'exécution du processus.
2. L'utilisateur ne perd pas le droit d'utiliser ces rôles pendant l'exécution de l'activité.
3. La précondition locale reste vraie pendant l'exécution des quatre opérations concernées. Ce qui signifie que les objets et associations concernées ne sont pas modifiés par les actions du diagramme d'activité ou par des activités extérieures qui seraient menées en parallèles de ce diagramme.

Si ces conditions ne sont pas garanties, il est nécessaire de vérifier les gardes plus souvent et d'encapsuler tout ou partie du diagramme dans une ou plusieurs transactions.

5. Extension du méta-modèle des diagrammes d'activité

Notre travail vise à intégrer les concepts de SecureUML au niveau des diagrammes d'activité. Pour ce faire, nous définissons la sémantique des liens entre ces modèles au moyen de leurs méta-modèles respectifs. L'explicitation de ces liens permet, outre la définition de leur sémantique, de définir les contraintes qui garantissent la cohérence entre ces modèles.

5.1. Liens entre méta-modèles

Le diagramme de la figure 5 représente les concepts de SecureUML en lien avec les notions d'activité abstraite et d'activité concrète utiles dans le présent travail. Les concepts de SecureUML sont : User, Role, Permission, Action, et AuthorizationConstraint. Les concepts inhérents aux activités sont : AbstractActivity, ConcreteActivity, Task, OperationAction et LocalPreCondition.

Une activité concrète est enclenchée par un ou plusieurs rôles et est nécessairement issue d'une activité abstraite. L'activité abstraite est réalisée par un ensemble de tâches (méta-classe Task) et l'activité concrète est réalisée par un ensemble d'actions (méta-

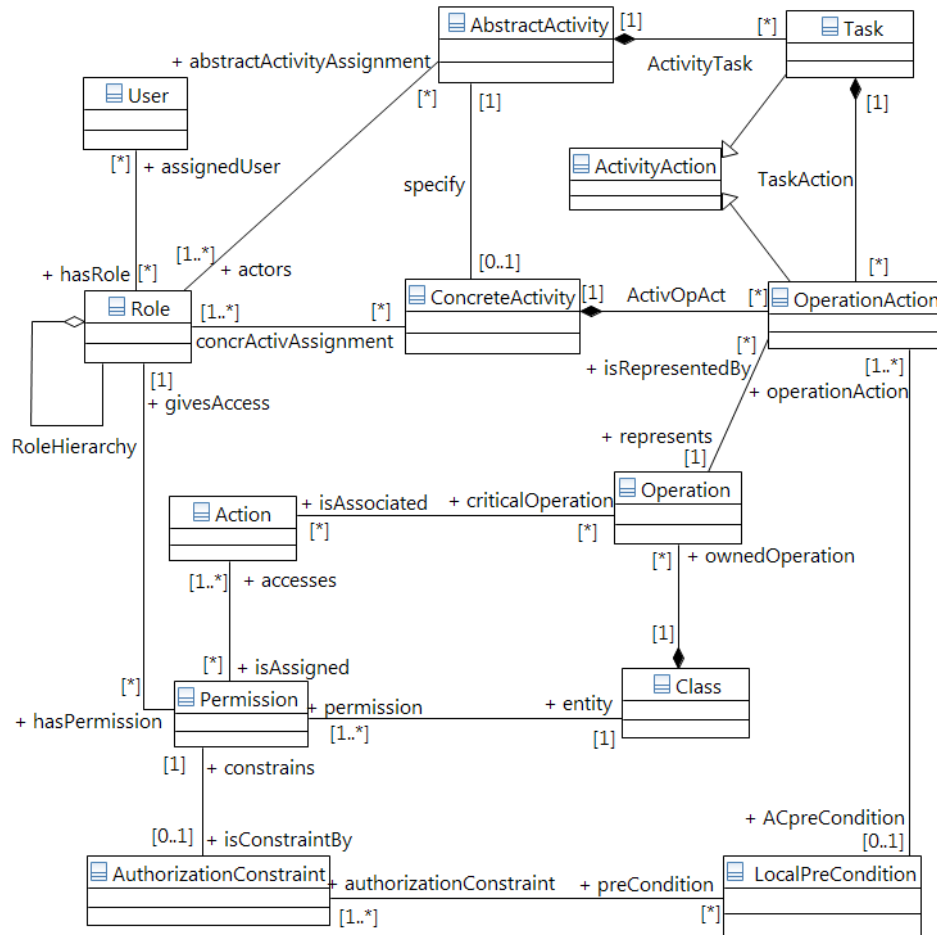


Figure 5. Métamodèle de contrôle d'accès aux activités

classe OperationAction) qui correspondent à des appels d'opérations fonctionnelles. Nous considérons que ces actions sont contenues dans les tâches réalisant l'activité abstraite. Ce lien de composition permet de garantir une certaine traçabilité entre les activités abstraites et concrètes tout en étant en relation avec les opérations fonctionnelles.

Les actions d'une activité concrète (OperationAction) font appel à deux types d'opérations : les opérations critiques auxquelles une ou plusieurs permissions sont associées dans le modèle SecureUML, et les opérations non protégées (comme par exemple *I.setConfirmationI* et *I.delete* pour la classe Invitation). Aucune vérification n'est effectuée pour les opérations non protégées. En revanche, pour les opérations critiques, nous projetons les contraintes d'autorisation qui leur sont associées, sous forme de préconditions des actions OperationAction. De ce fait, l'appel d'une opéra-

tion fonctionnelle dans une activité concrète devient préconditionné par la satisfaction de contraintes d'autorisation.

5.2. Règles de cohérence

Dans cette sous-section nous spécifions au moyen d'expressions OCL les règles de cohérence que les instances de notre méta-modèle doivent respecter. Pour ce faire, nous définissons d'abord deux requêtes OCL, `isCritical()` et `criticalOps()` qui permettent respectivement d'indiquer si une opération donnée est critique, et de récupérer pour une activité concrète l'ensemble de ses opérations critiques. Par exemple, les actions associées à des opérations critiques de l'activité *Follow answer* de la figure 4 sont : *M.getDateM*, *M.getPlaceM*, *M.getTimeM* et *M.applyChange*. Les autres actions sont en relation avec des opérations non protégées.

```
Context Operation::isCritical():Boolean
Body: self.isAssociated → notEmpty()
```

```
Context ConcreteActivity::criticalOps():Set(Operation)
Body: self.ActivOpAct.represents → select(o:Operation | o.isCritical())
```

5.2.1. Conformité des rôles

L'invariant *RoleConformance* ci-dessous permet de garantir que tous les rôles autorisés à enclencher une activité concrète disposent de permissions dans le modèle SecureUML leur donnant le droit de réaliser toutes les opérations critiques de l'activité.

```
Context ConcreteActivity inv RoleConformance :
self.role → (forall(r : Role | r.hasPermission.accesses.criticalOperation
→ includesAll(self.criticalOps())))
```

Il faut noter que cet invariant impose que le plus faible rôle associé à l'activité ait les droits sur toutes les permissions. Dans la suite de nos travaux, nous étudierons d'une part la possibilité de relâcher cette contrainte en s'assurant que l'utilisateur ait des droits pour chaque opération, mais en lui permettant d'utiliser des rôles différents et complémentaires. Il serait également intéressant d'exprimer explicitement le changement de rôle d'un utilisateur lors de l'exécution d'une activité concrète.

Dans notre démarche, les rôles associés à une activité concrète correspondent aux acteurs définis dans l'activité abstraite qu'elle spécifie. Nous établissons ainsi l'invariant suivant pour garantir un usage de rôles adéquat durant le processus de spécification allant des activités abstraites aux activités concrètes.

```
Context ConcreteActivity inv RoleSpecification :
self.concrActivAssignment = self.specify.actors
```

5.2.2. *Traçabilité des contraintes d'autorisation*

Les contraintes d'autorisation définies au niveau des permissions sont projetées au niveau des activités concrètes sous forme de préconditions d'actions. Ces préconditions, ainsi issues des permissions *SecureUML*, permettent de contrôler l'exécution des actions d'activité concrète. Elles peuvent soit autoriser le déclenchement des actions d'opération critiques si les contraintes d'autorisation sont vérifiées ou bien bloquer l'exécution de l'activité dans le cas contraire. Le lien entre les méta-classes *Precondition* et *AuthorizationConstraint* permet d'avoir une traçabilité entre ces deux notions.

6. Conclusion et perspectives

Le contrôle d'accès est devenu un souci majeur dans le développement des SI. Plusieurs domaines de l'activité économique et sociale sont maintenant sujets à des lois et des normes très strictes au niveau de la sécurité des données (Milhau, 2011). Cet article propose une approche de spécification d'une politique d'autorisation RBAC au niveau du Workflow des activités d'un processus métier. Ceci est réalisé en utilisant des règles de contrôle d'accès, exprimées dans une vue statique via le profil *SecureUML*, pour contrôler l'accès des utilisateurs aux actions des activités.

La séparation entre les préoccupations fonctionnelles et sécuritaires est au cœur de notre approche. Cette séparation des préoccupations vise à maîtriser la complexité du système, et peut également se montrer utile quand il s'agit de réaliser la sécurisation d'un système d'information pré-existant. De plus, elle facilite l'évolution de la politique de contrôle d'accès. Dans notre approche, les aspects fonctionnels sont exprimés dans les cas d'utilisation, le diagramme de classes et les diagrammes d'activité abstraits, et leur raffinement dans l'enchaînement des opérations concrètes. Les aspects sécuritaires sont exprimés statiquement, sous forme de permissions associées à des rôles, pour chaque classe protégée du diagramme de classes. A cet effet, nous nous basons sur le profil *SecureUML*. Les aspects sécuritaires sont ensuite exprimés dynamiquement par des préconditions associées localement ou globalement aux activités concrètes. En outre, les contraintes d'autorisation permettent des interactions entre les aspects fonctionnels et sécuritaires. Elles peuvent être spécifiées dans la vue statique et prendre la forme de gardes dans la vue dynamique.

Enfin, nous avons proposé un métamodèle, associé à des contraintes OCL, qui précise les relations et les contraintes entre la vue statique de la politique de contrôle d'accès, et son expression dynamique dans les activités concrètes. Ce métamodèle doit encore évoluer pour permettre plus de flexibilité dans l'utilisation des rôles. Il sera alors possible de définir un outillage associé à un environnement UML2, qui vérifie la conformité de nos diagrammes par rapport à ce métamodèle. Nous avons, en outre, appliqué notre démarche sur le système d'organisation de réunions, et développé ainsi tous les cas d'utilisation de la figure 1. Les règles de contrôle d'accès ont ainsi été exprimées sur les différentes activités du système. Cette étude de cas nous a permis de valider par la pratique le présent travail et montrer l'intérêt d'exprimer des règles de

contrôle d'accès au niveau des processus métiers. Des études de cas de plus grandes tailles seront étudiées dans de futurs travaux.

La démarche que nous proposons construit graduellement une politique de contrôle d'accès et identifie les gardes qui la mettront en oeuvre dynamiquement dans le déroulement des activités concrètes. Cette représentation est proche des scénarios d'exécution et facilite l'implémentation de la politique de sécurité dans des plateformes logicielles.

La suite de ce travail sera principalement consacrée à la mise en oeuvre de cette approche et du méta-modèle associé dans des outils. La plateforme B4MSecure² (Ledru Y. et al, 2015), qui supporte une variante de SecureUML et permet la traduction de ces modèles vers des spécifications formelles exprimées en B (Abrial, 1996), est un candidat naturel pour cette mise en oeuvre. Cette plateforme supporte aujourd'hui la vue statique (diagramme de classes et permissions). Il s'agirait d'y intégrer les vues dynamiques (diagrammes d'activités abstrait et concret), de vérifier la conformité de ces diagrammes à notre méta-modèle, et ensuite de traduire ces vues dynamiques en B. La traduction de ces vues dynamiques fournira une base intéressante pour l'animation des spécifications B et la validation de l'ensemble des modèles.

Bibliographie

- Abrial J.-R. (1996). *The B-book: assigning programs to meanings*. Cambridge University Press.
- Ahn G., Sandhu R., Kang M., Park J. (2000). Injecting RBAC to secure a web-based workflow system. In *the 5th acm workshop on role-based access control*, p. 1-10. New York, NY, USA, Morgan Kaufmann Publisher.
- Alghathbar K. (2012). Representing access control policies in use cases. *International Arab Journal of Information Technology*, vol. 9, n° 3.
- ANSI. (2004). Role based access control. *American national standard for information technology*, vol. 359, n° 2004, p. 1-47.
- Basin D. A., Clavel M., Doser J., Egea M. (2009). Automated analysis of security-design models. *Inf. & Softw. Technology*, vol. 51, n° 5, p. 815-831.
- Basin D. A., Doser J., Lodderstedt T. (2006). Model driven security: From UML models to access control infrastructures. *ACM Trans. Softw. Eng. Methodol.*, vol. 15, n° 1, p. 39-91.
- Bertino E., Ferrari E., Atluri V. (1999). The specification and enforcement of authorization constraints in workflow management systems. *ACM Transactions on Information and System Security*, vol. 2, n° 1, p. 65-104.
- BPMN2. (2011). *Business Process Modeling Notation (BPMN) Version 2.0*. Object Management Group. (<http://www.omg.org/spec/BPMN/2.0/formal-11-01-03.pdf>)
- Feather M., Fickas S., Finkelstein A., Lamsweerde A. (1997). Requirements and specification exemplars. *Automated Software Engineering*, vol. 4, n° 4, p. 419-438.

2. <http://b4msecure.forge.imag.fr/>

- Ferraiolo D., Kuhn D., Chandramouli R. (2003). *Role-based access control*. Artech House.
- Gaaloul K. (2010). *Une approche sécurisée pour la délégation dynamique de tâches dans les systèmes de gestion de workflow*. Thèse de doctorat. Nancy, France.
- Geambasu C. (2012). BPMN vs. UML activity diagram for business process modeling. *Accounting and Management Information Systems*, vol. 11, n° 4, p. 637–651.
- Jurjens J. (2010). *Secure systems development with UML*. Berlin, Heidelberg, Springer-Verlag.
- Kandala S., Sandhu R. (2002). Secure role-based workflow models. *The International Federation for Information Processing*, vol. 87, n° 2002, p. 45-58.
- Ledru Y. et al. (2015). Validation of IS Security Policies featuring Authorisation Constraints. *International Journal of Information System Modeling and Design (IJISMD)*, vol. 6, n° 1.
- Ma G., Wu K., Zhang T., Li W. (2011). A flexible policy-based access control model for workflow. *Computer Science and Automation Engineering*, vol. 2, n° 2011, p. 533 - 537.
- Milhau J. (2011). *Un processus formel d'intégration de politiques de contrôle d'accès dans les systèmes d'information*. Thèse de doctorat. Paris, France.
- Murata T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, vol. 77, n° 4, p. 541 - 580.
- OCL2. (2012). *Object Constraint Language (OCL) Version 2.3.1*. Object Management Group. (<http://www.omg.org/spec/OCL/2.3.1/PDF/>)
- Roques P. (2006). *UML 2 par la Pratique*. Paris, Eyrolles.
- Strembeck M., Mendling J. (2011). Modeling process-related RBAC models with extended UML activity models. *Information and Software Technology*, vol. 53, n° 2011, p. 456-483.
- UML2. (2011). *Unified modeling language: Superstructure(version 2.4)*. Object Management Group. (<http://www.omg.org/spec/UML/2.4/Superstructure/ptc-10-11-14.pdf>)
- Wainer J., Barthelmess P., Kumar A. (2003). W-rbac.a workflow security model incorporating controlled overriding of constraints. *International Journal of Cooperative Information Systems*, vol. 12, n° 4, p. 455-486.
- Wainer J., Kumar A., Barthelmess P. (2007). DW-RBAC: A formal security model of delegation and revocation in workflow systems. *Information Systems*, vol. 32, n° 2007, p. 365-384.
- WFMC. (1999). *Workflow management coalition Terminology and glossary*. Workflow Management Coalition. (http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf)

Un méta-modèle et un langage orientés patient pour la représentation de protocoles de soins auto-adaptatifs

Amira Derradji

1. Univ. Grenoble Alpes, LIG
220 rue de la chimie
F-38000 Grenoble
amira.derradji@imag.fr

2. ARCAN SYSTEMS
4 rue Edison
F-69500 Bron
amira.derradji@arcan.fr

RÉSUMÉ. La prise en charge médicale à domicile est actuellement très répandue. Les patients chroniques représentent une cible pertinente pour le suivi à domicile car (i) leur protocole de soins est précis, clair (dialyse péritonéale, diabète type 1), (ii) aucun séjour hospitalier n'est requis (hors situation d'urgence), (iii) les patients connaissent parfaitement bien leur maladie. Le problème majeur est que le protocole de prise en charge de maladies chroniques est souvent sous forme orale. Quand il est écrit, il est destiné aux tutelles ou à la sécurité sociale à des fins administratives. Nous proposons une recherche destinée à (i) réaliser un protocole de soins informatisé capable d'intégrer des données médicales, gérer des interventions entre patient et médecin par l'intermédiaire d'alertes par exemple, (ii) rendre ce protocole auto-adaptatif en introduisant l'expertise du patient préalablement filtrée pour améliorer le soin ou la prise en charge. Cet article présente un méta-modèle et un langage de représentation des protocoles. Des maquettes d'IHM (Interface Homme Machine) viennent compléter l'ensemble.

ABSTRACT. Home medical care is currently widespread. Chronic patients represent a good target for home-care because : (i) the care protocol is precise, clear (e.g peritoneal dialysis, type 1 diabetes), (ii) patients don't need hospital stay (except medical emergency), (iii) patients perfectly know their chronic disease. The main problem is that the care protocol is often oral and even in the case of written-based one, it is dedicated to government agencies or social security. Then, two problems arise : the lack of digital support for chronic diseases protocols and the impossibility to introduce in the care process the patient's expertise. We propose a research to address these two points : (i) a digital care protocol to help both patients and physicians to follow the care process, gather medical data, send alerts and help the patient in the today life and keep every care's stakeholder informed, (ii) an adaptive digital care protocol that introduce

filtered patients expertise to improve the protocol and improve the care. In this paper, we propose a metamodel and a language to formalize this research. We illustrate the proposition with some screenshots.

MOTS-CLÉS : e-santé, protocole de soins, patient expert, application logicielle médicale

KEYWORDS: e-health, care protocol, expert patient, medical software application

1. Introduction

Le transfert de soins de l'hôpital vers le domicile du patient a vu le jour depuis de nombreuses années avec la naissance du concept américain "HomeCare". En France en particulier, cette solution a été adoptée pour diverses raisons démographiques, économiques, sociales, politiques, etc. En revanche, il reste à l'heure actuelle un enjeu majeur pour assurer la continuité des soins et améliorer la qualité de vie du patient à l'extérieur du milieu hospitalier. Les maladies chroniques en particulier sont essentiellement vécues et prises en charge à domicile au jour le jour par le patient et éventuellement son entourage (famille, amis, voisins...). Quand la prise en charge n'est pas optimale, elle peut engendrer une perte d'autonomie voire une aggravation de la maladie.

Dans ce contexte, le patient est amené à suivre un protocole de soins défini et personnalisé par une équipe soignante. Ce protocole décrit les actions à réaliser à domicile afin de contrôler et d'améliorer sa santé. Le protocole de soins est souvent donné sous forme verbale, plus rarement par écrit (texte ou graphique simple). Cependant, le suivi à la maison ne suffit pas, il est indispensable que l'équipe médicale ait les moyens de réaliser des suivis réguliers à distance. Ce suivi est effectué aujourd'hui via des dispositifs de e-santé mais le plus souvent, de simples cahiers de liaison "papier" sont à remplir à la main par les différents intervenants à domicile.

Durant sa vie avec la maladie, le patient est souvent confronté à des imprévus non définis dans le protocole initial. Certains de ces imprévus sont dûs au patient (vécu quotidien...) et d'autres sont subis. La représentation actuelle du protocole n'est pas adaptée pour prendre en compte automatiquement ces imprévus et les intégrer dans le protocole initial s'ils permettent d'améliorer la santé ou la prise en charge.

Pourtant, le patient chronique connaît généralement parfaitement sa maladie, il a acquis et développé des connaissances fortes et a pu tirer une expertise très utile pour lui-même et l'équipe médicale. Il est ainsi une partie intégrante de sa propre santé, et de ce fait, est considéré comme un "patient expert".

L'idée générale de notre travail consiste à prendre en compte et intégrer l'expertise du patient chronique dans le protocole afin d'améliorer son suivi médical en lui offrant la possibilité de signaler tout imprévu (subi ou réalisé) non défini dans le protocole initial. L'équipe médicale pourra ensuite interpréter la situation, et mettre en oeuvre une action ciblée.

Les protocoles de soins devront être personnalisés et faciles à comprendre par le patient. Le patient devra pouvoir signaler des imprévus dans le protocole de soins. Ces imprévus devront être ensuite filtrés pour ne retenir que ceux qui ont médicalement un sens. Le protocole de soins sera adapté semi-automatiquement en fonction des imprévus retenus et réinjectés dans le système afin de permettre l'auto-enrichissement du protocole. Plusieurs défis devront être résolus tout au long de notre recherche tels que : Comment construire des protocoles de soins personnalisés et faciles à comprendre par le patient? Comment intégrer les imprévus dans le protocole de soins? Comment fil-

trer les imprévus pour ne retenir que ceux qui ont médicalement un sens? Comment automatiser le protocole de soins et l'adapter en fonction des imprévus? Comment permettre l'auto-enrichissement du protocole de soins?

De façon plus précise, l'objectif de notre travail est donc de réaliser un outil informatique simple et intuitif pour le suivi des maladies destiné aux patients ainsi qu'aux professionnels de santé qui permette :

- la construction simple de protocoles de soins;
- l'intégration de l'expertise du patient (les imprévus);
- le filtrage des imprévus;
- l'adaptation du protocoles de soins semi-automatiquement en cas de changement (après filtrage).

Cet outil présentera également un intérêt réel pour l'amélioration de la connaissance médicale car toutes les informations renseignées par le patient seront recueillies et restituées pour répondre à de nouvelles situations pathologiques et aider le professionnel de santé à mieux s'informer.

L'approche générale proposée dans notre travail s'articule ainsi autour de trois phases principales schématisées par la figure 1 : la première phase consiste à construire les protocoles de soins et faire participer le patient dans le protocole à travers le signalement d'imprévus via des interfaces IHM. La seconde phase définit le processus de filtrage qui permet de trier les imprévus et de ne garder que ceux qui peuvent aider le patient à avoir un meilleur suivi médical en dehors du cadre clinique. La dernière phase repose sur l'adaptation du protocole de façon semi-automatique afin d'être en conformité avec le vécu du patient. Dans cet article, nous développons uniquement nos propositions pour la première phase. Les phases 2 et 3 ainsi que le développement logiciel feront l'objet de travaux futurs.

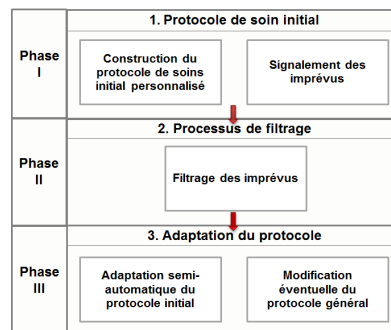


Figure 1. Approche globale de notre proposition

Cet article est organisé comme suit. La section 2 présente les travaux connexes. La section 3 détaille la notion de protocole de soins. La section 4 décrit le méta-modèle élaboré pour la description de protocoles de soins. La section 5 présente le langage que nous proposons pour la représentation de protocoles de soins à destination des patients et des médecins. La section 6 montre une partie de l'implémentation des interfaces de

la solution logicielle. Enfin, la section 7 présente la conclusion et les perspectives de nos travaux.

2. État de l'art

La santé est une ressource de la vie quotidienne, et non le but de la vie (OMS, 1986), et parce que la santé des individus est une priorité, il s'avère indispensable d'apporter des nouvelles solutions technologiques permettant d'offrir de meilleurs soins et d'améliorer les conditions de la prise en charge médicale notamment à domicile. En effet, le domaine de la santé a connu, et connaît toujours, un grand essor notamment avec l'évolution de l'informatique qui joue un rôle essentiel dans l'aide et l'amélioration de la vie quotidienne des patients. Dans (Picard, Salgues, 2008), les auteurs ont démontré ce rôle important à travers une enquête menée sur la participation des TIC (Technologies de l'information et de la communication) dans l'autonomie et la santé des patients et des handicapés dans leurs lieux de vie.

Aujourd'hui, l'usage des TIC dans la santé se manifeste par le développement des systèmes d'information de soins, de dossiers médicaux informatisés (Bernonville *et al.*, 2013), la représentation de l'information médicale (Flory *et al.*, 2006), etc. Il se traduit également par la mise en place d'outils de télémédecine pour permettre une prise en charge distante et partagée notamment avec la parution du décret de télémédecine du 19 octobre 2010 qui offre un cadre réglementaire aux actes réalisés par la télémédecine et décrit les conditions de mise en place d'un nouveau projet médical (Collectif-DGOS, 2012).

De nombreuses solutions permettant la prise en charge des patients à domicile sont mises en place. Dans (Zefouni, 2012), ces solutions sont classées en deux catégories. Certaines sont centrées sur *la gestion des activités de soins réalisées par les professionnels de santé*, notamment les travaux de (Bricon-Souf *et al.*, 2005), menés au CERIM¹ où les auteurs ont proposé une architecture intégrant les connaissances cognitives de la prise en charge. D'autres solutions sont orientées *processus de soins*, elles sont basées sur des services de télé-santé (télé-consultation, télé-assistance, télé-surveillance et télé-expertise). Parmi elles deux solutions sont interactives avec le patient : les projets de télé-consultation et de télé-surveillance à domicile. Les données recueillies sont issues de dispositifs médicaux (implantés ou non) et se présentent sous forme de mesures (glycémie, pouls, etc.), de signaux physiologiques (électrocardiogramme -ECG-, électro-encéphalogramme -EEG-, etc.) ou d'imagerie fixe et dynamique (image, vidéo) (Collectif-DGOS, 2012).

Le domaine médical a la particularité d'être très complexe en comparaison avec les autres domaines métiers (industrie, banque, assurance). Cette particularité réside principalement dans la personnalisation des soins pour chaque patient selon plusieurs facteurs pathologiques, médicamenteux, sociaux, etc. Dans le contexte de notre re-

1. Centre d'Études et de Recherche en Informatique Médicale

cherche, un autre problème est mis en exergue : celui de la continuité des soins à distance. Les systèmes de gestion de workflow (WFMC, 1999) qui sont un type spécial de systèmes de BPM (Business Process Management) semblent appropriés à l'amélioration de cette prise en charge. En effet, les solutions de BPM ont été mises en place pour la télémédecine (Ilahi, Ayachi Ghannouchi, 2013) et pour l'amélioration et la gestion des processus de base (administration, gestion, finance...) dans les établissements de soins. Les systèmes de workflows ont été également utilisés pour la gestion des hôpitaux (Dadam, Reichert, 2000), pour l'aide au diagnostic (Ardissono *et al.*, 2005) ou pour le traitement de pathologies.

Plusieurs projets autour de l'utilisation de workflow dans le secteur médical ont été proposés. Dans (Greiner *et al.*, 2005), les auteurs se sont intéressés à l'adaptation dynamique de protocole de chimiothérapie en cas d'exception (infection, toxicité, etc.) chez les patients concernés. Ils proposent la solution Adaptflow qui est basée sur les règles E-C-A et une base de connaissance. Les auteurs de (Leonardi *et al.*, 2007) proposent une solution basée sur le serviceflow (workflow basé sur les services) et les ontologies afin de traiter la coordination entre les différents intervenants appartenant à plusieurs unités organisationnelles. Cette solution a été mise en oeuvre pour la gestion du diabète. Dans (Mans, 2011), l'auteur s'intéresse aux processus organisationnels de santé. Il propose une extension d'un système de gestion de workflow avec deux nouvelles fonctionnalités : une fonctionnalité pour la planification du travail effectué par les utilisateurs du workflow de la santé basée sur un calendrier, et une autre fonctionnalité pour le traitement inter-organisationnel de workflows.

La majorité de ces travaux se concentrent sur la flexibilité des workflows et la coordination inter-organisationnelle des intervenants. De plus, ils ne traitent qu'une pathologie (diabète, etc.) -patient monopathologique- (Han *et al.*, 2006), alors qu'un patient chronique peut souvent souffrir de plusieurs maladies -patient polypathologique-. Quant au problème de la personnalisation des protocoles de soins, il est souvent traité avec des ontologies ou des bases de connaissances. Le tableau de la figure 2 résume ces travaux selon les processus ciblés par ces solutions et leurs objectifs poursuivis.

Travaux	Objectif			Processus ciblé	
	Flexibilité	Personnalisation	Coordination inter-organisationnelle	Processus de soins	Processus inter-organisationnel
(Greiner <i>et al.</i> , 2005),	√	√		√	
(Leonardi <i>et al.</i> , 2007)			√	√	
(Mans, 2011)	√	√			√

Figure 2. Les projets de workflow dans le domaine médical

Dans notre travail, nous nous intéressons à l'utilisation de la technologie du workflow dans la prise en charge médicale à domicile dans le but de réaliser et automatiser des workflows de soins personnalisés pour des patients chroniques (éventuellement

pour des patients chroniques polyopathologiques) et capables d'être flexibles et adaptable en fonction des imprévus signalés par les patients.

Dans la section suivante, nous nous focalisons particulièrement sur la représentation du protocole de soins dans le domaine médical.

3. Protocole de soins

Le terme "protocole de soins" se réfère à plusieurs définitions dans le domaine médical : (1) un formulaire de prise en charge défini par l'assurance maladie² ouvrant droit à un remboursement à 100% du tarif conventionnel aux patients chroniques, (2) un descriptif des techniques à appliquer et/ou des consignes à observer dans certaines situations de soins ou pour l'administration d'un soin par les professionnels de santé³, (3) un document informatif disponible sur internet principalement ou remis par le médecin qui a pour objectif d'informer le patient sur les principaux éléments du traitement et du suivi (ex. le guide du diabète de type 1 de l'adulte réalisé par l'HAS⁴ en coopération avec les professionnels de santé et les associations de patients directement concernées) dans le but de mieux informer les patients et de leur fournir des informations importantes sur leur maladie.

Dans notre cas, nous nous intéressons à un protocole défini par une équipe médicale, mais qui doit pouvoir être consulté, compris et modifié par le patient (ou ses aidants) à domicile. La troisième définition est, en cela, plus proche de nos préoccupations. Mais la forme dans laquelle le contenu du protocole est représenté n'est pas forcément complète ni facile à comprendre par les patients. À titre d'exemple, la figure 3 illustre un exemple de protocole de soins pour l'hypoglycémie chez un diabétique⁵.

De plus, chaque patient est unique, et la définition de son protocole de soins doit être personnalisée et appropriée à ses soins et traitements individuels, parfois même à ses souhaits et contraintes personnelles, ce qui n'est pas le cas avec les guides d'informations ou les référentiels de protocoles qui fournissent des informations très, voire trop générales. Au contraire, dans notre approche, nous souhaitons des protocoles de soins personnalisés simples à comprendre et sur lesquels le patient et l'équipe médicale pourront signaler tous les imprévus subis lors du déroulement du protocole.

Dans des travaux antérieurs (Derradji, 2014) nous avons étudié la possibilité de représenter un protocole de soins comme un processus, constitué d'un ensemble d'actions à réaliser par le patient ou l'équipe médicale. Nous avons ainsi constaté qu'une représentation graphique du protocole en utilisant des formalismes tels que UML (Morley *et al.*, 2006), BPMN (OMG, 2010) présentait l'avantage d'être plus concis

2. www.ameli.fr

3. Cours de soins infirmiers : quesnel.stephan.free.fr

4. Haute Autorité de Santé : <http://www.has-sante.fr/>

5. <http://www.diabhainaut.com/>

HYPOGLYCEMIE : Taux inférieur à 0,6 g/l voir 0,8g si hyperglycémies prolongées antérieures

SIÈNES

- ↳ Variable d'un sujet à autre, toujours les mêmes pour un même patient.
- ↳ Irritabilité, vision floue, fatigue, comportement inhabituel, trouble parole, somnolence, palpitations, pâleur, **sueurs**, **tremblements**, faim, céphalées.

ACTION

- ↳ Arrêt de l'activité et s'asseoir.
- ↳ Se resucrer avec 3 sucres soit 15 g de sucre ou équivalents jus de fruit, 1 càS de confiture ou de miel (éviter biscuits et chocolat).
- ↳ Tout de suite après faire glycémie dans les 15 mn suivant resucrage.
- ↳ Contrôler de nouveau la glycémie 30 mn plus tard si reste bas, reprendre 15 g de sucres et un sucre lent.
- ↳ **Si inconscient glucagon en IM.**
- ↳ Rechercher la cause de hypo.
 - quantité féculent insuffisant repas sauté
 - activité physique plus importante
 - stress infection
 - médicaments
 - erreur dans traitement etc
- ↳ Prévoir ultérieurement une adaptation du traitement ou alimentation en cas d'activité plus importante

Figure 3. Exemple d'un protocole de soins pour l'hypoglycémie chez un diabétique

qu'une représentation textuelle. Cependant, de tels langages sont mal adaptés pour des utilisateurs finaux (professionnels de santé, patient). En effet, ils sont très complexes et leur utilisation nécessite d'avoir des compétences particulières. Notre proposition consiste donc à proposer un langage graphique spécifique au domaine médical adapté aux médecins ainsi qu'à leurs patients. Ce langage devra être simple, intuitif et facile à comprendre. Or, un langage spécifique au domaine ou DSL (Domain Specific Language) doit être formalisé par une syntaxe abstraite (un méta-modèle) et une syntaxe concrète (un langage de représentation) (Fondement, Baar, 2005), (Dupuy-Chessa, 2006). La syntaxe abstraite capture le vocabulaire et la taxonomie (i.e. les concepts) du langage. La syntaxe concrète consiste en la représentation des éléments du langage, elle peut être graphique ou textuelle.

Les sections suivantes présentent nos propositions pour une syntaxe abstraite et une syntaxe concrète d'un langage de description et de représentation des protocoles de soins orientés patient.

4. Un méta modèle du protocole de soins

Le méta-modèle proposé décrit et formalise les concepts principaux des protocoles de soins de différentes maladies chroniques. Ce méta-modèle a été construit par étude de différents protocoles de soins desquels nous avons extrait les concepts principaux ainsi que les relations entre ces concepts. La complétude du méta-modèle obtenu présenté en figure 4 est en cours de validation auprès de professionnels de santé.

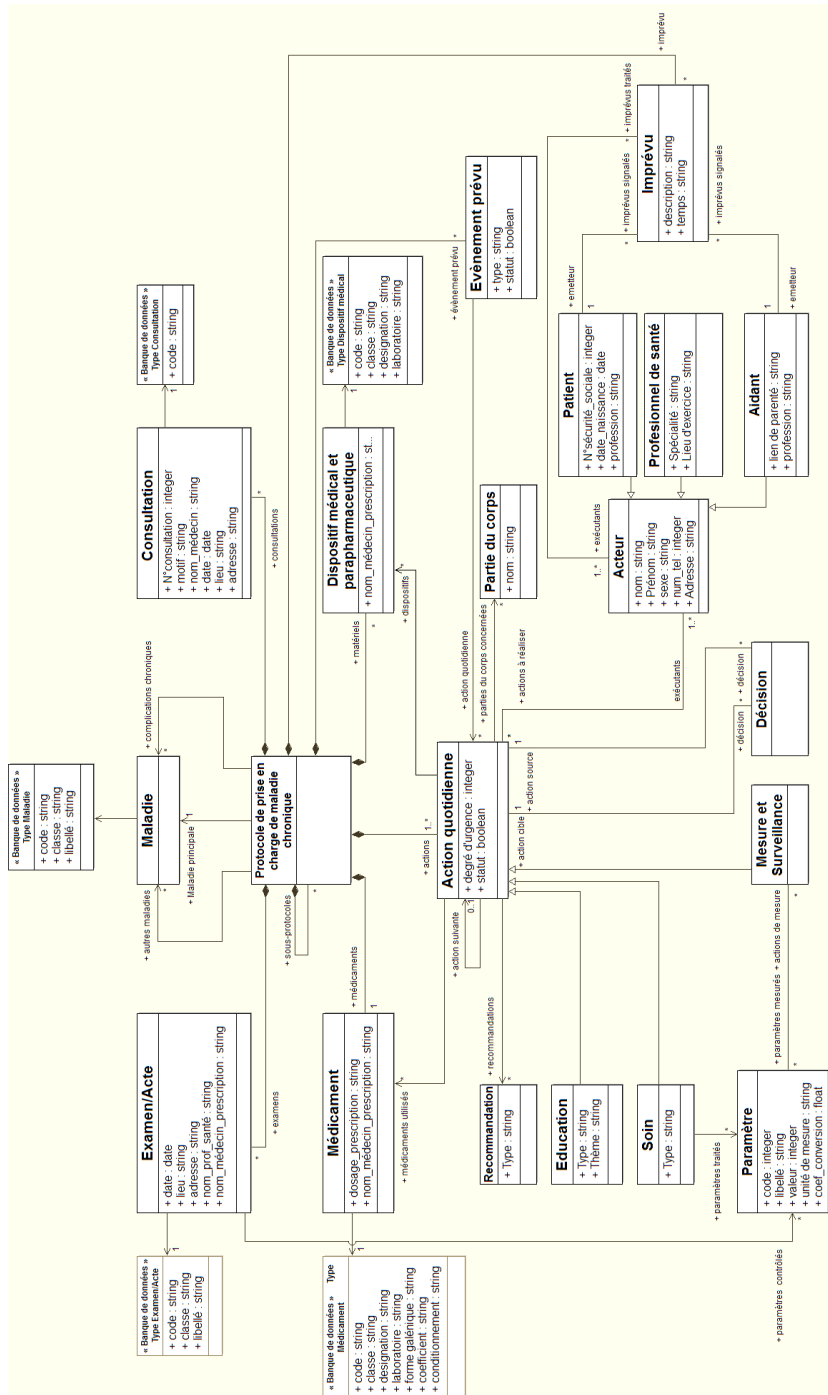


Figure 4. Méta-modèle du protocole de soins

La construction de protocoles de soins personnalisés au patient se fera donc conformément à ce méta-modèle.

La figure 4 présente le méta-modèle destiné à représenter des protocoles de soins associés à une *maladie chronique principale* (ex. le diabète), éventuellement à des *complications chroniques* causées par la maladie principale (ex. la néphropathie diabétique, l'une des complications chroniques du diabète) ou à d'*autres maladies* indépendantes de la maladie principale (ex. la grippe). Les complications chroniques et les autres maladies sont importantes car sont susceptibles de modifier le protocole de soins initial. Le protocole de soins initial est composé d'un ensemble de séquences d'*actions quotidiennes* que le patient a à réaliser le plus souvent à l'aide de *dispositifs médicaux et parapharmaceutiques* (ex. les aiguilles à insuline) et souvent reliées à des prises de *médicaments* (ex. l'insuline). Le patient doit également effectuer régulièrement des *examens/actes* et des *consultations*. Certaines informations seront extraites des *bases de données* existantes telles que la base VIDAL des médicaments et dispositifs médicaux et parapharmaceutiques ou encore une classification médicale comme la CIM (Classification Internationale des Maladies).

Les *actions quotidiennes* sont effectuées principalement par le *patient* dans le cadre de l'exécution de son protocole. Parfois, ces actions nécessitent l'intervention d'autres intervenants : les *aidants* ou les *professionnels de santé*. Une action en général, concerne une *partie du corps* (ex. l'abdomen pour l'injection de l'insuline). Elle peut être attachée à des *recommandations* qui fournissent des informations supplémentaires sur les dispositifs médicaux (ex. "les aiguilles à insuline sont à usage unique"), sur le temps de réalisation de certaines actions (ex. "en cas d'hypoglycémie, contrôler la glycémie 20 à 30 minutes après resucrage"), sur les médicaments (ex. "l'insuline rapide est mieux absorbée au niveau de l'abdomen") ou sur d'autres recommandations d'ordre général (ex. "en cas d'hypoglycémie, prendre 3 à 4 morceaux de sucre" ou encore "ne pas trop se resucrer pour ne pas entraîner une hyperglycémie", etc.).

Une *action quotidienne* peut être spécialisée en trois classes. Une *action de mesure et surveillance* consiste à contrôler des *paramètres* vitaux souvent à l'aide de *dispositifs médicaux* (ex. "contrôler la glycémie à l'aide du glucomètre"). Une *action de soin* consiste à réaliser des soins préventifs, curatifs et palliatifs (ex. "prendre un *médicament* ou injecter l'insuline"). Nous pouvons également trouver certains soins réalisés dans le but de traiter les paramètres vitaux obtenus à partir des actions de mesure et surveillance (ex. l'action curative "se resucrer" aura lieu si la valeur du paramètre "glycémie" est basse). Enfin une *action d'éducation* représente toute action qui permet au patient de réaliser les actions de mesure et surveillance ainsi que les actions de soin dans les meilleures conditions et le conseille sur les bonnes pratiques pour vivre le mieux possible sa maladie (ex. "se laver les mains avant le contrôle de glycémie et l'injection de l'insuline"). Les actions sont reliées entre elles par un flux *séquentiel* (une action peut être suivie par une autre action) car nous considérons dans les protocoles de soins à domicile que les actions quotidiennes ne peuvent être exécutées que d'une façon séquentielle par le patient ou ses aidants (ex. "contrôler la glycémie" et "se resucrer" ne peuvent pas avoir lieu en même temps). Les actions peuvent aussi

être suivies par des *décisions* qui impliquent d'autres actions. Dans notre cas, un seul chemin correspondant à la situation en cours sera affiché au patient. Enfin, un protocole de soins est susceptible d'être modifié par des événements définis au préalable par l'équipe soignante. Un événement par conséquent provoquera l'exécution d'une nouvelle action par le patient (ex. pour l'évènement "présence de lipodystrophie dans la zone d'injection", l'action prévue consiste à "injecter l'insuline sur peau saine").

Dans une optique d'adaptation de protocoles de soins, le patient ou ses aidants peuvent signaler un ou plusieurs imprévus. La classe *imprévu* sera précisée dans des travaux futurs, elle ne constitue pas l'objet principal de cet article.

La section suivante présente le langage de représentation (la syntaxe concrète) que nous proposons d'associer à ce méta-modèle.

5. Un langage de représentation du protocole de soins

Nous souhaitons accompagner le patient au jour le jour en lui proposant un accès à la consultation de son propre protocole de soins en permettant d'intégrer son expertise à travers le signalement d'imprévus non définis dans son protocole initial. Cela implique également que les professionnels de santé puissent construire les protocoles de soins personnalisés à leurs patients, de façon simple et intuitive. Pour cela, nous proposons un langage de représentation de protocole de soins, destiné aux professionnels de santé et aux patients, suffisamment simple, intuitif et facile à comprendre.

Ce langage correspond à la syntaxe concrète associée au méta-modèle, il décrit les *actions quotidiennes* réalisées par le patient ainsi que les autres éléments qui interviennent dans cette réalisation, en particulier les *acteurs*, les *recommandations* à suivre, les *parties du corps concernées*, les *dispositifs médicaux et parapharmaceutiques* à utiliser ainsi que les *médicaments* à prendre. Pour chaque concept du méta-modèle, nous associons un symbole graphique ou une étiquette comme le montre la figure 5. Chaque étiquette est cliquable et apporte une information liée à l'action quotidienne (ex. sur des dispositifs médicaux, des médicaments, etc.).

Une *action quotidienne* est représentée graphiquement par un rectangle dans lequel seront collées des étiquettes (cf. figure 6) afin d'apporter toute l'information utile au patient. Dans l'exemple de la figure 6, l'action "contrôler la glycémie" de type "mesure et surveillance" est réalisée par "le patient" à l'aide de dispositifs médicaux "glucomètre, auto-piqueur..." sur la partie du corps "le doigt" avec la recommandation de "se piquer le côté du doigt et éviter le pouce et l'index". Cette action donne lieu au paramètre "glycémie". Ce paramètre sera saisi par le patient dans la zone paramètre.

Ce langage est une première proposition de représentation de protocoles de soins. Il sera à évaluer et à améliorer dans nos futurs travaux de recherche.

La section suivante présente la conception logicielle du langage de représentation de protocole de soins.

Concept	Etiquette
Sous protocole	
Décision	
Action d'éducation	
Action de soin	
Action de mesure et surveillance	
Paramètre	
Acteur : patient	
Acteur : aidant	
Acteur : professionnel de santé	
Partie du corps	
Recommandation	
Dispositif médical et parapharmaceutique	
Médicament	

Figure 5. Représentation du protocole de soins : légende

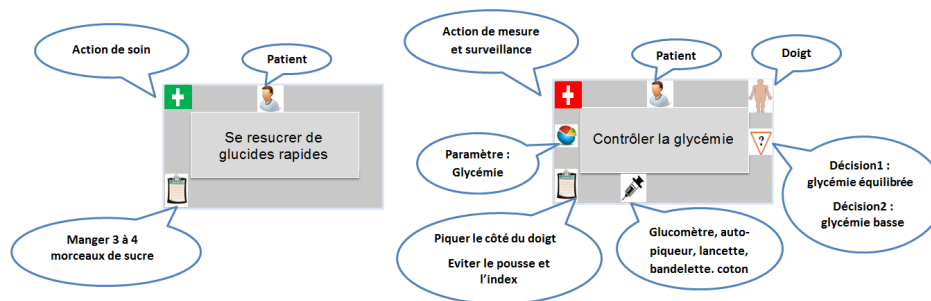


Figure 6. Représentation du protocole de soins : deux exemples d'actions

6. Conception de la solution logicielle

Le langage de représentation de protocoles proposé sera implémenté dans un prototype de la solution logicielle qui est en cours de conception sur un dispositif mobile dans l'environnement technologique de notre collaborateur industriel *Arcan Systems*⁶. Cette application proposera deux accès distincts, un pour le patient et l'autre pour le professionnel de santé. Selon chaque profil (patient ou professionnel de santé), un ensemble de fonctionnalités sera offert comme le montre le diagramme de cas d'utilisation de la figure 7.

6. www.arcan.fr

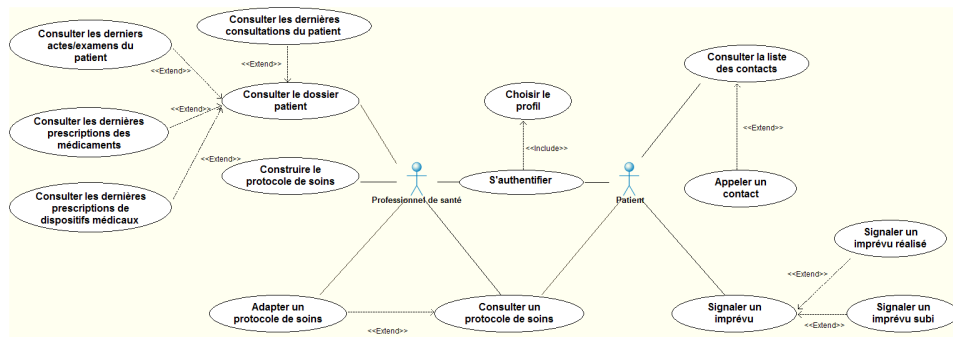


Figure 7. Diagramme de cas d'utilisation: fonctionnalités de l'application pour le professionnel de santé et le patient

– Profil "professionnel de santé" : les professionnels de santé concernés ont la possibilité de construire ou modifier le protocole de soins et de l'adapter en fonction des imprévus. La figure 8 présente un exemple d'interface de construction de protocole de soins.



Figure 8. Exemple de maquette d'interface "Profil professionnels de santé : construire un protocole"

– Profil "patient" : les patients concernés ont la possibilité de consulter leurs protocoles de soins (cf. figure 9) et d'intégrer leurs expertises à travers le signalement d'imprévus (cf. figure 10).

Dans la figure 9, le patient peut consulter l'action à réaliser "contrôler la glycémie". Cette action nécessite la saisie du paramètre "contrôler glycémie" dans la zone paramètre (l'interface de saisie fait partie de nos perspectives). En fonction de ce paramètre, l'action suivante "se resucrer de glucides rapides" s'affiche dans une nouvelle interface. Les étiquettes collées autour de l'action sont affichées sur la partie gauche.

Dans la figure 10, deux échelles de valeur varient de 0 à 5 pour représenter l'urgence ressentie (non urgent à extrêmement urgent) et la douleur ressentie (très modérée à

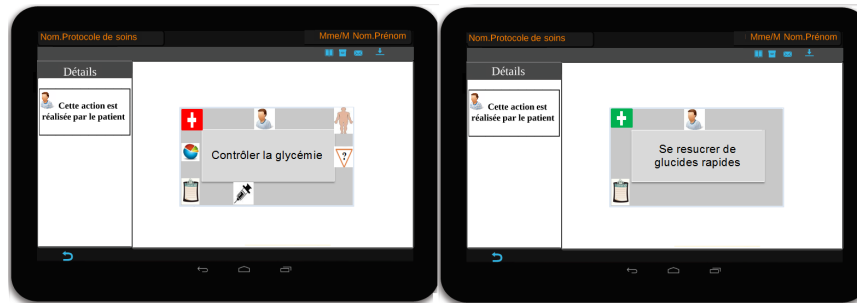


Figure 9. Exemple de maquette d'interface "Profil patient : suivre un protocole"

très importante). Le patient peut également décrire son problème imprévu subi dans la zone de texte correspondante.

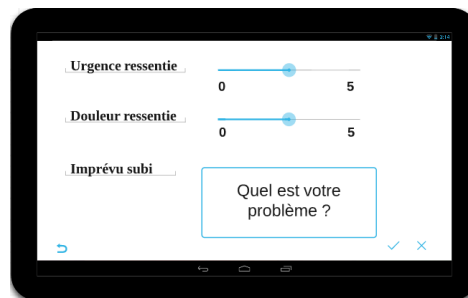


Figure 10. Exemple de maquette d'interface "Profil patient : signaler un imprévu subi"

Une fois l'imprévu signalé par le patient, le professionnel de santé recevra une notification de réception. De même pour le patient : dès que son imprévu aura été traité, il recevra une notification de traitement.

7. Conclusion et perspectives

La continuité des soins pour le patient chronique à domicile est essentielle. Elle lui permet de bien vivre sa maladie dans son environnement familial. En général, le patient connaît bien sa maladie et réalise quotidiennement un protocole de soins personnalisé. Néanmoins, la description de ce protocole reste souvent très informelle et peu guidée. Nous proposons dans nos travaux une nouvelle approche de construction de protocoles de soins. Pour ce faire, nous avons défini un DSL composé d'une syntaxe abstraite (un méta-modèle) et d'une syntaxe concrète (un langage graphique). Enfin, nous avons proposé des maquettes d'IHM que nous avons commencées à concevoir pour la construction de protocoles (par les professionnels de santé) et le signalement d'imprévus (par les patients). Cette solution n'est qu'une partie d'un ensemble plus large de travaux de recherche restant à effectuer.

Dans la suite de notre travail, un ensemble de perspectives répondant aux objectifs définis sera réalisé :

- Expérimentation pour validation du méta-modèle et du langage : il est nécessaire de confronter et valider notre proposition du méta-modèle, du langage de représentation et des maquettes d'IHM auprès des utilisateurs finaux. Nous envisageons de réaliser deux expérimentations de validation. La première sera réalisée auprès des professionnels de santé afin de valider les concepts et la terminologie employée dans le méta-modèle. La seconde cible les patients chroniques où un test de faisabilité et d'utilisabilité du langage et des maquettes proposées sera réalisé.

- Filtrage et adaptation du protocole : le concept d'imprévu intégré dans notre méta-modèle sera précisé par une étude approfondie de certaines solutions telles que les *ontologies* ou les *approches d'adaptation de processus*. L'objectif sera de pouvoir manipuler les imprévus signalés par le patient ou ses aidants, les filtrer et leur associer un raisonnement.

- Implémentation de la solution logicielle : un prototype de la solution sera implémenté sur un dispositif mobile en collaboration avec l'entreprise Arcan avec une amélioration de l'IHM dédiée aux patients en vue de simplifier son intégration (ex. utilisation de dispositifs de reconnaissance de la parole).

Bibliographie

- Ardissono L., Di Leva A., Petrone G., Segnan M., Sonnessa M. (2005). Adaptive medical workflow management for a context-dependent home healthcare assistance service. In *International workshop on context for web services (cws 2005)*, p. 59-68. Torino, Italy, Elsevier.
- Bernonville S., Vantourout C., Fendeler G., Beuscart R. (2013). A business process modeling experience in a complex information system re-engineering. In *stud health technol inform*. Pub Med.
- Bricon-Souf N., Dufresne E., Watbled L. (2005). Integration of the cognitive knowledge of activity in a service oriented architecture in the home care context. In *Studies in health technology and informatics*, vol. 116, p. 923-928.
- Collectif-DGOS. (2012). *Recommandations pour la mise en oeuvre d'un projet de télémédecine: déploiement technique: Urbanisation et infrastructure*. Web.
- Dadam P., Reichert M. (2000). Towards a new dimension in clinical information processing. In *stud health technol inform*, p. 259-301. Elsevier.
- Derradji A. (2014). Construction de workflow auto-adaptatif pour le suivi des maladies. In *Forum jeunes chercheurs du 32e congrès inforsid*. Lyon, France.
- Dupuy-Chessa S. (2006). *Modélisation en interaction homme-machine et en système d'information: à la croisée des chemins*. Habilitation à diriger des recherches, Université de Grenoble, France.
- Flory A., Verdier C., Sassi S. (2006). Nouvelle interfaces pour la représentation de l'information médicale: vers une prise en compte de l'approche métier. In *24e congrès inforsid*, p. 177-197. Hammamet, Tunisie.

- Fondement F., Baar T. (2005). Making metamodels aware of concrete syntax. In *First european conference on model driven architecture - foundations and applications (ecmda-fa 2005)*, p. 190-204.
- Greiner U., Mueller R., Rahmi E., Ramsch J., Heller B., Loeffler M. (2005). Adapflow: Protocol-based medical treatment using adaptive workflows. *Methods of Information in Medicine*, vol. 44, p. 80-88.
- Han M., Thiery T., Song X. (2006). Managing exceptions in the medical workflow systems. In *28th international conference on software engineering*, p. 741-750. Shanghai, China.
- Ilahi L., Ayachi Ghannouchi S. (2013, june). Improving telemedicine processes via bpm. In *Conference on enterprise information systems/ international conference on project management/ international conference on health and social care information systems and technologies*, p. 1209-1216. Montpellier, Procedia Technology.
- Leonardi G., Panzarasa S., Quaglini S., Stefanelli M., Aalst W. Van der. (2007). Interacting agents through a web-based health serviceflow management system. *Journal of Biomedical Informatics*, vol. 40, p. 486-499.
- Mans R. (2011). *Workflow support for the healthcare domain*. Thèse de doctorat, Université Eindhoven, Pays-Bas.
- Morley C., Hugues J., Leblanc B., Hernandez J., Joly R., Gardian G. (2006). *Uml 2 pour l'analyse d'un système d'information: Le cahier des charges du maître d'ouvrage n° 3*. Paris, Editions Dunod.
- OMG. (2010, juin). *Business process model and notation (bpnm 2.0)*. Rapport technique n° 2. Object Management Group.
- OMS. (1986). *Promotion de la santé: charte d'ottawa*. Web.
- Picard R., Salgues B. (2008). *Enjeux des tic pour l'aide à l'autonomie des patients et des citoyens en situation de handicap ou de fragilité dans leurs lieux de vie*. Web.
- WFMC. (1999). *Workflow management coalition terminology & glossary, document number wfmc-tc-1011*. Web.
- Zefouni S. (2012). *Aide à la conception de workflows personnalisés: application à la prise en charge à domicile*. Thèse de doctorat en informatique, Université de Toulouse, France.

Maintien de la cohérence entre les architectures pour une gestion autonome de systèmes pervasifs

Approche basée sur les méta-modèles des architectures de conception et d'exécution

Stéphanie Chollet¹, Philippe Lalanda², Catherine Hamon³

1. *Laboratoire de Conception et d'Intégration des Systèmes*

50, rue Barthélémy de Laffemas, BP 54

26902 Valence Cedex 9, France

stephanie.chollet@lcis.grenoble-inp.fr

2. *Laboratoire d'Informatique de Grenoble*

220, rue de la chimie, BP 53

38041 Grenoble Cedex 9, France

philippe.lalanda@imag.fr

3. *Orange Labs*

28, chemin du vieux chêne

38243 Meylan, France

catherine.hamon@orange.com

RÉSUMÉ. L'informatique orientée service est devenue très populaire ces dernières années, en particulier pour traiter des environnements dynamiques et imprévisibles. Cependant, cette nouvelle approche soulève de sérieux problèmes au niveau de l'administration et de la maintenance des systèmes. Dans cet article, nous développons une proposition où les architectures logicielles de conception et d'exécution sont utilisées pour gérer les systèmes orientés service et pour aider les administrateurs à suivre et comprendre les évolutions à l'exécution. Nous montrons comment des concepts liés à la conception et à l'exécution peuvent être liés et exploités par un gestionnaire autonome ou par un administrateur humain. Cette approche est validée sur un cas d'utilisation issu du domaine de la santé pervasive et développé avec Orange Labs.

ABSTRACT. Service-oriented programming has become very popular in recent years, in particular to deal with dynamic, unpredictable environments. However, this novel approach raises

major issues in term of administration. In this paper, we present an approach where design and runtime architectures are used to manage service-oriented systems and help administrators to follow and understand runtime evolutions. We show how concepts of design time and runtime can be linked and exploited by an autonomic manager or by a human administrator. This approach is validated on a real use case belonging to the pervasive health domain and built with the Orange Labs.

MOTS-CLÉS : Informatique autonome, Approche dirigée par les modèles, Approche orientée service

KEYWORDS: Autonomic Computing, Model-Driven Approach, Service-Oriented Computing

1. Introduction

L'informatique orientée service (Papazoglou, 2003) est fortement utilisée aujourd'hui pour permettre des adaptations à l'exécution. Cette approche est fondée sur les principes de faible couplage, de liaison retardée et de substituabilité de modules logiciels appelés service. L'informatique orientée service permet ainsi d'aborder de nouveaux domaines d'application caractérisés notamment par de fortes contraintes de dynamisme et par une faible prédictibilité. De façon non surprenante au vu de ses propriétés, l'informatique orientée service rencontre un réel succès pour la mise en place d'applications ambiantes (ou pervasives), de plus en plus répandues aujourd'hui.

Par ailleurs, les adaptations sont souvent réalisées de façon autonome (Lalanda *et al.*, 2013). Les propriétés autonomiques telles que l'auto-configuration, l'auto-optimisation, l'auto-réparation ou encore l'auto-protection peuvent permettre de faire face à la complexité croissante des besoins en adaptation. Ces propriétés peuvent permettre de diminuer significativement les coûts d'exploitation et de maintenance d'une part et palier la faible disponibilité des administrateurs d'autre part.

L'évolution autonome des applications soulève néanmoins des problèmes de suivi et de compréhension de l'état des applications. Notre objectif est de mieux formaliser la connaissance des applications dans le cadre des systèmes orientés services et, notamment, pour les applications pervasives. Ceci dans le but de faciliter le travail des administrateurs ou la construction de gestionnaires autonomiques. Nous nous focalisons sur le niveau architectural qui nous semble être le bon niveau de granularité pour mener à bien des adaptations. Disposer d'une formalisation claire de l'architecture exécutée est très précieux pour guider les adaptations. Cependant, cette architecture seule n'est pas suffisante. En effet, il est nécessaire de disposer d'une autre connaissance architecturale qui donne l'ensemble de propriétés que l'on souhaite garantir à l'exécution et qui exprime ce qu'est une architecture valide afin de conduire les adaptations conformément à cette description. Cette seconde moitié de connaissance qui est issue de la phase de conception, sera nommée par la suite architecture de conception. Cette complémentarité entre l'architecture de l'exécution et l'architecture de conception avait déjà été perçue par Garlan et Schmerl (2002).

Dans cet article, nous montrons comment simplifier les tâches des administrateurs d'applications orientées services en maintenant des liens de traçabilité entre les architectures réalisées à différents moments du cycle de vie de l'application. L'article est structuré de la façon suivante. La section 2 détaille un cas d'utilisation et ses besoins. La section 3 présente notre approche qui est ensuite développée dans la section 4. La section 5 traite de l'implantation de notre approche et de sa validation. Certains travaux connexes sont discutés en section 6. Enfin, cet article se termine par une conclusion en section 7.

2. Motivation et cas d'utilisation

Notre cas d'utilisation est issu du projet FUI Medical¹. C'est un cas d'utilisation fréquent dans le domaine de la santé ; il a pour objectif principal le maintien des personnes âgées à leur domicile. Plus précisément, nous avons travaillé sur l'application nommée actimétrie qui récupère un certain nombre de mesures et analyse l'activité physique d'une personne dans son environnement. L'idée est de tracer et enregistrer les habitudes d'un utilisateur pour détecter des changements d'habitudes imperceptibles par un médecin lors d'une consultation. Ces changements d'habitudes peuvent être un signe avant-coureur de sérieux problèmes de santé.

Pour mettre en place une telle application, il faut avoir des capteurs installés dans une maison permettant de récupérer des informations géolocalisées. Ces capteurs sont connectés par un réseau. Une de leurs principales caractéristiques est qu'ils sont hétérogènes et dynamiques. Pour ce type d'application (Figure 1), n'importe quel type de capteurs peut être utilisé : détecteurs de présence, tensiomètre et aussi des équipements comme une télévision, une machine à laver ou une machine à café connectées. Les données collectées sont régulièrement transmises à un serveur distant qui réalise une analyse. Cette analyse est utilisée pour détecter des modifications de comportements. Ces informations ne sont ensuite transmises qu'au médecin.

En pratique, la réalisation d'une telle application se base sur un ensemble de services logiciels. Les équipements et les applications sont exposés comme des services logiciels de différentes natures comme des services UPnP², des services Web³, des services DPWS⁴, etc. L'intégration de ces services distants ou locaux doit se faire de manière transparente (Lalanda *et al.*, 2010), (Escoffier *et al.*, 2014).

L'application d'actimétrie n'est pas contrainte au niveau du choix des équipements disponibles ; chaque maison peut être équipée d'un certain nombre de capteurs fournis par des fabricants différents. D'un point de vue logiciel, ceci est assez complexe à supporter. Il faut réaliser un code d'intégration conséquent et complexe. De plus,

1. Projet en collaboration avec Orange Labs, Université Joseph Fourier, Telecom ParisTech et ScalAgent : <http://medical.imag.fr>

2. <http://www.upnp.org>

3. www.w3c.org

4. <http://docs.oasis-open.org/ws-dd/ns/dpws/2009/01>

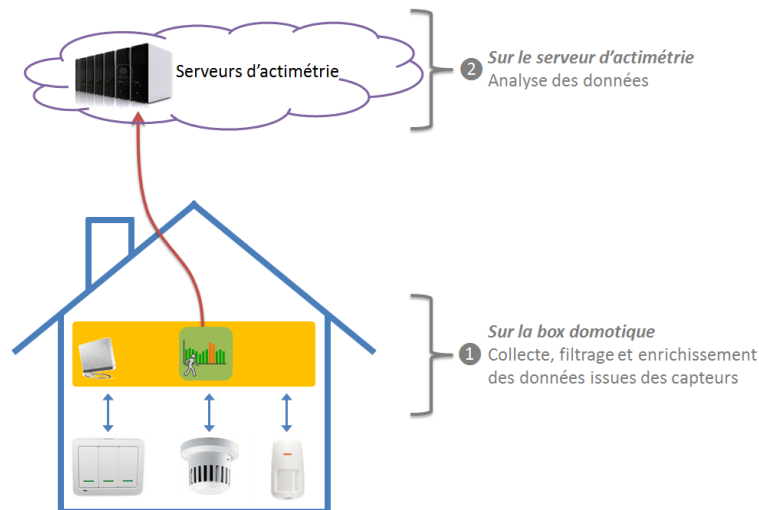


Figure 1. L'application Actimétrie.

il faut pouvoir gérer l'hétérogénéité et le dynamisme des capteurs. Toutes les situations ne peuvent pas être anticipées au moment de la conception de l'application et il faut apporter quelques informations supplémentaires au moment de l'exécution pour s'adapter à l'environnement changeant. C'est dans ce cadre que s'inscrit notre proposition.

3. Approche globale

3.1. Principes

Notre objectif est de simplifier les tâches des administrateurs d'application à base de composants orientés services en maintenant des liens pour la traçabilité entre les architectures réalisées à différent moment du cycle de vie de l'application. En effet, comme nous l'avons vu précédemment, l'architecture de l'exécution évolue suite à des modifications de l'environnement d'exécution (arrivée et/ou départ d'équipements). L'administrateur doit être capable de savoir si l'exécution qui est en cours est toujours valide par rapport à celle qui est souhaitée. Nous souhaitons automatiser la vérification de la validité de l'architecture à l'exécution. Pour ce faire, nous définissons trois types d'architectures qui sont utilisées à différents étapes du cycle de vie :

- **l'architecture de conception** qui est une composition de différents éléments liés les uns aux autres comprenant de la variabilité ; c'est-à-dire des contraintes lâches qui seront précisées dans les étapes suivantes du processus de développement.

- **l'architecture de déploiement** qui est issue de l'architecture précédente et qui contient en plus de nombreux éléments configurés. L'architecture de déploiement est celle qui doit être déployée sur la plate-forme cible. Les configurations sont dépendantes de cette plate-forme et du contexte. Cependant, dans cette architecture, il reste

encore un peu de variabilité qui sera résolue à l'exécution. Cette variabilité est restreinte et n'a de sens que pour une application basée sur une approche à composants orientés services.

– **l'architecture de l'exécution** qui est la représentation de l'application en exécution. Cette architecture est construite grâce à l'analyse de l'exécution avec des mécanismes de surveillance. Elle est une vue de l'exécution avec un certain niveau d'abstraction ; toutes les informations de l'exécution ne sont pas représentées. Dans cette architecture, il n'existe plus de variabilité ; ces dernières ont toutes été résolues par la machine d'exécution. La caractéristique principale de cette architecture est qu'elle évolue fréquemment, par opposition aux deux autres types d'architecture.

Pour maintenir le lien entre ces différentes architectures, nous proposons une approche basée sur les principes de l'informatique autonome. La Figure 2 illustre notre approche globale. Nous ajoutons au-dessus de notre système une boucle autonome MAPE-K (Kephart, Chess, 2003) et nous nous intéressons plus particulièrement à la structuration de la connaissance. C'est dans la base de connaissances que nous voulons maintenir le lien entre les architectures de conception et de l'exécution. Les phases d'analyse et de planification doivent permettre de vérifier la validité entre l'architecture de l'exécution et celle de la conception. Il faut noter que l'architecture de déploiement n'apparaît pas dans cette figure ; en effet, elle est utile uniquement pour le premier déploiement de l'application sur la plate-forme. Sitôt qu'elle est déployée, elle est réifiée sous la forme d'une architecture de l'exécution. Cette approche requiert de mettre en place des *touchpoints* (capteurs et actionneurs) sur le système permettant, à la phase de monitoring, de mettre à jour l'architecture de l'exécution et, à la phase d'adaptation, de modifier l'exécution suivant ce qui a été planifié (Morand *et al.*, 2011).

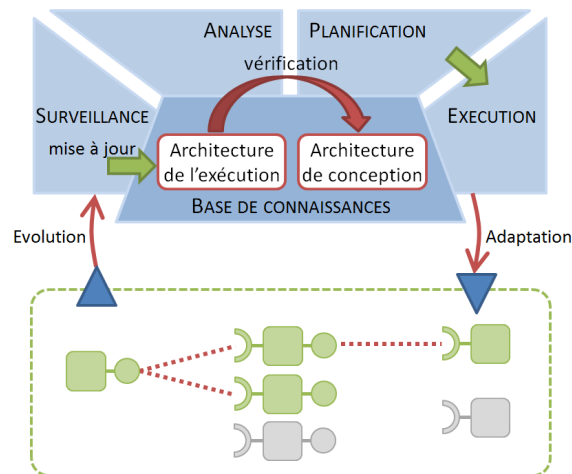


Figure 2. Principes de notre approche.

Dans les deux parties suivantes, nous allons présenter d'abord la formalisation de la connaissance, puis la méthode de vérification de la validité entre les architectures.

3.2. Formalisation de la connaissance à base de modèles et méta-modèles

La formalisation des architectures a pour objectif d'identifier clairement les différents éléments qui les composent. Pour ce faire, notre proposition est basée sur la définition d'un ensemble de méta-modèles (OMG, 2002 ; Favre, 2004). Un méta-modèle permet de spécifier de manière claire, précise et non-ambiguë le langage utilisé pour modéliser les architectures. Plus précisément, un méta-modèle sert à définir une grammaire et un vocabulaire afin de réaliser des modèles cohérents et conformes à celui-ci. Dans notre cas, nous proposons de définir un méta-modèle pour chaque type d'architecture afin de pouvoir construire des applications conformes et cohérentes.

Cependant, la définition des trois méta-modèles (architecture de conception, de déploiement et de l'exécution) pour exprimer les différentes architectures met en œuvre un ensemble de concepts communs dû au fait que se sont toutes les trois des architectures. Ces dernières comprennent certes des concepts de sens commun mais avec des particularités propres au domaine d'application de l'architecture. Nous proposons de concevoir un méta-modèle commun pour les trois architectures et qui sera hérité pour chacun des méta-modèle des architectures.

Un des principaux avantages de cette organisation est d'avoir une continuité entre les concepts. Il est clair qu'un composant dans l'architecture de conception doit être le même composant dans l'architecture de déploiement avec certes des caractéristiques supplémentaires pour la phase de déploiement. Un composant exécuté est, quant à lui, réifié dans l'architecture de l'exécution mais il reste intrinsèquement un composant comme défini dans les autres architectures.

Comme nous l'avons présenté précédemment, les différentes architectures sont définies avec des niveaux de variabilité qui diffèrent comme présenté dans la Figure 3. L'architecture de conception est la plus variable. Le méta-modèle doit donc supporter l'expression de cette variabilité. L'architecture de déploiement est un raffinement d'une architecture de conception ; elle contient moins de variabilité mais respecte les contraintes de l'architecture de conception dont elle est issue. Les seuls éléments de variabilité qui restent sont ceux qui pourront être résolus par la machine d'exécution. L'architecture de l'exécution est une vue de ce qui s'exécute et, par conséquent, il n'y a plus aucune variabilité dans ce modèle.

3.3. Vérification de la conformité entre modèles

La conformité entre les modèles est assurée en deux étapes :

– premièrement, par construction, l'architecture de déploiement est issue de l'architecture de conception. La variabilité exprimée dans l'architecture de conception est spécifiée et restreinte au niveau du déploiement. L'architecture de l'exécution est issue de l'architecture de déploiement au démarrage. Elle est uniquement une représentation de l'architecture de déploiement au moment du premier déploiement. Cependant, pendant l'exécution cette architecture de l'exécution évolue. Le lien de conformité est naturellement perdu.

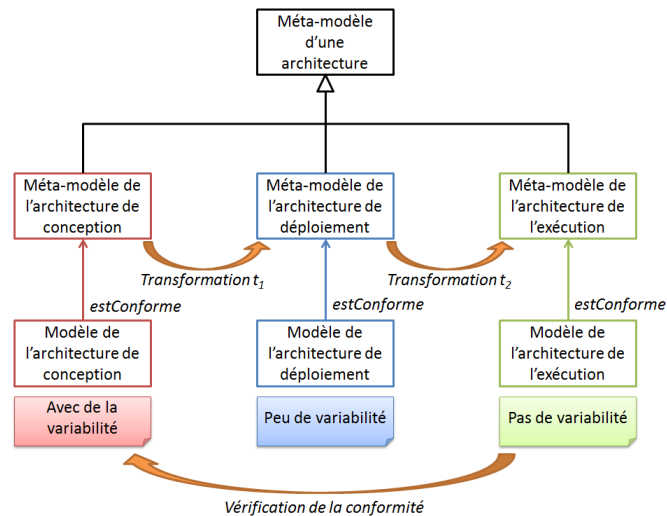


Figure 3. Modélisation des différentes architectures avec de la variabilité.

– deuxièmement, nous proposons un algorithme de vérification de la conformité entre l'architecture de l'exécution et l'architecture de conception. Cet algorithme permet de vérifier si l'architecture est toujours valide et sinon en informer soit l'administrateur soit le gestionnaire autonome de niveau supérieur pour qu'il soit alerté du problème.

Dans notre approche, nous proposons de définir deux transformations entre les méta-modèles pour assurer la conformité par construction, comme illustré dans la Figure 3. Ces deux transformations ne sont pas du même type. La première transformation entre le méta-modèle de l'architecture de conception et celui de l'architecture de déploiement est l'expression de la spécialisation/configuration d'un certain nombre d'éléments. La deuxième transformation entre le méta-modèle de l'architecture de déploiement et celui de l'architecture de l'exécution est la projection sur une plate-forme d'exécution. Evidemment, ces transformations exprimées au niveau du méta-modèle doivent être valides au niveau du modèle.

Pendant l'exécution, l'architecture de l'exécution évolue à cause de changements dans l'environnement d'exécution. Nous proposons de mettre en œuvre un algorithme de validation de la conformité qui s'assure que les éléments modifiés à l'exécution respecte les contraintes définies dans l'architecture de conception. La difficulté principale pour cette vérification est qu'il faut savoir à quel moment elle doit être opérée. Il faut que la vérification se fasse à un moment où l'état de l'exécution est valide ; c'est-à-dire que ce soit un état « stable » et non dans un état où des modifications « en cascade » ne sont pas terminées. De plus, l'algorithme de vérification surcharge le fonctionnement dit normal de l'application, il ne faut pas l'appeler trop fréquemment comme cela peut être le cas lorsqu'un équipement est « instable » ; par exemple, il est en limite de portée pour être détecté en continu.

4. Méta-modèles et modèles

4.1. Méta-modèle commun aux architectures

Le méta-modèle commun aux trois architectures est présenté à la Figure 4.

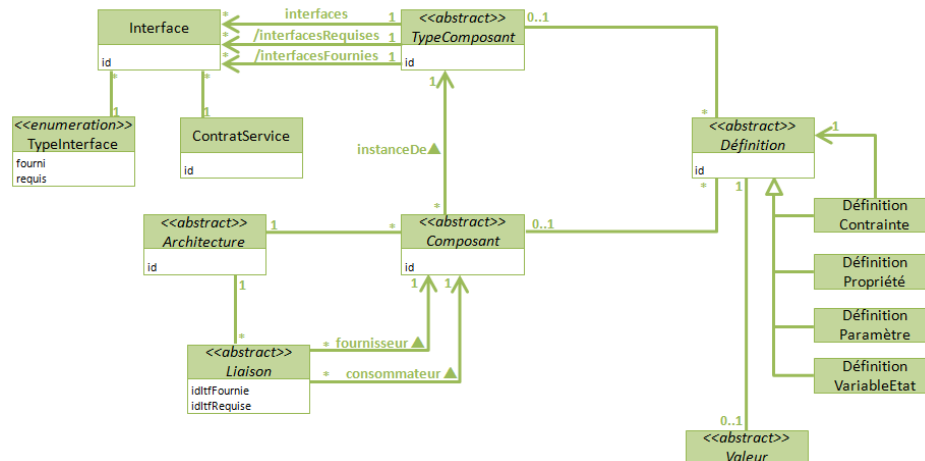


Figure 4. Méta-modèle commun définissant une architecture.

Une architecture est représentée par la classe abstraite *Architecture*. Chaque architecture est définie par un identifiant (*id*) unique, des composants et des liaisons. Chaque liaison et chaque composant a nécessairement une unique architecture d'appartenance. Chaque composant est défini par un identifiant (*id*) tel que tous les composants, qui font partie d'une même architecture, ont tous des identifiants différents. Chaque composant est une instance d'un type de composant, défini par un identifiant (*id*) unique. Chaque type de composant dispose d'un ensemble d'interfaces, qui ne peuvent pas être partagées entre composants. Chaque interface est identifiée avec un *id*, de sorte qu'un type de composant ne puisse pas disposer de deux interfaces avec le même identifiant.

Une interface peut être soit fournie, soit requise, comme l'indique son *TypeInterface*. Une interface fournie permet à un composant d'offrir un service, alors qu'une interface requise crée au contraire une dépendance de service. Les différentes interfaces d'un type de composant peuvent donc être réparties entre les interfaces fournies et les interfaces requises. C'est la raison pour laquelle nous avons les relations dérivées *interfacesRequises* et *interfacesFournies* (contrainte OCL 1).

Quel que soit le type de l'interface, le service dont il est question, est décrit grâce au *ContratService* de l'interface. Chaque *ContratService* dispose lui aussi d'un identifiant unique.

Une liaison permet de lier deux composants par l'intermédiaire de leurs interfaces, définies au niveau de leur type. C'est la raison pour laquelle une liaison est définie par

```

context TypeComposant
  inv: self.interfaces → select(i | i.TypeInterface = requis) →
    forAll(i | self.interfacesRequises → exists(ir | ir=i))
  inv: self.interfacesRequises →
    forAll(i | self.interfaces → select(i | i.TypeInterface = requis) → exists(ir | ir=i))
  inv: self.interfaces → select(i | i.TypeInterface = fourni) →
    forAll(i | self.interfacesFournies → exists(if | if=i))
  inv: self.interfacesFournies →
    forAll(i | self.interfaces → select(i | i.TypeInterface = fourni) → exists(if | if=i))

```

Contrainte OCL 1: Relations entre *interfacesRequises* et *interfacesFournies*.

les deux composants ainsi que par les identifiants des interfaces (présents dans leur type) à lier (contrainte OCL 2).

```

context Liaison
  inv: self.fournisseur.instanceDe.interfacesFournies → one(ili.id=self.idItfFournie)
  inv: self.consommateur.instanceDe.interfacesRequises → one(ili.id=self.idItfRequise)

```

Contrainte OCL 2: Types des interfaces pour les liaisons.

De plus, une liaison ne peut lier que des interfaces qui exposent le même contrat de service. Il est impossible d'établir une liaison entre deux interfaces d'un même composant. Enfin, il n'est pas possible de trouver deux liaisons reliant exactement les mêmes interfaces au sein d'une même architecture (contrainte OCL 3).

```

context Liaison
  inv: self.fournisseur.instanceDe.interfacesFournies →
    select(i | i.id = self.idItfFournie) → asSequence() →
      first().ContratService.id = self.consommateur.instanceDe.interfacesRequises →
        select(i | i.id = self.idItfRequise) → asSequence() → first().ContratService.id
  inv: self.fournisseur <> self.consommateur
context Architecture
  inv: self.Liaison → forAll(l1, l2 | l1 <> l2
  implies (l1.fournisseur <> l2.fournisseur)
    or (l1.consommateur <> l2.consommateur)
    or (l1.idItfFournie <> l2.idItfFournie)
    or (l1.idItfRequise <> l2.idItfRequise))

```

Contrainte OCL 3: Contraintes sur les liaisons.

Les composants peuvent disposer de propriétés, de paramètres de variables d'état et de contraintes, éventuellement valués selon le type de composant considéré. Ces éléments permettent de gérer une partie de la variabilité supportée. Par exemple, on peut définir une variable d'état *tempsTraitement* que l'on souhaite d'un niveau moyen à la phase de conception et, au moment de l'exécution, le temps de traitement est valué. Nous avons séparé l'aspect définition de la valeur pour ces quatre concepts. Une définition est en relation soit avec un *TypeComposant*, soit avec un *Composant* (de manière exclusive). De plus, toutes les définitions d'un composant ou d'un type de

composant ont des identifiants uniques au sein de celui-ci. Une définition de contrainte est en relation avec une définition, qui ne peut pas être une contrainte (contrainte OCL 4).

```

context DéfinitionContrainte
    inv: not(self.Définition.ocIsTypeOf(DéfinitionContrainte))
    
```

Contrainte OCL 4: Non récursivité des définitions de contraintes.

4.2. Méta-modèle de l'architecture de conception

L'architecture de conception est formalisée avec le méta-modèle présenté dans la Figure 5. Ce méta-modèle étend le méta-modèle commun présenté précédemment.

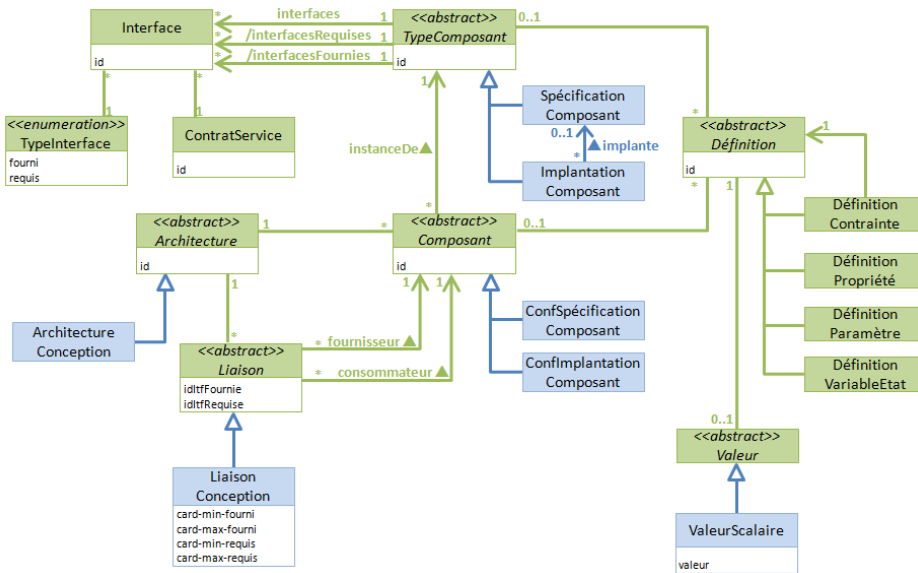


Figure 5. Méta-modèle de l'architecture de conception.

Ce méta-modèle représente les concepts nécessaires pour définir des modèles de l'architecture de conception. Nous devons retrouver dans ce méta-modèle les éléments qui sont particuliers à l'architecture de conception : les composants sont soit des spécifications, soit des implantations. Cette architecture peut disposer de configurations de spécifications et de configurations d'implantations de composant. Comme nous l'avons vu, une configuration de spécification (respectivement d'implantation) de composant instancie forcément une spécification (respectivement une implantation) de composant. Cela est exprimé avec la contrainte OCL 5.

De plus, cette architecture supporte une variabilité exprimée au niveau des liaisons avec des cardinalités. Ces dernières sont présentes sur chacune des extrémités des liaisons : côté interface fournie (`card-min-fourni` et `card-max-fourni`) et côté interface requise (`card-min-requis` et `card-max-requis`).

<p>context ArchitectureConception inv: self.Composant \rightarrow forAll(c.ocIsTypeOf(ConfSpécificationComposant) or c.ocIsTypeOf(ConfImplantationComposant))</p> <p>context ConfSpécificationComposant inv: self.reférence.ocIsTypeOf(SpécificationComposant)</p> <p>context ConfImplantationComposant inv: self.reférence.ocIsTypeOf(ImplantationComposant)</p>
--

Contrainte OCL 5: Contraintes sur les types de composants des architectures de conception.

Une architecture de conception peut disposer de configurations de spécifications et d'implantations de composants. Une configuration de spécification (resp. d'implantation) de composant instancie forcément une spécification (resp. d'implantation) de composant. De plus, une liaison permet d'exprimer qu'une implantation de composant peut ou non planter une unique spécification de composant. Cette liaison est orientée, car la spécification n'a pas besoin de savoir quelles sont les implantations qui l'implantent.

4.3. Méta-modèle de l'architecture de l'exécution

Le méta-modèle de l'architecture de l'exécution (Figure 6) vient, elle-aussi, étendre le méta-modèle commun. Il existe plusieurs architectures de l'exécution qui peuvent être issues d'une architecture de conception. Les architectures de l'exécution référencent une architecture de conception via une indirection. Cette dernière n'est navigable que dans un sens, car une architecture de conception n'a pas à savoir par quelles autres architectures elle est utilisée.

Une architecture de l'exécution ne peut disposer que d'instances de composants (contrainte OCL 6), chacune instanciant une implantation de composant.

<p>context ArchitectureExécution inv: self.Composant \rightarrow forAll(c.ocIsTypeOf(InstanceComposant))</p> <p>context InstanceComposant inv: self.instanceDe.ocIsTypeOf(ImplantationComposant)</p>

Contrainte OCL 6: Contraintes sur les types de composants des architectures de l'exécution.

Notons enfin que la hiérarchie de classes issue de la classe *Valeur* est différente par rapport aux autres architectures. En effet, dans le cas des variables d'état, nous souhaitons disposer d'un historique horodaté des valeurs successives prises par celle-ci. La classe *Collection* permet donc d'agréger un ensemble de valeurs scalaires, par l'intermédiaire de la classe horodatage.

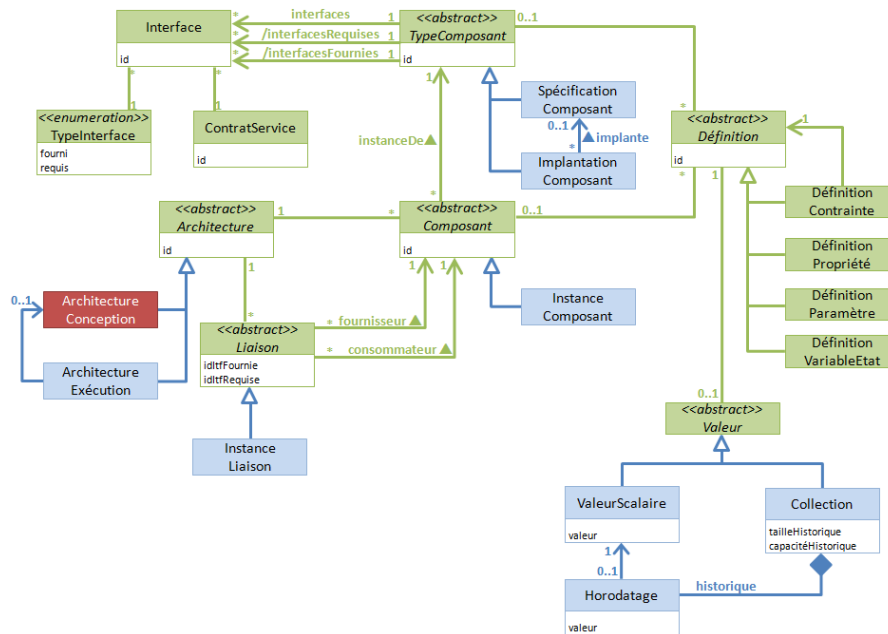


Figure 6. Méta-modèle de l'architecture de l'exécution.

5. Implantation et validation

5.1. Implantation

Notre approche a été mise en œuvre dans le cadre du projet FUI Medical pour l'application Actimétrie présentée précédemment. Notre solution se base sur la plateforme Cilia Mediation Framework⁵ qui permet de simplifier la construction d'applications de médiation. Cilia supporte l'évolution à chaud de son architecture à base de composants orientés services dynamiques et est capable de rendre compte de son état interne. De plus, Cilia embarque l'intergiciel RoSe⁶, dédié à la découverte et la communication avec des équipements et des services distants.

L'implantation de notre approche est résumée sur la Figure 7 : la plateforme d'exécution Cilia avec RoSe, une boucle de contrôle dans laquelle nous avons étendue la connaissance avec les différentes architectures, Cilia IDE qui est un outil de conception et de supervision des architectures.

L'atelier Cilia IDE, basé sur Eclipse IDE, permet de concevoir, de déployer et de superviser des architectures à base de composants Cilia soit au format XML soit de manière graphique. Il perçoit l'état de la plateforme Cilia via la base de connaissance

5. <http://wikiadele.imag.fr/index.php/Cilia>

6. <http://wiki.chameleon.ow2.org/xwiki/bin/view/Main/Rose>

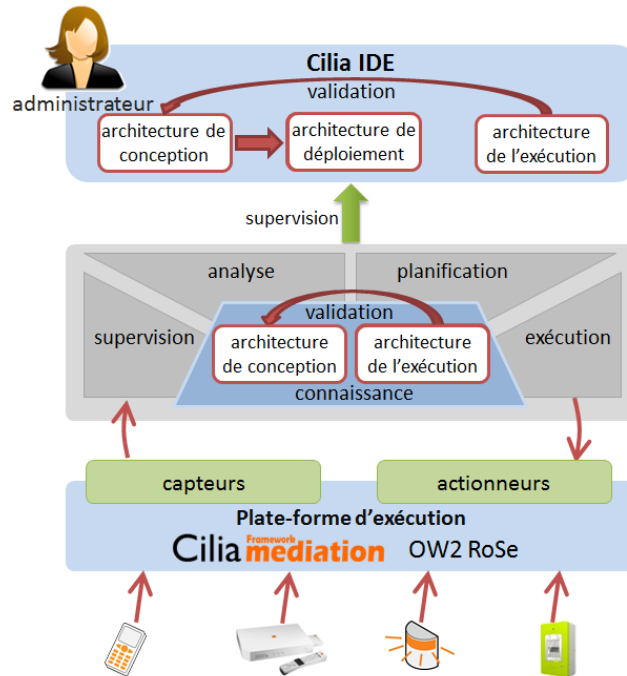


Figure 7. Architecture de l'implantation.

qu'il peut, par ailleurs, alimenter sur la partie conception. Enfin, il est en mesure d'envoyer des ordres de reconfiguration directement à la plate-forme Cilia. Et, il permet de montrer, s'il y a une violation de contraintes entre les architectures, où elle se trouve.

5.2. Validation

Dans le cadre de l'application actimétrie, la perception des habitudes des personnes âgées est effectuée au travers de leurs interactions avec l'écosystème numérique de leur habitat : utilisation de la *set top box*, manipulation d'interrupteurs connectés, passage devant des détecteurs de présence ou activation des détecteurs d'ouverture de porte. Les événements générés par ces équipements sont agrégés et permettent à l'application de localiser les personnes dans leur habitat et de déduire des activités sur leur activité courante. Dans le cadre de cette validation, l'application est séparée en deux parties : la collecte, le filtrage et l'enrichissement des données brutes issues des capteurs de l'habitat, d'un part, et l'analyse de ces données, d'autre part.

La première partie est donc une chaîne de médiation que l'on peut facilement développer avec Cilia. Elle a un tronc commun d'actions par contre la collecte d'informations est adaptée à l'habitat de chacun des utilisateurs, l'écosystème d'équipements étant différent d'un habitat à un autre.

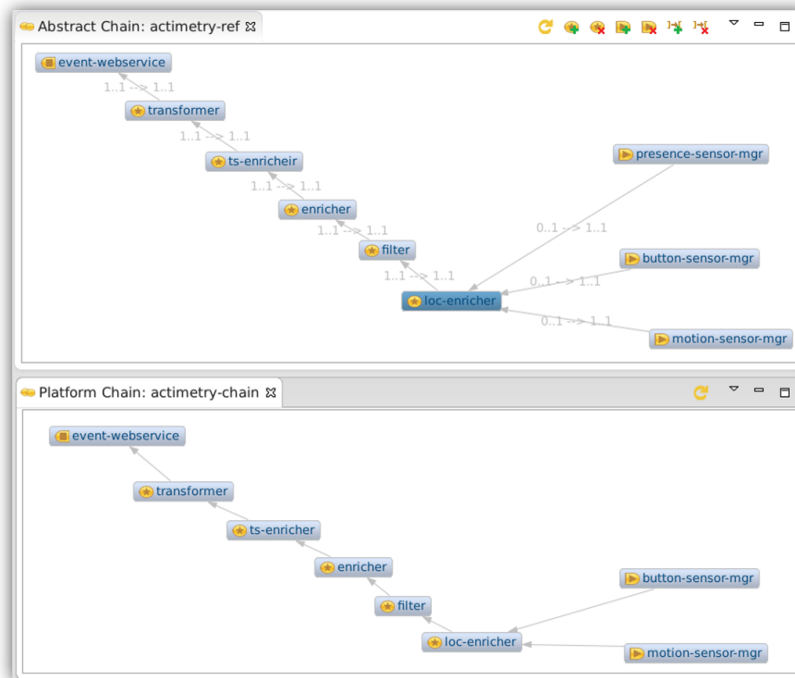


Figure 8. Capture d'écran de Cilia IDE avec deux architectures.

La Figure 8 montre un exemple pour l'application d'actimétrie. La partie haute représente l'architecture de conception qui comprend de la variabilité avec, par exemple, des cardinalités. La partie basse représente l'architecture de l'exécution. Lorsque l'on clique sur un élément, des fenêtres de propriétés permettent d'avoir l'ensemble des caractéristiques des composants. Ces deux vues permettent de détecter facilement les violations de contraintes. Les éléments qui ne respectent pas l'architecture de conception sont représentés en rouge.

6. Travaux connexes

Dans le cadre de l'informatique autonome, les notions de modèles des besoins/de conception (Elkhodary *et al.*, 2010 ; Baresi *et al.*, 2010) et de modèles des architectures (Garlan *et al.*, 2004 ; Floch *et al.*, 2006) sont très utilisées pour permettre une auto-adaptation des systèmes grâce aux liens entre les modèles. La connaissance est centrale dans la boucle autonome (Kephart, Chess, 2003). Dans nos travaux, nous avons fait le choix de garder dans la connaissance le modèle de conception et celui d'exécution pour vérifier la conformité entre ces modèles. La difficulté majeure est de savoir ce qui doit être représenté de manière explicite au niveau de la connaissance. Evidemment, plus il y a d'informations explicites, plus il est facile d'implanter et de maintenir une boucle autonome basée sur cette connaissance (Lalanda *et al.*,

2013). Un avantage de cette approche architecturale est que le modèle architectural peut être utilisé pour vérifier que l'intégrité du système est préservée lors des modifications. En effet, les évolutions sont prévues et appliquées dans un premier temps au niveau du modèle, ce qui permet de montrer quel sera l'état résultant du système et ainsi en déduire les violations éventuelles de contraintes (Oreizy *et al.*, 1998).

Une des limites de ces approches existantes est le manque d'expressivité possible au niveau des modèles de conception. Les modèles sont définis à base de composants, de connecteurs et de contraintes (*e.g.*, les plates-formes Acme (Garlan, Schmerl, 2002), C2/xADL (Oreizy *et al.*, 1998)) mais il y a relativement peu de variabilité supportée. De plus, une autre limite à l'utilisation de ces plates-formes est le manque d'outils facilitant la gestion complète du cycle de vie. Comme nous l'avons montré, il y a une cohérence entre les architectures de la conception et de l'exécution. Cependant, il n'existe pas aujourd'hui d'outils permettant aux concepteurs, développeurs, administrateurs de travailler sur un même système. L'administrateur doit se baser sur sa connaissance du système et sur les documents de conception qu'on lui fournit.

7. Conclusion

Les éléments manipulés au niveau de l'approche à services ont un niveau de granularité qui facilite le maintien du lien entre les modèles de la conception et de l'exécution. C'est pourquoi cette approche est utilisée aujourd'hui dans de nombreux domaines applicatifs comme pour l'informatique autonome. De plus, l'approche à services permet de gérer l'hétérogénéité et le dynamisme ce qui est très utile dans le cadre d'environnement changeant et imprévisible.

Dans cet article, nous avons proposé une approche qui permet de simplifier l'administration des applications pervasives en introduisant les modèles de la conception et de l'exécution au niveau de la connaissance d'une boucle autonome. Pour faciliter la tâche de l'administrateur, nous avons également proposé un algorithme de conformité entre les différents modèles pour signaler si des modifications de l'architecture de l'exécution n'ont pas respecté des contraintes de conception. Ces modèles doivent donc être compréhensibles par l'administrateur et être utiles pour un gestionnaire autonome qui automatise certaines tâches de l'administrateur. Les modèles que nous avons proposés, ont été définis de telle sorte que l'on garde une continuité entre les différentes activités (de la conception à l'exécution) tout en gardant la spécificité de chacune des étapes du cycle de vie. Cette approche a été implantée sous la forme d'une extension de la plate-forme Eclipse pour la partie modélisation (Cilia IDE). Elle a été testée avec la plate-forme logicielle Cilia qui permet de développer des chaînes de médiation dynamiques dans le cadre du projet FUI Medical.

Bibliographie

Baresi L., Pasquale L., Spoletini P. (2010). Fuzzy Goals for Requirements-Driven Adaptation. In *RE 2010, 18th IEEE International Requirements Engineering Conference, Sydney*,

- New South Wales, Australia, September 27 - October 1, 2010*, p. 125–134. IEEE Computer Society.
- Elkhodary A. M., Esfahani N., Malek S. (2010). FUSION: A Framework for Engineering Self-Tuning Self-Adaptive Software Systems. In G. Roman, K. J. Sullivan (Eds.), *Proceedings of the 18th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2010, Santa Fe, NM, USA, November 7-11, 2010*, p. 7–16. New York, NY, USA, ACM.
- Escoffier C., Chollet S., Lalanda P. (2014). Lessons learned in building pervasive platforms. In *11th IEEE Consumer Communications and Networking Conference, CCNC 2014, Las Vegas, NV, USA, January 10-13, 2014*, p. 7–12. IEEE.
- Favre J.-M. (2004). Foundations of Model (Driven) (Reverse) Engineering: Models - Episode I: Stories of the Fidus Papyrus and of the Solarus. In *Postproceedings of Dagstuhl Seminar on Model Driven Reverse Engineering*.
- Floch J., Hallsteinsen S., Stav E., Eliassen F., Lund K., Gjørven E. (2006, March). Using Architecture Models for Runtime Adaptability. *Software, IEEE*, vol. 23, n° 2, p. 62-70.
- Garlan D., Cheng S.-W., Huang A.-C., Schmerl B., Steenkiste P. (2004, Oct). Rainbow: Architecture-Based Self-Adaptation with Reusable Infrastructure. *Computer*, vol. 37, n° 10, p. 46-54.
- Garlan D., Schmerl B. R. (2002). Model-based adaptation for self-healing systems. In *Proceedings of the First Workshop on Self-Healing Systems, WOSS 2002, Charleston, South Carolina, USA, November 18-19, 2002*, p. 27–32. ACM.
- Keohart J. O., Chess D. M. (2003). The vision of autonomic computing. *Computer*, vol. 36, n° 1, p. 41–50.
- Lalanda P., Bourcier J., Bardin J., Chollet S. (2010, février). Building smart home pervasive services. In *Smart Home Systems*, p. 1-17. InTech.
- Lalanda P., McCann J. A., Diaconescu A. (2013). *Autonomic Computing - Principles, Design and Implementation*. Springer-Verlag.
- Morand D., García I., Lalanda P. (2011). Autonomic enterprise service bus. In *IEEE 16th conference on emerging technologies & factory automation, ETFA 2011, Toulouse, France, september 5-9, 2011*, p. 1–8. IEEE.
- OMG. (2002, avril). *Meta-Object Facility (MOFTM) Specification*. (version 1.4)
- Oreizy P., Medvidovic N., Taylor R. N. (1998). Architecture-based runtime software evolution. In *International conference on software engineering*, p. 177–186. IEEE Computer Society.
- Papazoglou M. P. (2003). Service-Oriented Computing: Concepts, Characteristics and Directions. In *Proceedings of the 4th international conference on web information systems engineering*, p. 3–12.

Session Données spatiales

Recommandation opportuniste de trajectoires pour l'accomplissement de tâches dans les systèmes crowdsourcing

André Sales Fonteles, Sylvain Bouveret, Jérôme Gensel

LIG, Université Grenoble Alpes,
F-38000 Grenoble, France

{andre.sales-fonteles,sylvain.bouveret,jerome.gensel}@imag.fr

RÉSUMÉ. Les systèmes de marché Crowdsourcing (CMS) sont des plateformes qui permettent à quelqu'un de publier des tâches afin qu'elles soient accomplies par d'autres. Récemment, un type de CMS est apparu dans lequel des tâches spatio-temporelles doivent être accomplies par des personnes dans une fenêtre de temps et un lieu précis. Dans cet article, nous présentons le Problème de Recommandation de Trajectoires Utiles (PRTU), qui permet à une personne qui veut se déplacer d'accomplir des tâches spatio-temporelles pour lesquelles elle montre une grande affinité et/ou aptitude, sans compromettre son arrivée dans les temps à destination. Nous démontrons que le PRTU est NP-complet (dans sa version décisionnelle) et lui proposons un algorithme exact. En outre, nous proposons une architecture de référence pour aider dans l'emploi de la recommandation de ces trajectoires dans un CMS réel. Enfin, nos expérimentations montrent que l'algorithme proposé est une solution faisable pour les instances de PRTU ayant quelques centaines de tâches ou moins.

ABSTRACT. Crowdsourcing market systems (CMS) are platforms that allow one to publish tasks in order to be accomplished by others. Recently, a type of CMS has appeared where spatio-temporal tasks are to be accomplished by persons at a specific time-window and location. In this paper, we present the Useful Trajectories Recommendation Problem (PRTU), that allows a person to accomplish tasks he has affinity and/or ability to, without compromising his arrival in time at the destination. We show that PRTU is NP-complet (in its decision version) and propose a exact algorithm for it. Further, we propose a reference architecture for the deployment of the recommendation of such trajectories in a CMS. Our experiments have shown that our algorithm is a feasible solution for instances of PRTU with a few hundreds tasks or less.

MOTS-CLÉS : ordonnancement de tâches spatio-temporelles, recommandation de tâches, crowdsourcing spatial.

KEYWORDS: spatial task assignment, task recommendation, spatial crowdsourcing.

1. Introduction

Au cours des dernières années, de nombreux systèmes de *crowdsourcing* (CMS, pour *Crowdsourcing Market Systems*) sont apparus, dans lesquels un groupe d'utilisateurs, appelés exécutants, contribuent pour atteindre des objectifs ou résoudre des tâches. Ces systèmes s'appuient sur l'intelligence humaine pour accomplir des tâches que les ordinateurs ne sont pas capables de réaliser seuls, ou pour lesquelles ils ne sont pas toujours les plus performants (Kittur *et al.*, 2011), telles que : la traduction de texte, la transcription de l'audio vers le texte, l'étiquetage sémantique, *etc.* Amazon Mechanical Turk est probablement l'exemple le plus connu de ce type de système, auquel se rattachent également des noms comme CrowdFlower¹ et oDesk².

Plus récemment, un autre type de CMS est apparu, où les commanditaires peuvent publier des tâches spatio-temporelles qui requièrent d'être dans un lieu et à un moment spécifique. Sereale³, TaskRabbit⁴ et Medusa (Ra *et al.*, 2012) sont des exemples de tels systèmes. Par exemple, une tâche spatio-temporelle peut consister à demander à quelqu'un de se rendre à l'intersection de deux rues à un moment donné de la journée pour filmer une courte vidéo qui permettra à quelqu'un d'autre, ou même à un système, d'analyser le trafic en ce lieu, à cette heure.

Dans ce travail, nous nous concentrons sur les CMS qui comportent des tâches spatio-temporelles. Plus spécifiquement, nous nous concentrons sur le scénario suivant: une personne, prête à contribuer au CMS, souhaite voyager d'un lieu à l'autre et le CMS, de façon opportuniste, recommande une trajectoire pour l'accomplissement de tâches spatio-temporelles sans compromettre son arrivée dans les temps à destination. En plus de respecter l'heure limite d'arrivée, d'autres aspects doivent être considérés. Par exemple, chaque tâche a une fenêtre de temps lors de laquelle elle peut être réalisée. Ainsi, une trajectoire doit garantir que pour chaque tâche proposée sur le parcours en un lieu donné, l'heure d'arrivée en ce lieu est bien située dans la fenêtre temporelle associée à cette tâche. De plus, au lieu d'essayer de recommander le chemin le plus court à l'exécutant, le but du système est de recommander une trajectoire avec des tâches pour lesquelles l'exécutant montre la plus grande affinité ou aptitude⁵, et par conséquent qui ont une grande probabilité d'être acceptées. Nous appelons le problème de trouver une telle trajectoire, un Problème de Recommandation de Trajectoires Utiles, ou bien PRTU.

La suite de cet article est organisée comme suit. La section 2 est consacrée aux travaux voisins. Dans la section 3, nous décrivons formellement le Problème de Recommandation de Trajectoires Utiles et étudions sa complexité. Dans la section 4,

1. <http://www.crowdflower.com>

2. <http://www.odesk.com>

3. <https://www.sereale.fr>

4. <https://www.taskrabbit.com>

5. L'affinité (ou l'intérêt) et l'aptitude d'un exécutant pour une tâche s sont représentés par une fonction d'utilité $u(s)$

nous présentons une méthode naïve pour solutionner le PRTU, avant de proposer un algorithme exact plus performant. Dans la section 5 nous proposons une architecture de référence pour l'utilisation de la recommandation de trajectoires utiles dans un CMS réel. La section 6 présente les résultats de nos expérimentations. La section 7 conclut l'article et esquisse nos travaux futurs.

2. Travaux voisins

Dans cette section, nous présentons tout d'abord un aperçu de la littérature sur la recommandation de tâches, puis nous discutons des travaux voisins traitant de l'ordonnement de tâches dans les CMS.

Afin d'augmenter le nombre total de tâches accomplies dans un CMS, quelques chercheurs ont proposé l'utilisation de systèmes de recommandation pour suggérer à une personne des tâches ayant une utilité élevée. Ces approches ne se préoccupent toutefois pas de l'ordonnement de ces tâches. Bien qu'au premier coup d'œil il puisse sembler simple de recommander des tâches selon leurs utilités, le véritable défi de cette problématique se trouve dans l'estimation de la valeur de chaque utilité. En général, la littérature (Lin *et al.*, 2014 ; Ambati *et al.*, 2011 ; Yuen *et al.*, 2012 ; Fonteles *et al.*, 2014) propose d'estimer ces valeurs en s'appuyant sur les intérêts/préférences et compétences de l'utilisateur qui sont implicitement découverts par l'analyse de ses interactions avec le CMS. Certains travaux prônent une autre approche. Par exemple, bien que Difalla *et al.* (2013) utilisent une approche similaire d'estimation selon les préférences et compétences d'un exécutant, ces informations sont découvertes à partir de son profil dans des réseaux sociaux, et non par l'analyse de son utilisation du CMS. La différence principale entre notre problème et celui traité dans ce type de recommandation de tâches est que l'ordre dans lequel les tâches doivent être accomplies n'est pas prise en compte : lorsqu'un CMS ne traite pas de tâches spatio-temporelles, cette séquence n'est pas pertinente. Cependant, lorsque les tâches proposées présentent des contraintes de temps et localisation, l'ordonnement des tâches joue un rôle essentiel. Dans ce travail, nous utilisons le concept d'utilité de tâches pour estimer l'utilité d'une trajectoire, en faisant l'hypothèse que la première est déjà estimée (par exemple en utilisant des méthodes similaires à celles proposées dans les références mentionnées ci-dessus) et donnée comme entrée au problème PRTU.

Un autre domaine lié à notre travail est celui de l'ordonnement de tâches spatio-temporelles dans les systèmes de *crowdsourcing*. Kazemi et Shahabi (2012) proposent une approche dans laquelle le système, à partir d'un ensemble d'utilisateurs ou exécutants potentiels localisés, et d'un ensemble de tâches spatio-temporelles, tente d'attribuer une séquence de tâches aux utilisateurs qui maximise le nombre global de tâches accomplies. Un autre travail de Kazemi *et al.* (2013) a comme but la maximisation de nombre de tâches spatio-temporelles effectuées en attribuant le plus possible de tâches. Cette approche s'appuie sur la localisation des tâches et des exécutants, ainsi que sur la notion de réputation des exécutants et de réputation requise pour l'accomplissement d'une tâche. Enfin, Deng *et al.* (2013) proposent une façon de recomman-

der un ordonnancement à un exécutant, étant donnés sa localisation et un ensemble de tâches spatio-temporelles, qui maximise le nombre de tâches à accomplir. L'originalité de notre travail réside dans les caractéristiques suivantes. Premièrement, aucun de ces travaux ne prend en compte, dans la recommandation des tâches, les préférences/aptitudes d'un exécutant, représentées par la notion d'utilité d'une tâche pour un exécutant. Au contraire, ces approches tentent de satisfaire seulement le système (CMS) et les contraintes de ses tâches. De plus, les tâches spatio-temporelles traitées dans ces travaux ne sont pas associées à des périodes de temps durant lesquelles elles doivent être réalisées (le plus souvent elles ne possèdent qu'une échéance donnée). Enfin, aucun de ces travaux ne se concentre sur un scénario particulier mais fréquent dans lequel l'utilisateur se déplace d'un point origine à un point destination qu'il doit atteindre avant une échéance fixée.

3. Problème de Recommandation de Trajectoires Utiles

Une tâche s correspond à la demande (ou requête) d'un service qui doit être accompli par une personne en un lieu et une fenêtre de temps. Nous caractérisons une tâche comme un quadruplet $s = (l^s, \tau_1^s, \tau_2^s, \delta^s)$, où l^s est le lieu où la tâche doit être réalisée, $\tau_1^s \in \mathbb{N}$ et $\tau_2^s \in \mathbb{N}$ sont respectivement la date de début au plus tôt et la date de début au plus tard de la tâche (en d'autres termes, sa fenêtre de temps), et δ^s est la durée de la tâche. Nous définissons ci-dessous une instance du PRTU.

DÉFINITION 1. — Une instance du Problème de Recommandation de Trajectoires Utiles PRTU est un tuple $(\mathcal{S}, u, l_o, l_d, \tau_o, \tau_d, \mathcal{G})$, où :

- $\mathcal{S} = \{s_1, \dots, s_n\}$ est un ensemble de tâches ;
- $u : \mathcal{S} \rightarrow \mathbb{R}^+$ est une fonction d'utilité qui associe à chaque tâche s_i une utilité, $u(s_i)$, qui traduit l'intérêt de l'exécutant pour la tâche ;
- l_o est le lieu initial (origine) de l'exécutant ;
- l_d est le lieu final (destination) que l'exécutant souhaite atteindre ;
- $\tau_o \in \mathbb{N}$ est le temps au plus tôt auquel l'exécutant peut quitter le lieu l_o ;
- τ_d est le temps au plus tard (échéance) auquel l'exécutant doit atteindre son lieu de destination l_d ;
- $\mathcal{G} = (V, E, d)$ est un graphe orienté pondéré, où $V = (\{l^s | s \in \mathcal{S}\} \cup \{l_o, l_d\})$ est l'ensemble des sommets (ou bien de lieux), $E \subset V^2$ est l'ensemble de arêtes avec $v \neq v'$ pour chaque $(v, v') \in E$ et d est une fonction de coût qui associe à chaque $(v, v') \in E$ un nombre dans \mathbb{N} qui spécifie le temps nécessaire pour voyager de v à v' .

Dans ce modèle, l'aspect spatial est pris en compte de manière indirecte sous la forme du graphe sous-jacent au problème. Cela permet de s'abstraire du calcul des trajectoires exactes dans l'espace (qui peut être réalisé dans une étape préliminaire au PRTU, aboutissant au graphe des coûts), tout en gardant un formalisme suffisamment général, expressif et abstrait pour englober tout une famille de problèmes pour lesquels on peut associer un coût de passage d'une tâche à l'autre à chaque paire de tâches.

Soit $\mathcal{I} = (\mathcal{S}, u, l_o, l_d, \tau_o, \tau_d, \mathcal{G})$ une instance de PRTU. Une trajectoire pour \mathcal{I} est une séquence $T = \langle (s_1^T, t_1), \dots, (s_m^T, t_m) \rangle$ de tâches distinctes devant être accomplies, chacune étant associée à un temps d'arrivée sur son lieu d'exécution. Le temps d'arrivée est défini comme suit :

$$t_i = \begin{cases} \max(\tau_1^{s_1^T}, \tau_o + d(l_o, l^{s_1^T})) & \text{si } i = 1 \\ \max(\tau_1^{s_i^T}, t_{i-1} + \delta^{s_{i-1}^T} + d(l^{s_{i-1}^T}, l^{s_i^T})) & \text{si } i > 1, \end{cases}$$

où \max représente le fait que l'exécutant doit attendre l'instant au plus tôt $\tau_1^{s_i^T}$ afin d'accomplir la tâche s_i^T .

Une trajectoire T est *valide* si et seulement si elle satisfait toutes les conditions suivantes :

1. pour toute $(s_i^T, t_i) \in T$, $t_i \in [\tau_1^{s_i^T}, \tau_2^{s_i^T}]$ (une tâche doit être accomplie dans sa fenêtre de temps), et
2. $t_m + \delta^{s_m^T} + d(l^{s_m^T}, l_d) \leq \tau_d$ (l'exécutant doit arriver à sa destination avant l'échéance).

Pour souci de simplicité, dans la suite de cet article, nous assimilons une trajectoire à une séquence de tâches sans les temps d'arrivée. Par exemple, soient A, B et C des tâches et soit $T = \langle (A, t_A), (C, t_C), (B, t_B) \rangle$ un exemple de trajectoire comprenant ces tâches. La trajectoire T peut être également représentée par $\langle A \rightarrow C \rightarrow B \rangle$.

Étant donné une trajectoire T , l'utilité de T correspond alors simplement à la somme de toutes les utilités associées à chaque tâche présente dans la trajectoire, à savoir :

$$u(T) = \sum_{(s_i^T, t_i) \in T} u(s_i^T)$$

Maintenant, nous pouvons assembler les concepts précédents et définir le Problème de Recommandation de Trajectoires Utiles de la manière suivante :

PRTU	
Entrée :	Un tuple $(\mathcal{S}, u, l_o, l_d, \tau_o, \tau_d, \mathcal{G})$.
Sortie :	Une trajectoire T valide d'utilité maximale parmi l'ensemble des trajectoires valides.

3.1. Complexité du problème

La complexité du problème PTRU dépend directement de celle des fonctions u et d . Sous l'hypothèse raisonnable que ces fonctions sont calculables en temps déterministe polynomial (hypothèse que nous ferons implicitement dans toute la suite

du document), nous pouvons montrer que la version décisionnelle du PRTU est NP-complète, c'est-à-dire que sous l'hypothèse $P \neq NP$, la résolution du PRTU est théoriquement difficile.

PROPOSITION 2. — *Étant donné un tuple $(\mathcal{S}, u, l_o, l_d, \tau_o, \tau_d, \mathcal{G})$ et un entier k , décider s'il existe une trajectoire valide T , telle que $u(T) \geq k$, est un problème NP-complet.*

PREUVE 3. — L'appartenance à NP est immédiate. On peut voir simplement qu'étant donné que d et u sont calculables en temps polynomial, on peut vérifier en temps polynomial également qu'une trajectoire T est valide et a une utilité supérieure à un entier k .

La NP-difficulté du problème peut se montrer par réduction depuis le problème du voyageur de commerce (TSP pour *Travelling Salesperson Problem*), dont la version décisionnelle peut s'énoncer comme suit :

TSP	
Entrée :	Un ensemble $C = c_1, \dots, c_m$, une fonction de distance $d : C \times C \rightarrow \mathbb{N}$, et un nombre entier k .
Question :	Existe-t-il une permutation σ de $[1, m]$ telle que $d(c_{\sigma(m)}, c_{\sigma(1)}) + \sum_{i=1}^{m-1} d(c_{\sigma(i)}, c_{\sigma(i+1)}) \leq k$?

À partir d'une instance $((C, d), k)$ de TSP, nous pouvons créer une instance $((\mathcal{S}, u, l_o, l_d, \tau_o, \tau_d, \mathcal{G}), k')$ de la version décisionnelle du PRTU comme suit :

- $\mathcal{G} = (V, E, d')$ est un graphe complet entre m sommets (v_1, \dots, v_m) , et $d'(v_i, v_j) = d(c_i, c_j)$;
- $\mathcal{S} = \{s_1, \dots, s_m\}$, où $l^{s_i} = v_i$, $\tau_1^{s_i} = 0$, $\tau_2^{s_i} = k$ et $\delta^{s_i} = 0$;
- $u(s_i) = 1$ pour tout i ;
- l_o peut être n'importe quel sommet du graphe;
- $l_d = l_o$;
- $\tau_o = 0$;
- $\tau_d = k$;
- $k' = m$.

Supposons qu'il existe une trajectoire valide T pour cette instance de PRTU ayant une l'utilité supérieure ou égale à $k' = m$. C'est-à-dire que toutes les tâches dans T sont présentes dans la trajectoire. Étant donnée la définition d'une trajectoire valide et de son utilité, on peut affirmer que (i) la trajectoire passe sur tous les sommets du graphe, et (ii), la durée totale de la trajectoire est inférieure à k (l'échéance). Cela correspond à la solution pour l'instance $((C, d), k)$ du problème TSP initialement présenté.

Inversement, supposons qu'il existe une solution σ pour l'instance initiale de TSP. Maintenant, supposons, sans perte de généralité que $v_{\sigma(1)} = l_o$ et que $s_{\sigma(i)}$ soit une tâche dont la localisation est $v_{\sigma(i)}$. Alors, on peut facilement voir que la trajectoire :

$$\langle s_{\sigma(1)} \rightarrow s_{\sigma(2)} \rightarrow s_{\sigma(3)} \rightarrow \cdots \rightarrow s_{\sigma(m)} \rangle$$

est une trajectoire valide (ayant une durée totale d'au plus k). ■

4. Algorithmes

Une approche de type « force brute » pour résoudre le PRTU consiste à parcourir l'ensemble des trajectoires possibles pour en extraire celles qui sont valides, et déterminer, parmi ces trajectoires, celle qui a l'utilité la plus élevée. Bien que l'algorithme de force brute garantisse la détermination d'une solution exacte au problème, son exécution reste très coûteuse en temps. En effet, le nombre total de trajectoires qui doivent être créées et analysées par l'ordinateur est la permutation de k tâches parmi un total de n , où $n = |S|$ et k variant à partir de 1 à n , depuis une trajectoire à une seule tâche jusqu'à une trajectoire à n tâches.

$$\sum_{k=1}^n \frac{n!}{(n-k)!} = \frac{n!}{0!} + \frac{n!}{1!} + \cdots + \frac{n!}{(n-1)!}$$

Par exemple, à partir d'un ensemble $S = \{s_1, s_2\}$ peut être dérivé un total de 4 trajectoires possibles : 2 avec deux tâches ($k = 2$): $\langle s_1 \rightarrow s_2 \rangle$ and $\langle s_2 \rightarrow s_1 \rangle$; et 2 avec seulement une tâche ($k = 1$): $\langle s_1 \rangle$ and $\langle s_2 \rangle$.

4.1. Algorithme exact

Dans ce scénario, nous présentons une approche améliorée qui donne aussi la réponse exacte du PRTU lorsque l'on considère qu'aucune tâche n'a d'utilité négative. L'algorithme proposé s'appuie sur le lemme suivant, dont la preuve, évidente, est omise :

LEMME 4. — *Pour toute trajectoire $T \subset T'$, $u(T') \geq u(T)$ lorsque l'utilité d'une tâche ne peut pas être négative.*

Par exemple, la trajectoire $T' = \langle A \rightarrow B \rightarrow C \rangle$ dérivée de la trajectoire $T = \langle A \rightarrow B \rangle$, i.e., $T \subset T'$, a une utilité $u(T')$ au moins aussi élevé que $u(T)$, puisque $u_A + u_B + u_C \geq u_A + u_B$. Par conséquent, si une trajectoire valide T peut être étendue par l'ajout d'une nouvelle tâche telle que la trajectoire résultante soit encore valide, la première peut être éliminée en tant que réponse au PRTU, car il existe une trajectoire $T' \supset T$ telle que $u(T') \geq u(T)$. Ainsi, pour définir la réponse au PRTU, l'algorithme trouve l'ensemble des trajectoires valides qui ne peuvent plus être étendues par l'ajout de nouvelles tâches et compare l'utilité de ces éléments. L'essence de cette approche consiste en un algorithme de parcours en profondeur qui utilise des

stratégies d'élagage spécifiques pour le PRTU. Nous présentons dans l'algorithme 1 une vue d'ensemble (la fonction principale) de l'algorithme proposé.

<p>Entrée : Une instance de PRTU Sortie : Une trajectoire d'utilité maximale</p> <pre> 1 pour $s \in \mathcal{S}$ faire 2 si $\tau_1^s > \tau_d$ OU $\tau_2^s < \tau_o$ alors 3 $\mathcal{S} = \mathcal{S} - \{s\}$; 4 fin 5 fin 6 pour $s \in \mathcal{S}$ faire 7 Trajectoire T = new Trajectoire(); 8 Ensemble tâchesCand = \mathcal{S}.cloner(); 9 expand(T, s, tâchesCand, u, l_o, τ_o, l_d, τ_d, *meilleureT, \mathcal{G}); 10 fin 11 retourner *meilleureT;</pre>
--

Algorithme 1 : Algorithme proposé

Dans la fonction principale, la première action effectuée consiste à élaguer l'ensemble de tâches \mathcal{S} en supprimant toutes celles qu'il est impossible d'accomplir pendant la période $[\tau_o, \tau_d]$. Puis, pour chaque trajectoire conservée dans l'ensemble, la fonction la plus importante de l'algorithme est exécutée : *expand*. La fonction *expand* se charge de créer, ou trouver, de manière récursive, toutes les trajectoires valides dérivées à partir d'une trajectoire de base. L'intuition derrière *expand* est simple. Tout d'abord, elle tente d'ajouter une nouvelle tâche à la fin de la trajectoire de base et puis vérifie si la nouvelle trajectoire est valide. Si elle l'est, la fonction tente de l'étendre davantage récursivement. Si elle ne peut pas être étendue, l'algorithme la compare avec la meilleure trajectoire trouvée jusqu'alors et élague toutes les trajectoires dérivées, puisqu'elles sont invalides. Nous présentons dans l'algorithme 2 la fonction *expand* en détails.

Tout d'abord, la fonction *expand* vérifie s'il est possible pour la personne (l'exécutant) de se déplacer à partir de la localisation, appelée *position*, de la dernière tâche de la trajectoire jusqu'à la nouvelle tâche, de l'accomplir et d'arriver à destination avant l'échéance τ_d . Si ce n'est pas possible, la tâche analysée est retirée de l'ensemble des tâches candidates (*tâchesCand*) et de ses dérivées. Cette décision est prise parce que si une tâche E ne peut pas être ajoutée à la fin d'une trajectoire $\langle A \rightarrow C \rightarrow B \rangle$ sans compromettre l'échéance, alors elle ne peut pas non plus être ajoutée à la fin d'une version étendue $\langle A \rightarrow C \rightarrow B \rightarrow D \rangle$. Ainsi, cette tâche n'est jamais analysée à nouveau lorsque l'algorithme tente plus tard d'étendre la trajectoire de base ou ses dérivées.

Dans le cas où une nouvelle tâche s' est ajoutée à la fin d'une trajectoire T sans compromettre l'échéance, la fonction vérifie si l'exécutant qui suit la trajectoire T est capable d'arriver dans la fenêtre de temps $[\tau_1^{s'}, \tau_2^{s'}]$. Si oui, la nouvelle tâche est

```

1 Fonction expand(T, s, tâchesCand, u, position,  $\tau$ ,  $l_d$ ,  $\tau_d$ , *meilleureT,  $\mathcal{G}$ )
2    $\tau = \max((\tau + \mathcal{G}.\text{distance}(\text{position}, l^s)), \tau_1^s)$ ;
3    $d\text{TâcheDestination} = \mathcal{G}.\text{distance}(l^s, l_d)$ ;
4   si  $\tau + \delta^s + d\text{TâcheDestination} \leq \tau_d$  alors
5     si  $\tau \leq \tau_2^s$  alors
6       position =  $l^s$ ;
7       T.ajouter(s,  $\tau$ );
8        $\tau = \tau + \delta^s$ ;
9       tâchesCand = tâchesCand.cloner();
10      tâchesCand = tâchesCand - {s};
11      peutÉtendre = faux;
12      pour  $s' \in \textit{tâchesCand}$  faire
13        peutÉtendre = peutÉtendre OU expand(T.cloner(),  $s'$ ,
14        tâchesCand, u, position,  $\tau$ ,  $l_d$ ,  $\tau_d$ , *meilleureT,  $\mathcal{G}$ );
15      fin
16      si NOT peutÉtendre alors
17        si *meilleureT = null OU  $u(T) > u(*meilleureT)$  alors
18          *meilleureT = T;
19        fin
20      fin
21      retourner vrai;
22    fin
23  sinon
24    tâchesCand = tâchesCand - {s};
25  fin
26  retourner faux;

```

Algorithme 2 : Fonction *expand*

ajoutée à la fin de la trajectoire, retirée de *tâcheCand* et la fonction est exécutée à nouveau récursivement pour chaque tâche encore dans *tâcheCand*. Finalement, si la trajectoire n'est étendue par aucune tâche candidate, l'algorithme compare son utilité avec l'utilité de la meilleure trajectoire (*meilleureT*) trouvée jusqu'alors. Dans le cas où la nouvelle trajectoire a une utilité plus élevée, elle devient la meilleure et ainsi de suite.

On peut observer que d'autres stratégies d'élagage peuvent être considérées dans l'algorithme. Par exemple, la meilleure utilité possible est celle d'une trajectoire qui passe par toutes les tâches $s \in \mathcal{S}$ (dans n'importe quelle ordre). Lorsqu'une telle trajectoire est trouvée, l'algorithme peut s'arrêter.

Exemple: Considérons une instance de PRTU comme le montre la figure 1 où $\mathcal{S} = \{A, B, C\}$, $\tau_o = 1:15$ pm et $\tau_d = 2:00$ pm. Dans ce cas, l'algorithme tente

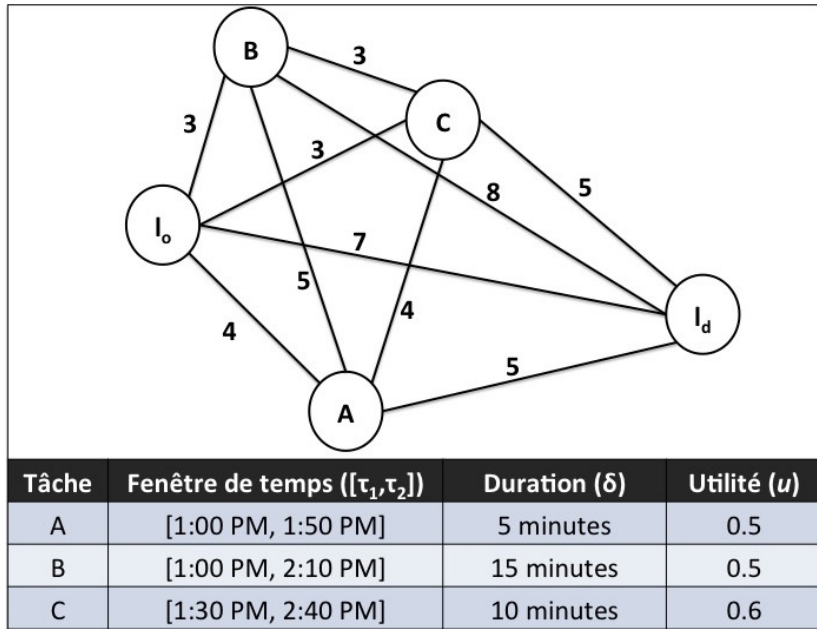


Figure 1. Configuration de tâches dans un exemple d'instance de PRTU

d'abord de retirer de S toutes les tâches ayant une fenêtre qui n'a pas d'intersection avec $[1:15 \text{ pm}, 2:00 \text{ pm}]$ ($[\tau_o, \tau_d]$), aucune dans cet exemple. Ensuite, pour chaque tâche de S , la fonction *expand* est exécutée avec la tâche elle-même et une trajectoire vide comme paramètre. Lorsque la fonction est exécutée avec la tâche A , l'algorithme vérifie si la trajectoire $\langle A \rangle$, résultat de l'ajout de A à la fin de la trajectoire vide, est valide. Comme c'est le cas, il essaie d'étendre la trajectoire en exécutant *expand* récursivement pour chaque tâche $s \in \{B, C\}$ en utilisant $\langle A \rangle$ comme trajectoire de base. La trajectoire $\langle A \rightarrow B \rangle$ est aussi valide et par conséquent l'algorithme essaie de l'étendre à son tour. Cependant, cette fois, $\langle A \rightarrow B \rightarrow C \rangle$ est invalide parce que l'exécutant ne peut pas suivre cette trajectoire et arriver en l_d avant l'instant t_d . Puisque $\langle A \rightarrow B \rangle$ ne peut pas être étendue et qu'aucune autre trajectoire n'a encore été analysée, elle devient la meilleure trajectoire trouvée jusque là. Puis, la récursion revient à $\langle A \rangle$ et essaie de l'étendre en $\langle A \rightarrow C \rangle$, qui est aussi valide et ne peut plus être étendue. Son utilité est donc comparée avec celle de $\langle A \rightarrow B \rangle$ et comme elle est plus élevée, $\langle A \rightarrow C \rangle$ devient la meilleure trajectoire. Ensuite, la récursion amène à considérer une nouvelle recherche avec $\langle B \rangle$ comme trajectoire de base. Une nouvelle meilleure trajectoire est trouvée : $\langle B \rightarrow C \rightarrow A \rangle$. Finalement, l'algorithme cherche des trajectoires meilleures en utilisant $\langle C \rangle$ comme trajectoire de base, mais il n'en trouve aucune et $\langle B \rightarrow C \rightarrow A \rangle$ est renvoyée comme réponse au PRTU. La figure 2 présente l'espace complet de recherche de trajectoires pour cet exemple. Les tâches grises foncées sont des tâches qui font partie de trajectoires invalides qui ont été élaguées alors que les grises claires font partie aussi de trajectoires invalides, mais ont été analysées. Bien que dans cet exemple seulement une trajectoire invalide a été

élaguée, nos expérimentations avec des ensembles de données synthétiques nous ont montré que cet algorithme est bien plus performant que l'approche de type force brute lorsque le nombre de tâches est élevé.

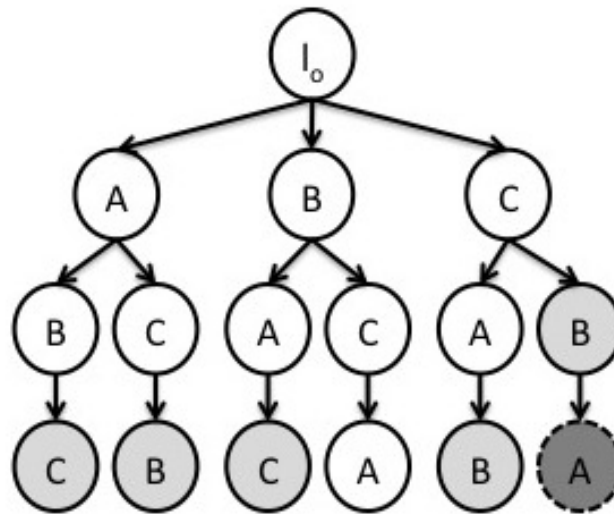


Figure 2. Espace de recherche de trajectoires.

5. Architecture de référence pour l'emploi de la recommandation

Dans cette section, nous proposons une architecture de référence pour l'emploi de la recommandation de trajectoires utiles dans un CMS réel. La figure 3 présente cette architecture, qui consiste en les principales entités et composants qui jouent un rôle important dans la recommandation de trajectoires utiles : le Gestionnaire de Contexte, le Gestionnaire de RTU, la Base de Données et l'exécutant.

Afin de proposer une recommandation, un algorithme pour le PRTU doit être muni d'une quantité considérable d'informations. Une partie de cette information peut être facilement récupérée à partir de la Base de Données du CMS. Par exemple, pour récupérer l'ensemble de tâches \mathcal{S} , une requête peut être effectuée pour demander toutes les tâches qui sont géographiquement et/ou temporellement pertinentes (une tâche au Japon ne concerne pas une trajectoire qui se passe dans une ville située en France). Les informations qui regardent l'exécutant requièrent des opérations plus sophistiquées pour être découvertes. Par exemple, l'intention de voyage $(l_o, \tau_o, l_d, \tau_d)$ d'un utilisateur peut être obtenue par l'analyse d'un agenda virtuel, ou inférée par l'histoire de ses positions sur le GPS, ou nombreuses techniques différentes. Le rôle du Gestionnaire de Contexte est de faire face à ces complexités en observant l'utilisateur en permanence, détecter toute nouvelle intention de voyage (c.-à-d. : un changement de contexte) et notifier le Gestionnaire de RTU en passant l_o, τ_o, l_d, τ_d comme paramètres.

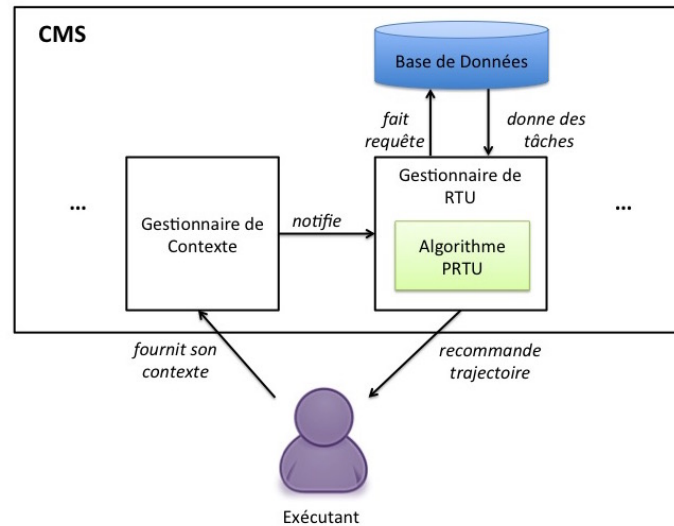


Figure 3. Architecture de référence pour l'emploi de la recommandation.

Le Gestionnaire de RTU (**R**ecommendation de **T**rajectoires **U**tiles) est le composant ayant l'implémentation d'un algorithme pour le PRTU. Lorsque ce composant est notifié à propos d'une intention de voyage en recevant l_o, τ_o, l_d, τ_d , il fait une requête à la Base de Données en demandant toutes les autres informations pour l'exécution de l'algorithme, (c'est-à-dire, les autres paramètres de l'instance PRTU). Puis, il exécute l'algorithme et envoie une alerte qui recommande le résultat à l'exécutant, qui à son tour peut accepter ou non la proposition.

6. Expérimentations

Afin d'analyser la faisabilité de l'utilisation de notre algorithme pour solutionner le PRTU, nous avons réalisé deux expérimentations sur des jeux de données synthétiques. Ces données ont été obtenues à travers un programme qui, à partir d'un nombre de tâches donné en entrée, produit en sortie une instance de PRTU générée de façon aléatoire. Dans tous les jeux de données, nous avons fixé la fenêtre de temps entre l'origine et la destination $[\tau_o, \tau_d]$ à 1 heure. En outre, nous nous sommes assurés que toutes les fenêtres de temps de toutes les tâches générées avaient une intersection non vide avec la fenêtre de temps de l'exécutant $[\tau_o, \tau_d]$. Nos expérimentations ont été effectuées sur un Mac Book Pro Retina avec un processeur Intel Core i7 3GHz et une mémoire vive de 8Go 1600MHz DD3. Le langage de programmation utilisée a été Java.

Dans notre première expérimentation, nous avons comparé le temps d'exécution de l'approche de type force brute à celui de l'algorithme. Pour cela, nous avons exécuté chacun des deux algorithmes sur 5000 instances aléatoires de PRTU. Plus spécifique-

ment, nous avons exécuté chaque algorithme 1000 fois avec des entrées aléatoires de 5 tâches, puis la même procédure pour des entrées de 6 tâches, et ainsi de suite jusqu'à 9 tâches. La figure 4 montre une comparaison entre les temps d'exécution moyens, en millisecondes, des deux approches. On peut voir qu'avec 9 tâches, le temps d'exécution pour solutionner PRTU en utilisant la force brute est déjà de 250 millisecondes, alors qu'avec notre approche il est encore proche de zéro (0,064 millisecondes). Bien que dans le pire cas (quand toutes les permutations possibles de tâches sont valides), notre approche peut générer autant de trajectoires que la force brute, dans la pratique, la stratégie d'élagage se montre sensiblement performante dans la réduction de l'espace de recherche, ce qui explique la différence entre les deux temps d'exécution.

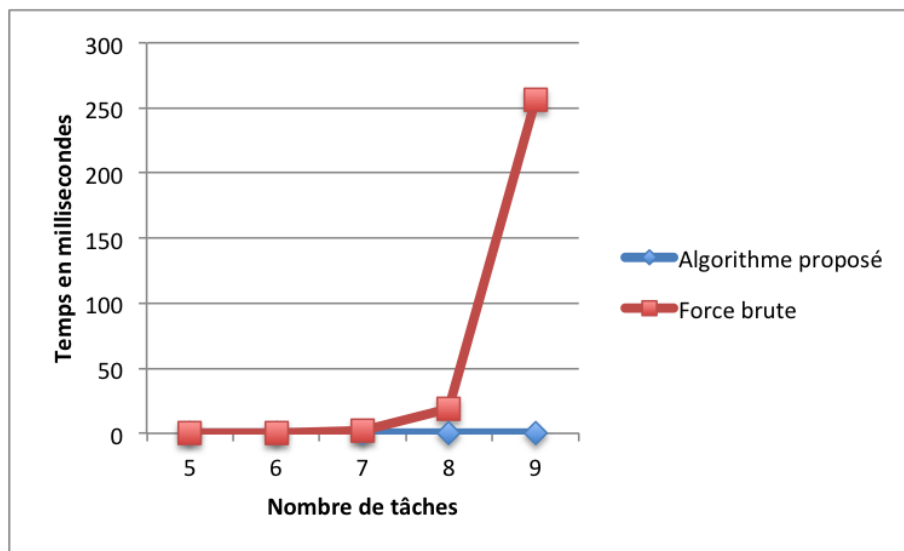


Figure 4. Comparaison entre notre approche et la force brute.

Dans notre deuxième expérimentation, nous avons analysé le temps d'exécution de notre approche prise isolément. Nous avons exécuté l'algorithme 500 fois avec des entrées aléatoires composées de 10 tâches, puis 500 fois pour des entrées de 20 tâches et ainsi de suite jusqu'à des entrées de 100 tâches. Au total, l'algorithme a été exécuté 5000 fois en faisant varier la taille des entrées. La figure 5 présente le résultat de l'expérimentation en millisecondes. Même avec 100 tâches, notre algorithme s'est montré capable de donner une réponse au PRTU en un temps moyen de 80 millisecondes. Nous avons également soumis notre algorithme à des instances de PRTU composées de 200 tâches, mais le temps d'exécution moyen s'est élevé à 2,5 secondes.

Nos expérimentations montrent que l'algorithme proposé est une solution faisable pour le PRTU offrant une réponse exacte pour des entrées comptant jusqu'à 100 tâches dans un temps moyen de moins de 80 millisecondes. De plus, si un CMS n'a pas besoin d'une réponse immédiate pour la recommandation de la trajectoire, même des entrées avec quelques centaines de tâches peuvent être traitées dans quelques secondes. Comme dit précédemment, bien que dans le pire de cas l'algorithme génère

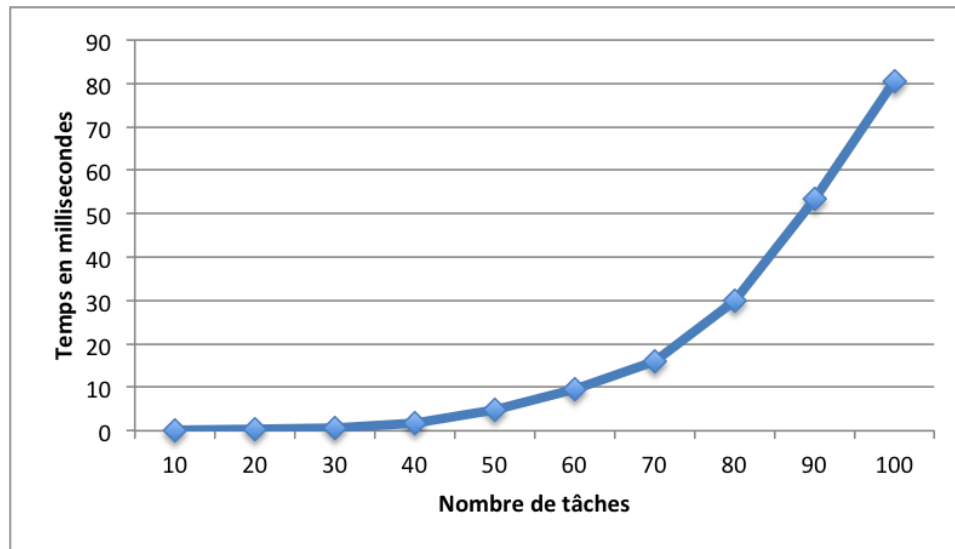


Figure 5. Temps d'exécution de notre algorithme.

autant de trajectoires que la force brute, l'expérimentation a montré que sa stratégie d'élagage réduit sensiblement l'espace de recherche et le temps d'exécution.

7. Conclusion

Dans ce travail, nous avons introduit et formalisé le Problème de Recommandation de Trajectoires Utiles (PRTU). Le PRTU permet à une personne d'accomplir des tâches pour lesquelles elle possède un intérêt et/ou des aptitudes, sans compromettre l'arrivée à sa destination avant un temps limite. Par ailleurs, nous avons prouvé que PRTU est NP-complet (dans sa version décisionnelle) et nous avons proposé un algorithme exact pour le résoudre. Enfin, nous avons aussi proposé une architecture de référence pour l'emploi de la recommandation de trajectoires utiles dans un CMS réel. Nos expérimentations ont montré que l'utilisation de notre algorithme est performant dans un scénario comportant jusqu'à cent tâches en donnant une réponse en quelques millisecondes, et quelques secondes lorsque le nombre de tâches s'élève à quelques centaines.

Comme travaux futurs, nous proposerons des algorithmes d'approximation qui donnent une réponse immédiate au PRTU lorsque plusieurs tâches sont impliquées. Nous prévoyons aussi d'expérimenter le taux d'acceptation des utilisateurs pour les trajectoires recommandées (c'est-à-dire, à quelle fréquence l'utilisateur accepte de suivre les trajectoires qui lui sont recommandées). Enfin, nous proposons d'étudier si la similarité entre une trajectoire recommandée et la trajectoire que l'exécutant emprunte normalement peut influencer ce taux d'acceptation.

Remerciements

Ce travail est soutenu par le Ministère de l'Enseignement Supérieur et de la Recherche.

Bibliographie

- Ambati V., Vogel S., Carbonell J. G. (2011). Towards task recommendation in micro-task markets. In *Human computation*.
- Deng D., Shahabi C., Demiryurek U. (2013). Maximizing the number of worker's self-selected tasks in spatial crowdsourcing. In *Proceedings of the 21st acm sigspatial international conference on advances in geographic information systems*, p. 324–333. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2525314.2525370>
- Difallah D. E., Demartini G., Cudré-Mauroux P. (2013). Pick-a-crowd: Tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on world wide web*, p. 367–374. Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee. Consulté sur <http://dl.acm.org/citation.cfm?id=2488388.2488421>
- Fonteles A. S., Bouveret S., Gensel J. (2014). Towards matching improvement between spatio-temporal tasks and workers in mobile crowdsourcing market systems. In *Proceedings of the third acm sigspatial international workshop on mobile geographic information systems*, p. 43–50. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2675316.2675319>
- Kazemi L., Shahabi C. (2012). Geocrowd: Enabling query answering with spatial crowdsourcing. In *Proceedings of the 20th international conference on advances in geographic information systems*, p. 189–198. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2424321.2424346>
- Kazemi L., Shahabi C., Chen L. (2013). Geotrucrowd: Trustworthy query answering with spatial crowdsourcing. In *Proceedings of the 21st acm sigspatial international conference on advances in geographic information systems*, p. 314–323. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2525314.2525346>
- Kittur A., Smus B., Khamkar S., Kraut R. E. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual acm symposium on user interface software and technology*, p. 43–52. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2047196.2047202>
- Lin C. H., Kamar E., Horvitz E. (2014). Signals in the silence: Models of implicit feedback in a recommendation system for crowdsourcing.
- Ra M.-R., Liu B., La Porta T. F., Govindan R. (2012). Medusa: A programming framework for crowd-sensing applications. In *Proceedings of the 10th international conference on mobile systems, applications, and services*, p. 337–350. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2307636.2307668>
- Yuen M.-C., King I., Leung K.-S. (2012). Task recommendation in crowdsourcing systems. In *Proceedings of the first international workshop on crowdsourcing and data mining*, p. 22–26. New York, NY, USA, ACM. Consulté sur <http://doi.acm.org/10.1145/2442657.2442661>

Approche Guidée par les Structures Hiérarchiques Spatiales pour la Généralisation à la Volée dans SOLAP

Ziouel Tahar* — **Khalissa Derbal**** — **Kamel Boukhalfa****

(*) *École Nationale Supérieure d'Informatique, Algérie.*

Email : t_ziouel@esi.dz

(**) *Université des Sciences et de la Technologie Houari Boumediene, Algérie*

Email : {kderbal,kboukhalfa}@usthb.dz

RÉSUMÉ. La généralisation à la volée désigne l'utilisation de techniques de généralisation automatique en temps réel. Les données générées sont temporaires et exclusivement destinées à la visualisation et non pas au stockage ou à d'autres fins. Ceci, rend ce processus bien adapté aux contextes hautement interactifs tels que la cartographie en ligne et le SOLAP. BLG tree et GAP tree sont des structures hiérarchiques spatiales largement utilisées en généralisation cartographique et plus particulièrement dans le contexte du webmapping. Cependant, ces structures s'avèrent insuffisantes dans le contexte des applications SOLAP. Dans cet article, nous proposons une approche de généralisation à la volée qui intègre l'aspect décisionnel interprété par la mesure. De nouvelles structures adaptées au contexte SOLAP sont définies: SOLAP BLG tree et SOLAP GAP tree. Le processus mis en œuvre implémente certains opérateurs de généralisation notamment la simplification et la fusion. Nos expérimentations ont été effectuées sur un jeu de données vecteurs lié au phénomène du risque routier.

ABSTRACT. On-the-fly map generalization refers to the use of automated generalization techniques in real time. The data generated is temporary and intended solely for viewing and not for storage or other purposes. This makes the process well suited for highly interactive contexts such as online mapping and SOLAP. BLG tree and GAP tree are spatial hierarchical structures widely used in cartographic generalization and particularly in the context of web mapping. However, these structures are insufficient in the context of SOLAP applications. In this paper, we propose an on-the-fly map generalization approach that integrates decisional aspect interpreted by the measure. New structures adapted to the SOLAP context are defined and now called SOLAP BLG tree and SOLAP GAP tree. The process implements generalization operators such as simplification and fusion. Our experiments were performed on a set of data vectors related to the phenomenon of road risk.

MOTS-CLÉS : Généralisation à la volée, SOLAP, Structures hiérarchiques spatiales.

KEYWORDS: On-the-fly map generalization, SOLAP, Hierarchical spatial structures.

1. Introduction

L'informatique décisionnelle permet l'exploitation des données d'une organisation afin de faciliter la prise de décision. Elle a connu, et continue de connaître encore aujourd'hui, un essor important.

Les Systèmes d'Information (SI) classiques avaient pour simple vocation la production de données. Les entrepôts de données constituent une solution adéquate pour construire un système d'aide à la décision pour ces SI. Les analystes et décideurs de l'entreprise, peuvent ainsi, analyser interactivement et itérativement les données multidimensionnelles à un niveau détaillé ou agrégé grâce aux outils de traitement analytique en ligne, OLAP [CHA 97, KIM 02, THO 02].

Néanmoins, ces données peuvent avoir une vision géographique or, les systèmes OLAP n'offrent aucun outil permettant la gestion des données spatiales. C'est pourquoi une nouvelle technologie a émergé, issue à la fois de la technologie SIG (Système d'Informations Géographiques) et de l'OLAP, soit l'OLAP Spatial (SOLAP) [RIV 05, MAL 08].

Le SOLAP a été défini comme une plateforme visuelle conçue spécialement pour supporter une analyse spatio-temporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographiques, graphiques et tabulaires [BÉD 97]. Le SOLAP enrichit les capacités d'analyse des systèmes OLAP classiques, entre autres, la visualisation des mesures sur une carte permet de comprendre la distribution géographique d'un phénomène, de mettre en relation les différents phénomènes spatiaux par rapport aux axes d'analyse alphanumériques et de comparer ces phénomènes à diverses granularités géographiques. De plus, la composante cartographique dans l'OLAP représente une interface vers l'entrepôt de données spatiales (EDS). En d'autres termes, l'utilisateur peut accéder aux opérations de navigation multidimensionnelle à travers la simple interaction avec la composante cartographique. Dans ce contexte, l'analyse des données spatiales multidimensionnelles nécessite souvent le passage par différents niveaux de détails, ce qui permet d'étudier l'évolution d'un phénomène (fait) et permettre ainsi, une prise de décision efficace. Le processus de généralisation s'avère donc bien adapté à ce contexte aux besoins des décideurs [BIM 12]. Or, ce dernier adresse uniquement l'aspect cartographique, au détriment de l'aspect décisionnel essentiel dans toute analyse multidimensionnelle.

Largement abordé en cartographie, la généralisation à la volée consiste à générer des données temporaires à différents niveaux de détail à partir du niveau le plus détaillé. Ces données sont exclusivement destinées à la visualisation et non pas au stockage ou à d'autres fins [WEI 08], ce qui la rend bien adaptée au contexte SOLAP. Différentes approches de généralisation à la volée ont été développées [BER 12] notamment celles basées sur la manipulation des structures hiérarchiques [OOS 11, OOS 14]. À notre connaissance aucun travail n'a pris en considération l'aspect décisionnel (la mesure) dans le processus de généralisation à la volée.

Dans cet article, nous proposons une approche de généralisation à la volée adaptée au SOLAP. Elle consiste en une prise en compte simultanée de l'aspect décisionnel et l'aspect cartographique en intégrant la mesure du fait lors de la généralisation.

Le présent papier est organisé en plusieurs sections. Dans la section 2 nous présentons quelques travaux liés à la problématique abordée. La section 3 est organisée en plusieurs sous-sections, elle présente une description détaillée de l'approche proposée. Les différentes étapes de nos expérimentations ainsi les résultats obtenus sont présentés dans la section 4. Enfin nous terminons cet article par une conclusion et quelques perspectives.

2. Travaux connexes

De nombreux travaux ont abordés la généralisation depuis plus de trois décennies [SAR 07, GAF 08, STA 14]. La généralisation à la volée qui consiste à générer des cartes à différents niveaux de détails destinées exclusivement à la visualisation et non pas au stockage, a émergé avec le développement de la cartographie en ligne.

Selon [WEI 08], l'approche de généralisation à la volée suit deux voies, la première exploite des algorithmes de généralisation rapides qui permettent une génération grossière des niveaux de détail en temps réel, et la deuxième utilise des structures de données spatiales hiérarchiques.

Les algorithmes rapides de généralisation s'appuient sur des algorithmes et heuristiques simples mais efficaces. La préférence est pour les opérateurs de généralisation peu sophistiqués tels que la sélection et la simplification [BER 12]. Les travaux qui ont eu le plus d'impact dans cette approche sont ceux de [LEH 05, FOE 10] réalisés dans le cadre du projet européen GiMoDig [GIM 01]. L'objectif de ce projet est de développer et tester des méthodes pour délivrer des données spatiales aux utilisateurs mobiles par des moyens d'intégration et de généralisation en temps réel.

Vu que la généralisation engendre une hiérarchie de niveaux de détails, il est donc naturel d'exploiter des structures hiérarchiques tels que les arbres pour le stockage de la géométrie (point, ligne, polygone) relative à un objet dans le niveau de détail le plus élevé. Cette structure est enrichie d'informations supplémentaires qui reflètent l'importance du niveau de la hiérarchie, à partir de laquelle, des niveaux de détails sollicités peuvent être générés. Le processus de généralisation à la volée est ainsi accéléré grâce à l'accès rapide aux structures élaborées.

Pour chaque type de donnée spatiale, il existe une structure hiérarchique adéquate qui permet la généralisation interactive et rapide des objets géographiques. Le BLG tree (Binary Line Generalization tree) a été proposé pour la simplification des objets linéaires [VAN 91], il utilise une variante de l'algorithme de Douglas-Peucker [DOU 73]; au lieu de supprimer les sommets les moins importants, il les stocke dans la structure. Le GAP tree (Generalized Area Partitionning) a été proposé pour la sélection et la fusion des polygones [VAN 05, OOS 11, OOS 14]. Les Quadrees ont

été proposées pour la sélection, la simplification, l'agrégation et le déplacement des objets de type points [BER 12, BER 13].

Par ailleurs, et à notre connaissance, il n'existe pas de travaux qui ont abordé l'intégration de la généralisation dans le contexte SOLAP. Néanmoins certains travaux se sont intéressés à l'aspect spatial [BÉD 97, BIM 12].

Dans cet article, nous proposons d'intégrer un processus de généralisation à la volée dans le SOLAP, afin de s'adapter au niveau de détail répondant au besoin exprimé par le décideur. L'approche que nous proposons se base sur les structures BLG tree et GAP tree dédiées à la généralisation cartographique. Cependant, ces structures ne peuvent servir efficacement dans un processus de prise de décision, car elles ne considèrent pas l'information décisionnelle intrinsèque aux SOLAP. Pour mieux appréhender ce problème, nous proposons l'exemple illustré par la figure 1. La carte analysée comporte six objets, à chacun d'entre eux est associée une mesure. Par exemple, à l'objet *A* est associée la mesure 20, à l'objet *B*, la mesure 10, à l'objet *C*, la mesure 30, etc. Ce dernier possède la plus grande mesure parmi les objets présents au niveau le plus détaillé (voire figure 1.a). Lors de la réduction d'échelle, le processus de généralisation classique est déclenché. En prenant en considération seulement l'aspect topographique, l'objet *C* n'est plus visible (voire figure 1.b) malgré sa pertinence décisionnelle (plus grande mesure) vis-à-vis de l'objectif de l'analyse effectuée par le décideur.

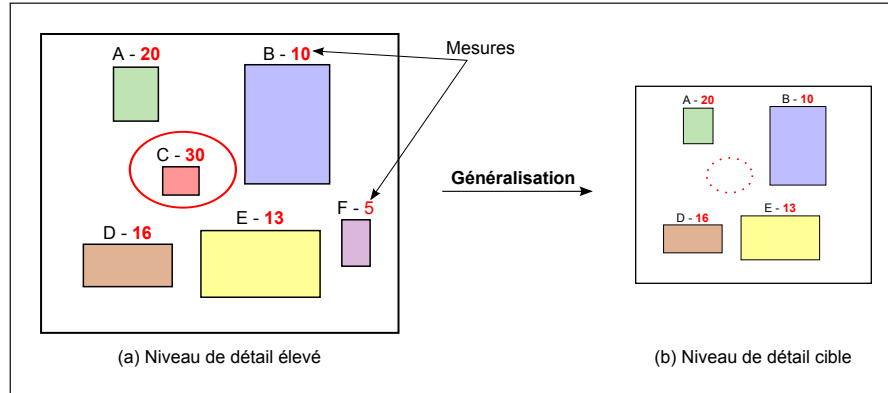


Figure 1. Exemple de résultat de généralisation classique

Pour remédier à cette limite, nous intégrons la mesure dans les structures BLG tree et GAP tree, que nous appelons, SOLAP BLG tree et SOLAP GAP tree. Ces structures sont utilisées dans notre approche de généralisation à la volée que nous présentons dans la section suivante.

3. Approche proposée

L'objectif principal de notre approche est de développer un système de généralisation à la volée adaptée aux applications d'aide à la décision SOLAP. Ce système doit

être capable de combiner les aspects décisionnel et cartographique afin de produire des cartes adaptées aux besoins des décideurs. La figure 2 illustre l'architecture globale de notre approche. En effet, l'entrepôt de données spatiales stocke des données de type décisionnel (mesure du fait) et des données de type cartographique. Ces derniers sont représentés dans un seul et unique niveau de détail (le niveau de détail le plus élevé). Lorsque l'utilisateur émet sa requête, le résultat sera initialement extrait des données stockées. Ces derniers ne correspondent pas nécessairement au niveau de détail sollicité par le décideur, un processus de généralisation à la volée s'avère donc nécessaire pour adapter le résultat au besoin exprimé.

Le processus de généralisation que nous proposons est basé sur les structures SOLAP BLG tree et SOLAP GAP tree. Un ensemble de paramètres liés à l'aspect décisionnel est intégré à ces structures notamment, la fonction d'importance, d'agrégation etc. Tous ces concepts seront détaillés dans les sections suivantes.

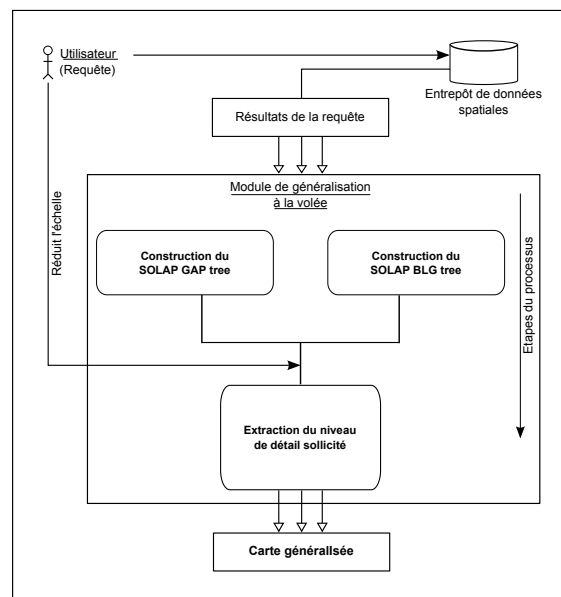


Figure 2. Architecture globale de notre approche

Les structures hiérarchiques spatiales que nous avons proposées servent à stocker les objets géographiques de la zone analysée ainsi que d'autres informations décisionnelles (les mesures). À chaque type d'objet géographique correspond une structure adéquate, le SOLAP BLG tree est proposé pour les objets de type polyligne et le SOLAP GAP tree pour les objets de type polygone. La construction de ces structures se fait selon une suite d'étapes que nous décrivons en détail pour chacune des structures proposées.

3.1. Construction du SOLAP BLG tree

La construction du SOLAP BLG tree s’articule autour de deux étapes principales : (1) l’attribution d’une valeur d’importance à chaque point de la polyligne et (2) création de la hiérarchie en considérant l’importance de chaque point.

En effet, un objet de type polyligne (route, cours d’eau, etc.) est constitué d’un ensemble de points (vertex). Le SOLAP BLG tree stocke ces différents points dans une structure hiérarchique. Chaque nœud de la structure consiste en un point de la polyligne conjointement avec sa valeur d’importance élaborée selon la fonction suivante.

L’importance $I(p)$ pour chaque point p sera calculée en fonction de sa distance $D(p)$ et de sa mesure $M(p)$ comme suit : $I(p) = f(D(p), M(p))$.

Cette fonction peut être décrite par la somme de son importance cartographique (distance) et de son importance décisionnelle (mesure) :

$$I(p) = D(p) + M(p)$$

La distance $D(p)$ est la distance orthogonale entre le segment reliant les deux points d’extrémité et le point p de la polyligne. $M(p)$ représente la mesure au point p .

Un nœud dans la structure est créé pour représenter un point p_i ayant une valeur d’importance $M(p_i)$. Ainsi, la polyligne (p_1, p_n) sera traitée comme suit : Si le nœud racine est représenté par p_k (un point de la polyligne) ayant la plus grande valeur d’importance, la création des autres nœuds se fait selon un processus itératif traitant tous les points des segments $[p_1, p_k]$ et $[p_k, p_n]$. Afin d’illustrer ce processus nous proposons l’exemple suivant relatif au phénomène d’analyse du risque routier ; nous nous intéressons aux nombres d’accidents constatés sur les segments de routes reliant dix communes représentées par les points de c_1 jusqu’à c_{10} (voir figure 3.). Chaque segment est porteur d’une mesure qui représente le nombre d’accidents relevés sur le segment reliant la commune c_i à la commune c_{i+1} .

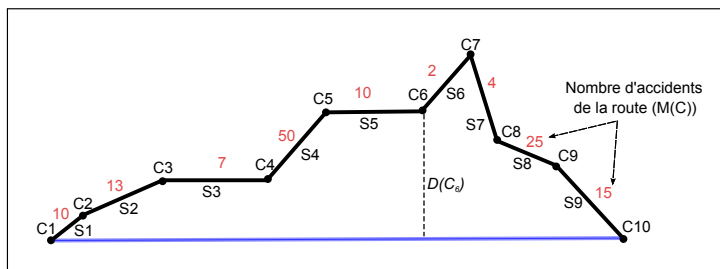


Figure 3. Représentation d’une polyligne (route) dans le contexte SOLAP

Nous soulignons que dans ce cas précis, les mesures sont associées aux segments de routes, or le traitement nécessite leurs transpositions vers les points extrémités constituant ces segments. Pour ce faire nous proposons que la valeur de la mesure du

point p_i soit déterminée par la valeur maximale des mesures des segments auxquels il appartient :

Soit $M(p_i)$ la mesure du point p_i .

$$M(p_i) = \text{Max}(M([p_{i-1}, p_i]), M([p_i, p_{i+1}]))$$

Par ailleurs, les valeurs des mesures et des distances telles que relevées sont hétérogènes. Il existe un écart important entre ces deux paramètres. Une étape de normalisation est donc nécessaire, dans le but de rendre les valeurs comparables les unes aux autres. Pour ce faire nous allons restreindre les valeurs entre 0 et 1. Pour chaque valeur V d'une mesure ou d'une distance, sa valeur normalisée V' est calculée comme suit :

$$V' = \frac{V - V_{\min}}{V_{\max} - V_{\min}}$$

Ainsi, la structure relative à l'exemple de la figure 3 se présente comme illustré dans la figure 4.

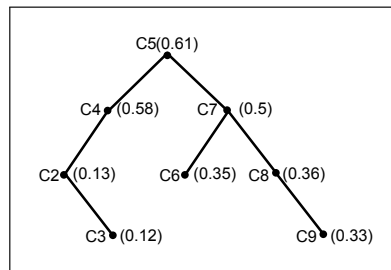


Figure 4. La structure SOLAP BLG tree construite

3.2. Construction du SOLAP GAP tree

De la même manière que le SOLAP BLG tree, la construction du SOLAP GAP tree passe par une suite d'étapes qui commence par le calcul de la valeur d'importance de chaque objet polygonal de la zone traitée.

À chaque polygone a est attribué une valeur d'importance $I(a)$ en fonction de sa surface $S(a)$, du poids de sa classe $P(a)$ déterminé par rapport à l'objectif d'analyse et de la mesure qui lui est associée $M(a)$:

$I(a) = f(S(a), P(a), M(a))$. Cette fonction d'importance, peut être décrite de la manière suivante :

$$I(a) = S(a) * P(a) + M(a)$$

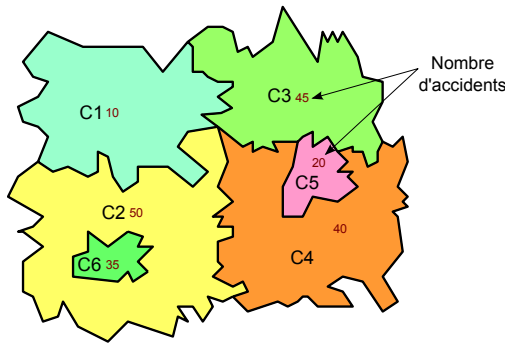


Figure 5. Zone analysée

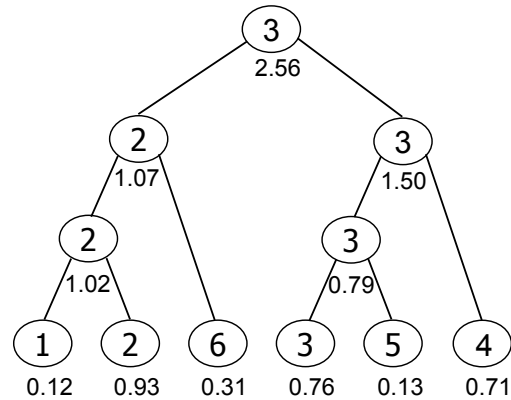


Figure 6. La structure SOLAP GAP tree construite

Ainsi définie, elle est utilisée pour trouver le polygone le moins important a , dans quel cas il sera éliminé et sa surface sera fusionnée avec un objet voisin b . Ce dernier est déterminé selon une fonction de fusion $F(a, b) = f(Bord(a, b), Sim(a, b), M(b))$, tel que :

$$F(a, b) = Bord(a, b) * Sim(a, b) * M(b)$$

Où, $Bord(a, b)$ est la longueur de la bordure commune entre deux polygones a et b . $Sim(a, b)$ est la similarité entre les types des polygones, par exemple (communes rurales, communes citadines). $M(b)$ est la valeur de la mesure d'un polygone voisin de l'élément sélectionné.

Pour illustrer la construction du SOLAP GAP tree, nous utilisons l'exemple de la figure 5 qui décrit le nombre d'accidents pour six communes. Le SOLAP GAP tree construit après normalisation est présenté dans la figure 6.

4. Processus de généralisation proposé

Le processus de généralisation à la volée adapté au contexte SOLAP est guidé par les structures SOLAP BLG tree et SOLAP GAP tree selon la méthodologie décrite ci-dessous.

Une fois ces structures construites, elles sont enregistrées au cours d'une session de travail. Lors de la navigation entre les niveaux de détail, un processus de généralisation à la volée est déclenché. Ce dernier consiste à récupérer uniquement les objets géographiques visibles dans le niveau de détail requis. La sélection de ces objets est effectuée en comparant les valeurs d'importance préalablement stockées avec une valeur seuil. Ce dernier est déterminé en fonction de l'échelle de visualisation et d'autres paramètres cartographiques qui nécessitent l'intervention d'un expert cartographe. Dans

le cadre de ce travail nous nous sommes basés sur des seuils fixés expérimentalement. En effet, le seuil est utilisé pour déterminer la profondeur de parcours des structures et en sélectionner uniquement les nœuds dont la valeur d'importance est supérieure à celui-ci.

Les mesures associées aux différents objets sont soumises à une fonction d'agrégation qui permet de déterminer les mesures des objets résultants. Ceci permet de maintenir l'importance de l'aspect décisionnel dans les différents niveaux de détail sollicités. Cette fonction d'agrégation est élaborée selon le fait analysé. À titre d'exemple pour l'analyse du phénomène du risque routier, la fonction d'agrégation proposée prend la forme d'une sommation, et ce, afin de préserver l'information sur le nombre total d'accidents relatif à l'objet généralisé.

4.1. Agrégation des mesures associées aux polygones

Le calcul des valeurs des mesures, s'effectue selon le type de l'objet traité ; dans le cas d'un objet linéaire, nous proposons la création d'une structure de données tableau, qui contiendra les valeurs des mesures associées aux différents segments constituant la polygône initiale de l'exemple de la figure 3 tel qu'illustré par le tableau 1.

Au cours de la généralisation, la polygône résultante où les segments S_i à S_k peuvent être supprimés et remplacés par un nouveau segment formé par le premier point de S_i et le dernier point de S_k . La nouvelle mesure est calculée à partir des valeurs des mesures associées aux segments S_i à S_k préalablement stockées dans le tableau 1.

Segment	S1	S2	S3	S4	S5	S6	S7	S8	S9
Mesure	10	13	7	50	10	2	4	25	15

Tableau 1. Exemple de structure de données dédiée au stockage des mesures

4.2. Agrégation des mesures associées aux polygones

Pour les objets de type polygône, les mesures sont stockées dans le SOLAP GAP tree. Ainsi, lors de la généralisation, l'agrégation des polygones entraîne l'agrégation des mesures qui leurs sont associées. La mesure $M(c)$ du nouveau polygône c résultant de l'agrégation des polygones a et b , sera égale à la somme des mesures des polygones a et b : $M(c) = M(a) + M(b)$.

Les exemples illustrés par la figure 7 décrivent quelques résultats de généralisation d'une polygône par le SOLAP BLG tree avec l'utilisation de la valeur seuil. Il faut préciser qu'une manière simple de définir le seuil est de le fixer statiquement par

l'utilisateur. Un seuil réduit implique peu de polygones/segments sont éliminés par contre un seuil plus important engendre l'élimination de plusieurs polygones/segments en même temps. Dans la figure 7.a, l'échelle demandée correspond à un seuil égal à (0.6), il n'y a que la valeur d'importance du point c5 qui soit plus grande que ce seuil. Ce point est sélectionné et affiché sur la carte avec les deux points d'extrémités de la polyligne initiale. Dans la figure 7.b, le seuil est égal à 0.48. Les points c4, c5 et c7 sont sélectionnés et affichés dans cette échelle.

De même, les exemples illustrés par la figure 6 décrivent quelques résultats de généralisation d'une région géographique par le SOLAP GAP tree.

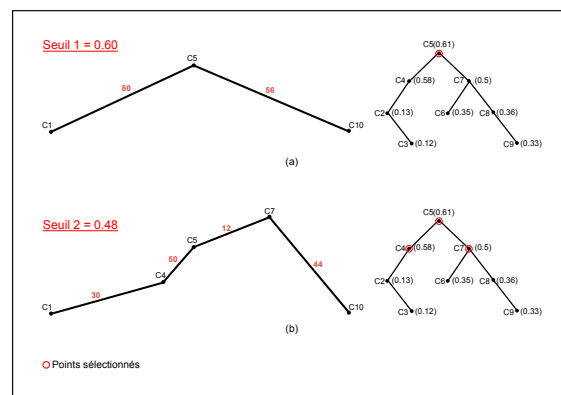


Figure 7. Résultats de la généralisation par le SOLAP BLG tree

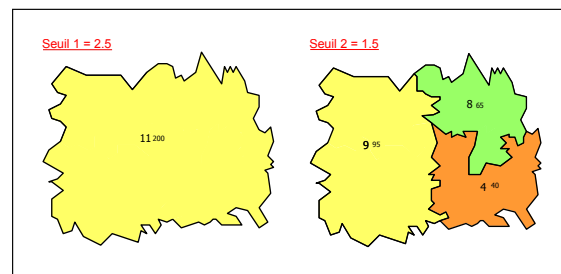


Figure 8. Résultats de la généralisation par le SOLAP GAP tree

5. Expérimentation

La validation de notre approche passe par la construction des structures spatiales proposées soient : le SOLAP BLG tree et le SOLAP GAP tree, que nous utilisons dans le processus de généralisation à la volée mis en œuvre. Nous avons opté pour le phénomène du risque routier comme sujet d'analyse, vu son impact socio-économique à l'échelle mondiale. En effet, selon les statistiques de l'organisation mondiale de la Santé [BEN 12], la route provoque chaque année plus de 1,2 million de décès et

entre 20 à 50 millions de blessés. Dans des travaux antérieurs, nous avons abordé ce phénomène en intégrant l'information spatiale [AMI 13, DER 14].

Nos tests sont effectués sur des données spatiales de type vecteur ; le thème linéaire routier et le thème de découpage administratif pour tester respectivement les structures SOLAP BLG tree et SOLAP GAP tree. Pour la réalisation de notre prototype de généralisation, nous avons eu recours à plusieurs ressources logicielles et matérielles : (1) Oracle 11G Entreprise Edition comme SGBD (Système de Gestion de Base de données) via sa composante ORACLE SPATIAL. (2) Oracle MAPVIEWER pour la visualisation de la carte géographique de la zone analysée. (3) Oracle WEBLOGIC SERVER sur lequel sont déployés les composants de MAPVIEWER. (4) L'outil Oracle MAP BUILDER a été utilisé pour charger les données géographiques dans le SGBD et la construction de la carte. (5) L'outil JDeveloper d'Oracle comme éditeur de code.

Nos expérimentations ont été réalisées sur deux jeux de données différents : le premier jeu représente le thème routier de la commune de Dar El Beida à Alger, enrichi des différentes mesures représentant le nombre d'accidents relevés sur les routes considérées (figure 9). Le second jeu représente le découpage administratif de l'Algérie où chaque commune comporte comme mesure le nombre d'accidents survenus.

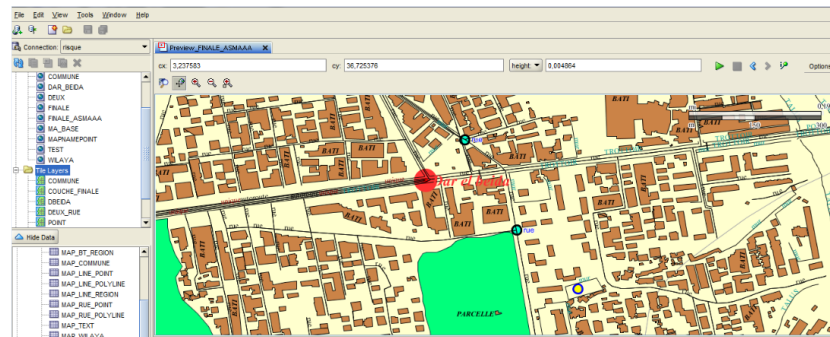


Figure 9. Zone d'étude (commune de Dair El-Beida)

5.1. Test du SOLAP BLG tree

Pour tester le SOLAP BLG tree, nous avons sélectionné une route de la commune de Dar El Beida. Cette route est illustrée en pointillés sur la figure 10. Elle comporte 28 segments chacun de ces segments comportant une mesure. Notre système de généralisation permet de simplifier cette route à des échelles plus réduites, toute en prenant en considération l'aspect décisionnel (les mesures).

La figure 11, illustre une représentation détaillée de la route sélectionnée ; les segments ayant les plus grands nombres d'accidents sont mis en évidence.



Figure 10. Réseau Routier de Dar El Beida et route sélectionnée

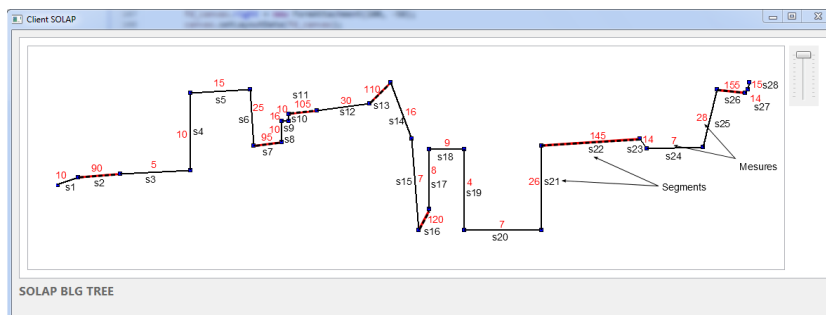


Figure 11. Représentation détaillée de la route sélectionnée

Le SOLAP BLG tree correspondant à la route sélectionnée est illustré par la figure 12. Le nœud racine contenant le point sommet p_{14} a la plus grande valeur d'importance par rapport aux autres points de la route. Les points stockés dans les niveaux hiérarchiques les plus élevés sont les points ayant les valeurs d'importance les plus élevées.

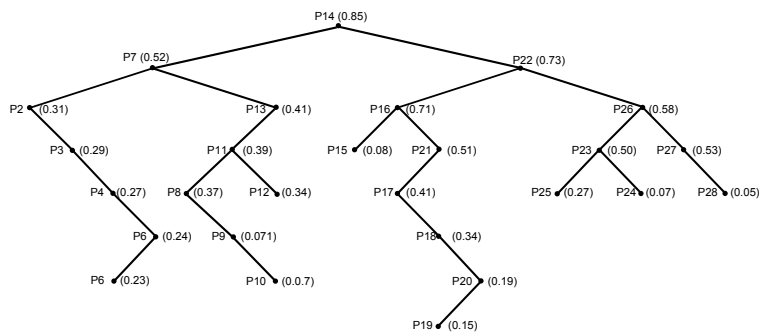


Figure 12. SOLAP BLG tree de la route sélectionnée

Selon l'approche proposée, la généralisation de la route sélectionnée permet de restituer les points des segments de route visible à l'échelle souhaitée. À partir d'une échelle plus détaillée, nous pouvons analyser les résultats obtenus à des échelles dif-

férentes (voir la figure 13). À titre d'exemple à l'échelle 1 :5000 tous les segments pertinents de la route à l'exception du segment S_2 sont présents sur la carte, par contre à l'échelle 1 :10 000, il ne reste que quatre segments pertinents et à l'échelle 1 :50 000 les segments pertinents ne sont plus perceptibles.

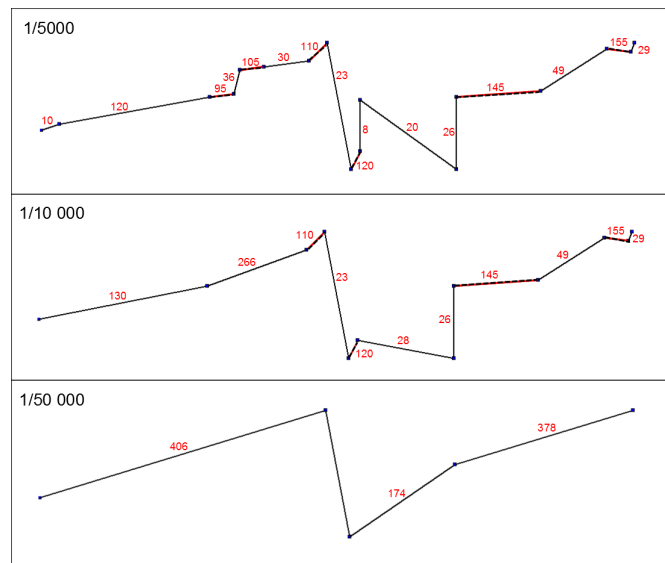


Figure 13. Résultats de la généralisation avec le SOLAP BLG tree

La figure 14, illustre une comparaison entre les résultats de la généralisation avec le SOLAP BLG tree que nous proposons dans le cadre de ce travail et les résultats de la généralisation avec le BLG tree classique. On constate qu'à la même échelle 1 :10 000, des segments de routes pertinents sont visibles dans le cas du SOLAP BLG tree (comme les segments S_{13} , S_{16} , S_{22} et S_{26}), alors qu'ils ne le sont plus dans le cas du BLG tree, et ce, malgré leur pertinence décisionnelle, d'où l'intérêt de la généralisation avec le SOLAP BLG tree dans le contexte SOLAP.

5.2. Test du SOLAP GAP tree

L'implémentation du SOLAP GAP tree nécessite la prise en compte de quelques paramètres tels que la surface des polygones, l'identification des polygones voisins, la longueur de leur bordure commune, etc. Lors de la généralisation, le polygone résultant de l'agrégation de deux polygones voisins est dotée d'une mesure calculée par application de la fonction d'agrégation définie dans la section 3.2 et qui prend la forme d'une sommation dans ce cas d'étude (risque routier). A titre d'exemple, la mesure *nombre d'accidents* associée au polygone c , $M(c)$ résultant de l'agrégation des polygones voisins a et b est : $M(c) = M(a) + M(b)$. C'est ainsi que la pertinence de l'objet du point de vue décisionnel est conservé. La figure 15, illustre notre démarche.

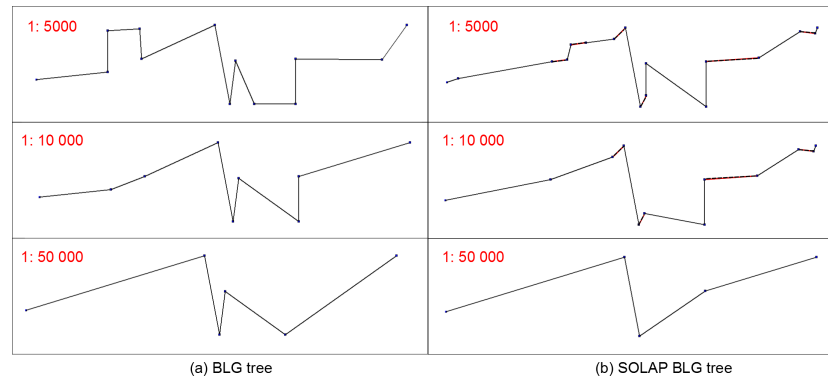


Figure 14. Comparaison des résultats de généralisation avec (a) BLG tree et (b) SOLAP BLG tree

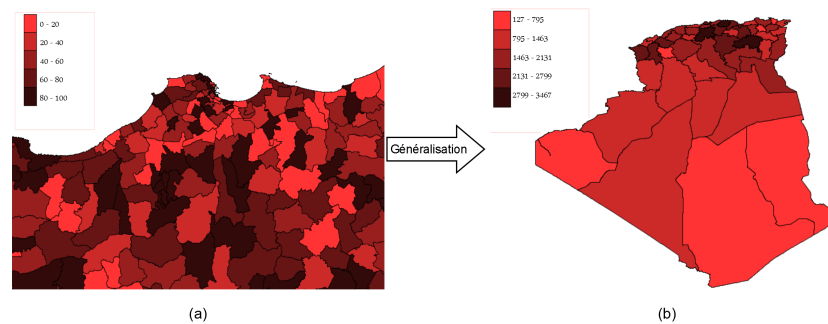


Figure 15. Résultats de la généralisation selon SOLAP GAP Tree

6. Conclusion et perspectives

Le travail présenté dans cet article, consiste en la mise en œuvre d'une approche de généralisation à la volée adapté au contexte SOLAP. La généralisation de l'information géographique, largement abordée dans le domaine cartographique, vise à améliorer la lisibilité du contenu d'une carte lors de la réduction d'échelles. Cet objectif est atteint par l'application d'une série d'opérations de généralisation basées sur des critères géométriques tels que la forme et la taille des objets géographiques. La généralisation à la volée en est une variante, adaptée aux contextes hautement interactifs tels que la cartographie en ligne et les applications SOLAP.

BLG Tree et GAP tree sont des structures hiérarchiques initialement dédiée à la généralisation cartographique à la volée. Les utiliser dans le contexte SOLAP risque de nuire à l'aspect décisionnel. Le processus de généralisation adapté aux SOLAP, que nous proposons intègre le concept mesure, essentielle dans de telles applications. Pour cela, de nouvelles structures basées sur BLG tree et GAP tree sont définis et désormais baptisées SOLAP BLG tree et SOLAP GAP tree.

Pour valider notre approche, nous avons choisi le risque routier comme phénomène à analyser vu son impact socio-économique à l'échelle mondiale. De plus, l'utilisation de la cartographie dans l'analyse de ce phénomène est d'un intérêt majeur pour les décideurs car il est intimement lié à l'information géographique représentée par le tronçon de route et la localité auquel il appartient. Dans nos expérimentations, nous avons montré l'apport des structures proposées dans le contexte SOLAP à travers les différentes fonctions mises en œuvre et dont l'application a permis de conserver les mesures tout en assurant une perceptibilité cartographique.

Comme axes de continuité à ce travail, nous proposons : (1) mener une étude expérimentale massive sur plusieurs jeux de données et différents utilisateurs afin de valider les choix effectués sur l'approche ;(2) adapter les autres opérateurs de généralisation (tel quel le lissage, le déplacement, la typification, l'exagération, etc.) aux applications SOLAP. Cela permet de générer des cartes de meilleure qualité cartographique et faciliter ainsi l'étude et l'analyse de données spatiales ; (3) Adapter le processus généralisation aux EDS en utilisant une autre approche de généralisation à la volée, développée dans le contexte du webmapping et qui se base sur les algorithmes rapides de généralisation et (4) Élaborer une étude comparative entre les approches implémentées selon des critères bien définis afin d'évaluer leurs efficacités par rapport à un contexte d'utilisation donné ;

Bibliographie

- [AMI 13] AMIEUR K. D., FRIHI I., BOUKHALFA K., ALIMAZIGHI Z., « De la Conception d'un Entrepôt de Données Spatiales à un Outil Géo-Décisionnel pour une Meilleure Analyse du Risque Routier », , 2013.
- [BÉD 97] BÉDARD Y., « Spatial olap », *Forum annuel sur la RD, Géomatique VI : Un monde accessible*, 1997, p. 13–14.
- [BEN 12] BENCHERIF H., BOUBAKOUR F., BELKACEM N., « Les accidents de la route dans les médias de masse en Algérie. Du traitement de l'information à sa diffusion », *Communication. Information médias théories pratiques*, vol. 30, n° 1, 2012, Editions Nota bene.
- [BER 12] BEREUTER P., WEIBEL R., « Algorithms for On-the-Fly Generalization of Point Data Using Quadtrees », *Proceedings AutoCarto 2012*, , 2012.
- [BER 13] BEREUTER P., WEIBEL R., « Real-time generalization of point data in mobile and web mapping using quadtrees », *Cartography and Geographic Information Science*, vol. 40, n° 4, 2013, p. 271–281, Taylor & Francis.
- [BIM 12] BIMONTE S., BERTOLOTTI M., GENSEL J., BOUSSAID O., « Spatial olap and map generalization : Model and algebra », *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 8, n° 1, 2012, p. 24–51, IGI Global.
- [CHA 97] CHAUDHURI S., DAYAL U., « An overview of data warehousing and OLAP technology », *ACM Sigmod record*, vol. 26, n° 1, 1997, p. 65–74, ACM.
- [DER 14] DERBAL K., IBTISSEM F., BOUKHALFA K., ALIMAZIGHI Z., « Spatial data warehouse and geospatial decision making tool for efficient road risk analysis », *Information and Communication Technologies for Disaster Management (ICT-DM), 2014 1st International Conference on*, IEEE, 2014, p. 1–7.

- [DOU 73] DOUGLAS D. H., PEUCKER T. K., « Algorithms for the reduction of the number of points required to represent a digitized line or its caricature », *Cartographica : The International Journal for Geographic Information and Geovisualization*, vol. 10, n° 2, 1973, p. 112–122, UT Press.
- [FOE 10] FOERSTER T., LEHTO L., SARJAKOSKI T., SARJAKOSKI L. T., STOTER J., « Map generalization and schema transformation of geospatial data combined in a Web Service context », *Computers, Environment and Urban Systems*, vol. 34, n° 1, 2010, p. 79–88, Elsevier.
- [GAF 08] GAFFURI J., « Généralisation automatique pour la prise en compte de thèmes champ : le modèle GAEL », PhD thesis, Université Paris-Est, 2008.
- [GIM 01] GIMODIG, « Geospatial Info-Mobility Service by Real-Time Data-Integration and Generalisation », 2001.
- [KIM 02] KIMBALL R., ROSS M. et al., « The data warehouse toolkit : the complete guide to dimensional modelling », *Nachdr.] New York [ua] : Wiley, , 2002.*
- [LEH 05] LEHTO L., SARJAKOSKI L. T., « Real-time generalization of XML-encoded spatial data for the Web and mobile devices », *International Journal of Geographical Information Science*, vol. 19, n° 8-9, 2005, p. 957–973, Taylor & Francis.
- [MAL 08] MALINOWSKI E., ZIMÁNYI E., *Advanced data warehouse design : from conventional to spatial and temporal applications*, Springer Science & Business Media, 2008.
- [OOS 11] VAN OOSTEROM P., MEIJERS M., « Towards a true vario-scale structure supporting smooth-zoom », *Proceedings of the 14th ICA/ISPRS workshop on generalisation and multiple representation, Paris*, vol. 48, 2011.
- [OOS 14] VAN OOSTEROM P., MEIJERS M., STOTER J., ŠUBA R., « Data Structures for Continuous Generalisation : tGAP and SSC », *Abstracting Geographic Information in a Data Rich World*, p. 83–117, Springer, 2014.
- [RIV 05] RIVEST S., BÉDARD Y., PROULX M.-J., NADEAU M., HUBERT F., PASTOR J., « SOLAP technology : Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data », *ISPRS journal of photogrammetry and remote sensing*, vol. 60, n° 1, 2005, p. 17–33, Elsevier.
- [SAR 07] SARJAKOSKI L., « Conceptual models of generalisation and multiple representation », *Generalisation of geographic information : Cartographic modelling and applications. Amsterdam, The Netherlands : Elsevier, , 2007.*
- [STA 14] STANISLAWSKI L. V., BUTTENFIELD B. P., BEREUTER P., SAVINO S., BREWER C. A., « Generalisation Operators », *Abstracting Geographic Information in a Data Rich World*, p. 157–195, Springer, 2014.
- [THO 02] THOMSEN E., *OLAP solutions : building multidimensional information systems*, John Wiley & Sons, 2002.
- [VAN 91] VAN OOSTEROM P., « The reactive-tree : A storage structure for a seamless, scale-less geographic database », *AUTOCARTO-CONFERENCE-*, vol. 6, ASPRS AMERICAN SOCIETY FOR PHOTOGRAMMETRY AND, 1991, p. 393–393.
- [VAN 05] VAN OOSTEROM P., « Variable-scale topological data structures suitable for progressive data transfer : The GAP-face tree and GAP-edge forest », *Cartography and Geographic Information Science*, vol. 32, n° 4, 2005, p. 331–346, Taylor & Francis.
- [WEI 08] WEIBEL R., BURGHARDT D., « Generalization, on-the-Fly », *Encyclopedia of GIS*, p. 339–344, Springer, 2008.

Session Ingénierie des exigences

Combining Models, Diagrams and Tables for Efficient Requirements Engineering: Lessons Learned from the Industry

Christophe Ponsard¹, Robert Darimont², Arnaud Michot¹

1. *CETIC Research Centre*

Rue des frères Wright 29/3, 6041 Gosselies, Belgique

{christophe.ponsard,arnaud.michot}@cetic.be

2. *Respect-IT S.A.*

Place des Peintres, 7/201, 1348 Louvain-la-Neuve, Belgique

robert.darimont@respect-it.be

ABSTRACT. Requirements Engineering (RE) involves a number of artifacts of different nature (models, structured text, tables, diagrams). Capturing and maintaining the relationships between those artifacts is currently not an easy task but definitely helps in both improving the quality and level of automation of requirements specification. Through several industrial applications, this paper looks at how the interactions between RE artifacts happens in practice and how to provide better tool support for them. Our experimental framework used KAOS model-based goal-oriented approach to investigate different industry use scenarios like identifying concepts within source documents or producing requirements/evaluation/cost estimates tables from the RE model. As a result of our findings, we have implemented related extensions to a goal-oriented RE tool.

RÉSUMÉ. L'ingénierie des exigences (IE) implique de nombreux artefacts de nature diverse tels que modèles, texte structuré, diagrammes et tables). Capturer et maintenir les relations entre ces artefacts reste une tâche difficile mais qui a un impact véritable à la fois sur la qualité et le niveau d'automatisation de la production d'un cahier des charges. Sur base de nombreuses applications industrielles, ce papier décrit comment les interactions entre ces différents artefacts s'organisent en pratique et comment leur fournir un meilleur soutien outillé. Notre cadre expérimental s'est appuyé sur la méthodologie orientée but KAOS afin d'étudier différents scénarios industriels tels que l'identification de concepts dans les documents sources ou la production de tables d'exigences/évaluation/estimation de coûts. Le résultat de cette étude à mené à l'implémentation d'extensions dans un outil d'ingénierie d'exigence.

KEYWORDS: Requirements Engineering, goal-orientation, industry case studies, tool support

MOTS-CLÉS: Ingénierie des exigences, orientation buts, cas d'étude industriels, support outillé

1. Introduction

Requirements Engineering (RE) is a process involving many activities such as elicitation, analysis, specification, validation and management as shown in Figure 1. Several artifacts are involved in this process either as input, output or internal part of the process (Abran *et al.*, 2001)(Sommerville, 2010). Those artifacts can take various forms in order to best support the related activity. The elicitation phase typically relies on interview transcripts whiteboard pictures or existing documentation from which requirements are identified. The analysis phase will structure those requirements either directly in a text processor or using an explicit model supported by a dedicated tool. The specification phase will produce system and/or software requirements specification usually taking the form of well structured documents according to company specific templates or standards like IEEE830 (IEEE, 1998). Those documents are generally not only composed of text but also typically rely on diagrams and tables. In order to convey the information in the most convenient format for its accurate description and understanding. A number of diagram notations has been agreed upon and standardised for this purpose, for example UML use cases/sequence diagrams/statecharts, SysML requirements diagrams, Business Process Modelling Notation (BPMN). Likewise, tabular notations are very popular in order to list requirements together with their characteristics. Dedicated tabular notations are also widely used to precisely describe specific aspects of the system such as its behaviour. For example state transition table, SCR tables, or IO tables.

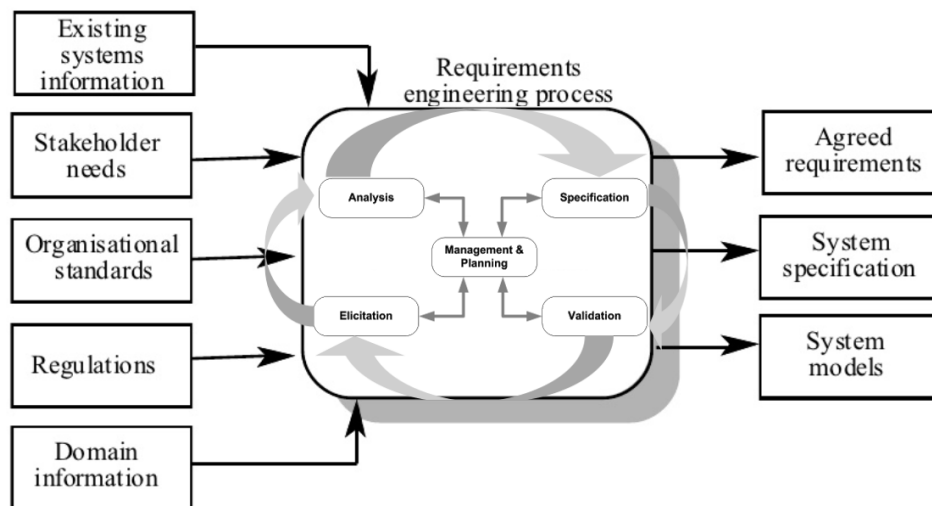


Figure 1. Requirements Engineering Process (adapted from Sommerville)

Across the different activities of the requirements engineering process, several artifacts of various forms are produced, processed, transformed and combined often iteratively. For example considering the validation of the first version of the requirements specification, some new requirements might be identified and need to be included in the requirements model. From there, specific diagrams and tables need to

be updated and re-included in an updated version of the requirements specification. Although a number of existing tools can trace the impact of such changes, they are often mostly limited to the management of documents and some kind of tables (IBM, 2002)(Dassault Systems, 2010). However they fail to address a number of scenarios, like involving diagrams, language translation, and performing quantitative processing like costs estimates.

The purpose of this paper is to present a consolidated view of interaction scenarios involving those different RE artifacts. Those scenarios were identified after conducting several real world requirements specification projects together with a systematic review of popular diagrammatic and tabular notations used in the field. Those experimentations also drove the development of specific extensions to a model based toolset supporting the KAOS goal oriented requirements engineering method (Lamsweerde, 2009)(Respect-IT, 2005). In order to ease the application of what we learned to other methods and tools, our paper describes our implementation using some abstractions for the requirements meta-model, document structure and way to query information.

Table 1. Overview of the Industrial Case Studies

Year	Domain	Location	Purpose	#reqs	#pages	Notations used
2004	Accessibility	Belgium	Domain model	185	108	goal diagrams, structured text, ranking table
2006	Grid computing	Europe	SRS	52	137	Structured text, req. tables, sequence diagrams
2007	Child care	Brussels	As-Is	126	182	BPMN, statecharts, tables (high level goals)
2009	Parliaments	Belgium	Call for tender	187	127	Structured text, req. tables, goal diagrams
2010	Banking	Brussels	internal templates	N/A	N/A	Structured text, context/goal/BMPN/information diagrams
2011	Smart Cards	Brussels	test plan	N/A	230	Command tables, Finite State Machines
2012	Child care	Brussels	Call for tender	173	267	Structured text, Event Process Chains, Use cases, req tables
2013	Electronic Nurse record	Brussels	Call for tender	223	151	Structured text, process models, req tables and diagrams
2014	Cloud computing	Europe	SRS, arch.	200	187	Structured text, req. tables, sequence diagrams
2014	Accounting	Belgium	SRS	210	170	Structured text, req tables and diagrams

As a support for our survey and experimentation work, our paper relies on about ten large requirements specifications involving big industrial players in Belgium and at international level. Table 1 give a overview of the size and contest of each case study. For confidentiality reasons, only the application domain and geographic location is mentioned. Our paper will also focus more specifically on three of those cases in order to illustrates some scenarios: the Belgian parliamentary administrations, an industrial Cloud interoperability project and a smart cart development project.

This paper is structured as follows. In section 2, we state the problem by performing a survey of popular ways to describe artifacts used within the requirements engineering process. Specific scenarios are identified and specified as user stories. In section 3, a consolidated view of all those user stories is proposed and some abstraction in order to be able to describe their relations in an implementation indepen-

dent way is introduced. In section 4, we describe the concrete implementation of the prototype we used to carry out our experimentations. In section 5, we illustrate our experimentation on a number of scenarios using excerpts of real world requirements specifications. In section 6, a critical discussion is carried out to point the achievements and areas of improvements, also in the light of related work. Finally in section 7, we draw some conclusions and discuss future work.

2. Survey of Artifacts used in the Requirements Engineering Process

This section reviews different kinds of artifacts of different forms already mentioned in the introduction. The identification is strongly based on the artifacts used in our industrial experience as described in Table 1. It is consolidated with literature material. The goal is not to be exhaustive in identifying those artifacts but rather to identify interesting ways those artifacts combine and interact within the requirements engineering process. For this purpose, we will make use of a user story template of the following form which will enable to capture both the interaction requirement but also the key actor and the underlying purpose.

As < agent >, I need < interaction_requirement > so that < goal >

2.1. Document Structure

Having a well-defined structure for the Requirements Document (RD) is important both from the RD user and RD producer. Standardised structures have been published as international standard such as the IEEE830, now replaced by the (IEEE, 1998; ISO/IEC/IEEE, 2011) or public templates like Volere (Atlantic Guild, 2014). Specific translation constraints may also apply to multilingual requirements documents. This results in the following user stories:

- US1: **As** RD User, **I need** to have a agreed well-structured document **so that** I can easily find my way in it and have confidence about the document quality.
- US2: **As** RD Producer, **I need** to have an agreed well-structured document **so that** I can make sure everything is covered and structured according to domain usage.
- US3: **As** RD Translation Manager, **I need** to extract textual output in a specific table format **so that** the translation process can be easily managed.

2.2. Diagram Notations

Diagram notations are very popular in software and system engineering and several notations have been standardised like UML for modelling software related aspects (OMG, 2011b), SysML for system-wide modelling (OMG, 2011b) or BPMN for modelling business processes (OMG, 2011a). Specific diagram notations of those standards are useful for requirements engineering such as the UML use case, UML and SysML state diagrams and BPMN. SysML also includes a simple requirements

diagram. However more elaborated requirements models are supported by specific methods such as i*/URN (Yu, Mylopoulos, 1997), KAOS (Lamsweerde, 2009) or GSN (Kelly, Weaver, 2004). Figure 2 illustrates a part of the goal model for a Cloud Management System used as case study in section 5. It shows the refinement of a goal into requirements and their assignment to specific software components.

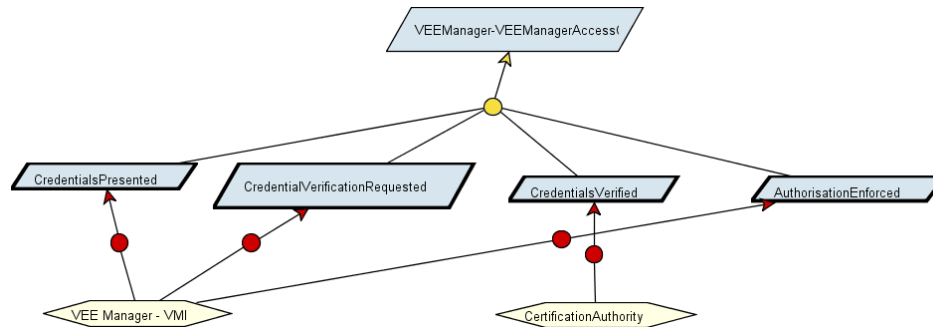


Figure 2. Example of Goal Model for a Cloud Management System

Considering the availability of such models it is also important in order to ease the way they are inserted in text documents. It also helps to automate the generation of specific byproducts, for example if the diagram is providing a specific view on information available in the requirements engineering model. This results in the following user stories:

- US4: **As** RD Producer, **I need** to be able to easily insert and update manually maintained figures in dedicated part of the RD **so that** the update effort is low.
- US5: **As** RD Produced, **I need** to be able to automate the generation of the diagrams that can be derived from information available in the RE model **so that** the update is efficient and consistency is guaranteed.

2.3. Tabular Notations

Tabular notations are very popular in requirements engineering and specification because they carry a lot of information in one structural form that the human reader can easily scan, understand and check. Just like the case of diagrams, an number of standard tabular notations are commonly used. The most common form of table is the requirements table listing requirements by tailing specific attributes such as their ID, short name, definition, priority, etc. Table 2 shows a typical requirements table found in the CCDS case.

Table 2. Example of Requirements Table

Id	Name	Definition	Resp.	Prior.	Ref.
R22	Initial Risk Assessment Provided	The Reservation Scheduling maintains for each running job a risk data structure that was initialised during the negotiation.	Risk Assessment	High	P1
R23	Dynamic Change in Risk Reported	The Extended Monitoring process monitors critical parameters about the resources and reports dynamic changes.	Extended Monitoring	High	P3

More specific notations are also widely used, such as state transition tables which describes state diagrams or SCR tables based on the Parnas four variables model or IO tables (Herrmannsdorfer *et al.*, 2008) (Hummel, Thyssen, 2009). Those tabular notations are more specific to the specification of control systems. Notice also that such tables can be redundant with diagrams and should be consistent with them. An example of a state-transition table extracted from the EMV case is shown in Figure 3.

GetProcessingOptions	Selected Initiated if SW = 9000 Selected if SW != 900	Initiated NOTALLOWED	Online NOTALLOWED	Script NOTALLOWED
InternalAuthenticate	NOTALLOWED	Initiated	NOTALLOWED	NOTALLOWED
PinChangeUnblock	NOTALLOWED	NOTALLOWED	NOTALLOWED	Script
PutData	NOTALLOWED	NOTALLOWED	Online	Script
ReadRecord	Selected	Initiated	Online	Script
Select	Selected	Selected	Selected	Selected
UpdateRecord	NOTALLOWED	NOTALLOWED	Online	Script
Verify	NOTALLOWED	Initiated	NOTALLOWED	NOTALLOWED
GetData	Selected	Initiated	Online	Script
GetChallenge	Selected	Initiated	Online	Script
GenerateAC	NOTALLOWED	Online if SW = 9000 and SW = ARQC Script if SW = 9000 and SW != ARQC Selected if SW != 9000	Script if SW = 9000 Selected if SW != 9000	NOTALLOWED
ApplicationUnblock	NOTALLOWED	NOTALLOWED	Online	Script

Figure 3. Example of State-Transition Table for the EMV Smart Card Specification

This results in the following user stories:

- US6: **As** RD Producer, **I need** to be able to automatically generate requirements tables from the RE model **so that** the document can easily be updated.
- US7: **As** RD Producer, **I need** to be able to maintain the consistency between table, RE model and related diagrams **so that** the document consistency is guaranteed.

3. Consolidated Interaction Scenarios and Problem Abstraction

3.1. Summary table

Based on the previous section, we identified four main artifact types which are: RE models, text documents, diagrams and tables. Using the identified user stories we can populate a table showing how each kind of artifact can interact with another one in an input/output relationship. Furthermore, we can also systematically question all the gaps to identify interesting missing interactions. Scenarios identified during that "gap analysis" mainly relate to model transformations, to the use of table to automate large update of specific model attributes (e.g. validated priority information) and to more quantitative reasoning typically carried out through some spreadsheet output.

The final result of this process is shown in Table 3 which references the above user stories. In this table, the lines are the source artifacts while the columns are the target artifacts. Previously identified user stories are referenced using their identification number.

Table 3. Input/Output scenario combinations

In/Out	RE Model	Text Document	Diagram	Table
RE Model	Export (package, variant)	US1/2: explanation Concept reference	US5: generate goal breakdown Responsibility...	US6:Req. table Traceability Matrix US3: translation export
Text Document	Page reference Concept creation (Model editing)	US4: illustration Extraction	–	–
Diagram	Automate update (batch import)	Referenced concepts	(Format conversion)	US7:Content listing
Table	Automate update (batch import)	US3: translation import	Layout view from query	Cost estimates

3.2. Problem Abstraction

In order to ease the description of how the above interaction scenarios can be best supported, we introduce some abstractions for the description of the different kinds of artifacts as well as some operations that can be carried out on them. This will also enable a better reuse of the experiment approach using different tools.

RE Models. The requirements engineering model is always structured according to some underlying meta-model. Such a meta-model will typically capture concepts like goals, requirements, software components (or agents), and processed information (e.g using a class model). In the scope of this paper, we will use the KAOS meta-model which is depicted in Figure 4.

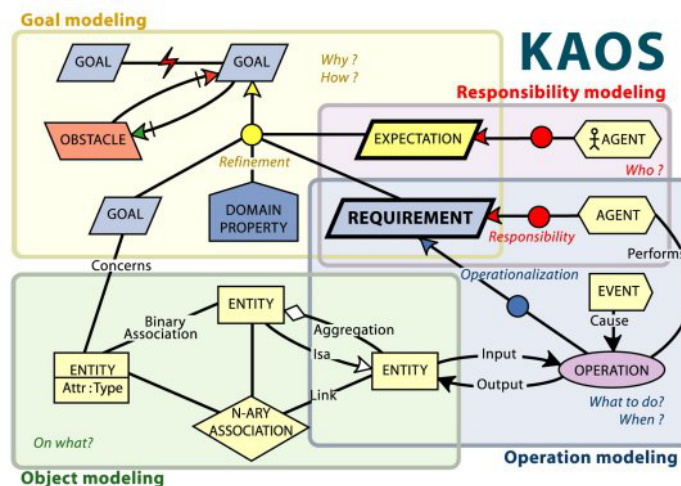


Figure 4. The KAOS Meta-model

Documents. We consider documents as a linear structure composed of a sequence of sections. Each section can start with a (sub)title and combine contents composed of texts, diagrams or tables as described by the following simple BNF grammar.

```
Document_Structure = Section*
section = [title], Content*
Content = Text | Diagram | Table
```

A typical operation is the generation of a concrete document based on a given document structure:

```
GENERATE Document_Structure AS Target_Format  
Target_Format = Text | Table
```

Diagrams can either be manually produced or automatically generated. An manually produced diagram has just some given reference whereas an automatically generated diagram results from the layout of concepts extracted from the RE model typically using some query which produces the list of identifiers to show in the diagram. A standard query language is OCL but in the scope of this paper we will use the OQL query language as it is the query language supported in our prototype and is quite similar to SQL (ODMG, 1998). Based on the identifier list, an adapted layout algorithm should be applied using a reliable layout engine (like GraphViz (AT&T Research, 2014)).

```
LAYOUT { Model_Concepts } AS Layout_Type  
Layout\_Type = Hierarchical | Orthogonal | Circular  
Model_Concepts = QUERY ON Model
```

Tables are also expressed as queries over the model as for diagrams. Those queries can specify a number of attributes that will be arranged in columns where each query result will form a line.

4. Prototype Description

4.1. Existing Tool Base

The prototype we used for our experimentation was implemented on top of the Objectiver goal oriented requirements engineering tool (Respect-IT, 2005).

Objectiver has a full support of the KAOS meta-model. It also includes a document meta-model as well as a report model. This was mapped on simple structure text in versions 2 and extended to OpenOffice (ODT format) in version 3. This also means that ODT documents can embed live references to model concepts using an URL of the type: "obj:\\unique_id". Thus a goal or requirement can be directly created from a source document or cross-referenced in some output documents. Those references are consistently and efficiently updated, including the production of a **[deleted concept]** reference in case a concept was removed from the model.

4.2. Extensions

A first extension we developed was adding spreadsheet support to the tool. We used the ODS format to keep it consistent with ODT textual format and ease the integration of both kind of content. ODS documents have a more limited meta-model than ODT which is mainly directed towards using it as output. For example, it is also not possible to embed concept reference within an ODS.

A second major work was on the report generator. This component takes as input a report instantiating the report model and produces either a linear textual document in the open document format or spreadsheet binder in the open spreadsheet format. The UNO API is used to produce those ODS based either Open Office or Libre Office (SUN Microsystems, 2009). Depending on the target format, the behaviour of the generator will be different. For example a query will result in the production of an additional spreadsheet tab in an ODS binder while it will result in the inclusion of the table in an ODT document.

Specific information is also stored back into the model for example the page where a requirement is first described is available from the model and will enable the production of traceability tables precisely pointing to that requirement.

Depending on the target format there might be some limitation for example as a binder are composed of a sequence of spreadsheets only the leaf content of the report structure is considered.

5. Experimentation

This section reports on experimentations on the interaction scenarios described in the previous section. The results of some scenarios are also reused in other scenarios to further demonstrate the chaining possibilities. In order to be concrete, we rely on the following three specific cases extracted from the list of presented earlier.

- The Parliamentary Access Control eXchange System (PACXS) enabled to share access control credential across different parliamentary assemblies in Belgium in order to share infrastructure costs and ease the authentication of deputies involved in multiple assemblies (CETIC, Respect-IT, 2009).

- a Cross Cloud Deployment System (CCDS) developed by the PaaSage FP7 project and synthesizing the requirements of four supporting case studies (PaaSage Consortium, 2013).

- The (public) specification of a smart card security module and of the agent interaction with the Europay-Mastercard-VISA environment (EMVCo, 2011) which was fed into a toolset dedicated for test plan development (Devos *et al.*, 2012).

5.1. Requirements tables

A requirements table can easily be generated using an OQL query. Depending on the structure of the report, requirements tables will be presented using different decompositions. They can follow either a goal breakdown structure, an agent breakdown structure or some other structure as recommended by standards such as the IEEE830 or ISO29148 (IEEE, 1998; ISO/IEC/IEEE, 2011). This will simply affect the way requirements are being fitted in the query. For example the following query returns all the requirements that are present in a specific goal diagram. The output of this query is similar to Table 2.


```
SELECT req.id, req.name, req.informalDef, req.priority
FROM Diagram AS d, d.elements AS g, g.concept AS req, req.directIsOf as t
WHERE d.name="MyGoalDiagram" and t.name="Requirement"
```

5.2. Generating the Traceability for Requirements Assignment to Agents

Producing a traceability matrix of how requirements(expectations) are assigned to specific software components (human agents) is important for different reasons:

- all requirements should be assigned to exactly one responsible agent
- figuring the agent load is important to assess the effort to develop it (for software) or the validate how the load can be managed by a responsible human agent

Such a traceability matrix can be produced by the simple following query with requirements as rows and agents as columns.

```
SELECT row.id, col.id
FROM Responsibility AS res, res.parent AS row, res.sons.son AS col
WHERE col.Category="Software"
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		IT-Manager	EndUser	Application-developper	SpeculativeProfiler	CrossCloudDeploymentManage	ExecutionWare	System-administrator	MetaDatabase	StochasticReasoner	HumanAnalyst	Adapter		TOTAL
71	CloudServiceIntegratedWithCustomerApplication	0	0	0	0	0	0	0	0	0	1	0		1
72	WorkflowKnown	0	0	0	0	0	0	0	0	0	0	0		0
73	DeploymentLocationPreferencesSpecified	0	0	0	0	0	0	0	0	0	1	0		1
74	DataReplicationDefined	0	0	0	0	0	0	0	0	0	0	0		0
75	NearOptimalDeploymentCalculated	0	0	0	0	0	0	0	0	1	0	0		1
76	DatabaseScalabilityDefined	0	0	0	0	0	0	0	0	0	1	0		1
77	MinCloudAdministrativeOverhead	0	0	0	0	0	0	0	0	1	0	0		1
78	LegacyApplicationsDeployed	0	0	0	1	0	0	0	0	0	0	0		1
79	TargetDeploymentEnvMappedToCloudProviders	0	0	0	0	0	0	0	0	1	0	0		1
80	CrossCloudDeployment	0	0	0	0	0	0	0	0	0	0	0		0
81														
82	TOTAL	0	0	0	7	7	0	0	0	34	17	3		

Figure 5. Responsibility Traceability Matrix for the CCDS Case Study

The resulting ODS document is shown in Figure 5 for our CCDS case. A total column is produced to easily check for the above constraints. In this case some requirements are not yet assigned to a responsible agent (see right column). Some identified agents have also currently no assigned requirements, while other are quite complex like the *StochasticReasoner*.

5.3. Generating Agent Assignment Diagrams

The DIAGRAM query to produce the list responsibility of responsibility is quite similar as the one for the traceability matrix except that it is filtering on a single agent and just returns a list of identifiers.

```
SELECT req.id
FROM Responsibility AS res, res.parent AS req, res.sons.son AS ag
WHERE ag.name="CrossCloudDeploymentManager"
```

A circular layout is then applied to this list to produce the diagram shown in Figure 6. Notice that for agents with a lot of responsibilities (e.g. more than 20) such diagrams become overloaded and too large so it is advised to keep to a table layout in that case.

```
LAYOUT { DIAGRAM } AS Circular
```

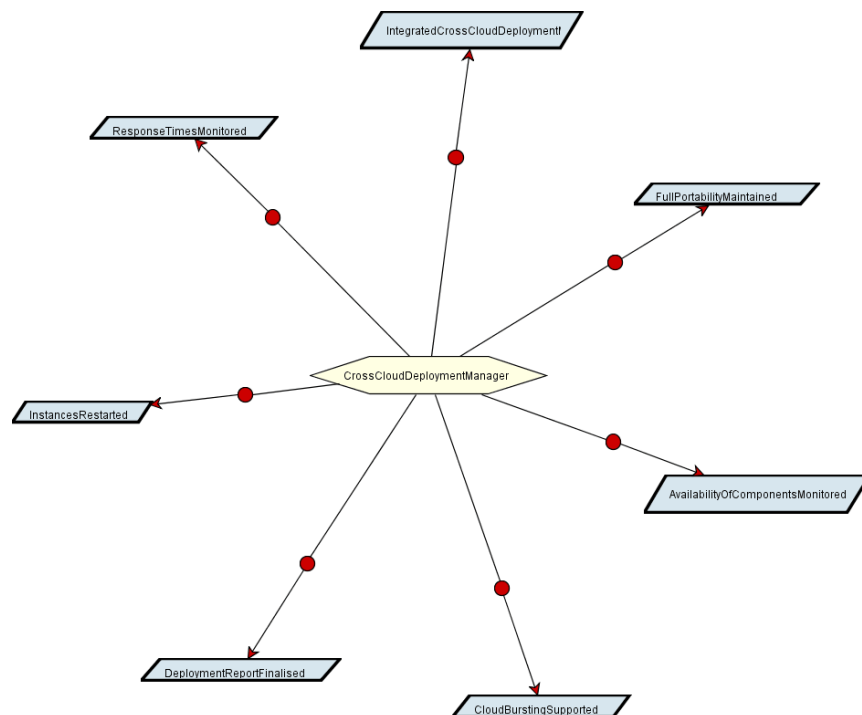


Figure 6. Responsibility Assignment for the Cross Cloud Deployment Management

5.4. Generating Evaluation Questionnaires

Another interesting scenario is to consider the procurement process, where candidates have to be evaluated against their ability to fulfil the requirements with their solution. Nice evaluation grids can be produced by grouping requirements by agents,

querying standard attributes like name or priority but also more specific attribute related to the actual reference of the requirement in the specification such as its number and page in the document.

```
SELECT req.numberInReport, req.name,
       req.priority, req.pageInReport, ag.name
FROM Responsibility AS res, res.parent AS req, res.sons.son AS ag
ORDER BY ag.name ASC
```

By using stylesheets and some macros, it is then possible to produce the kind evaluation questionnaire shown in Figure 7 where specific columns should be filled by the applicants. The returned results can then merged and compared in specific consolidation spreadsheets.

	Reference	Name	Page	Module	Coverage	Parametrisation	Standard development	Specific development	Availability	Comments
1										
2					NWLT	-/1	-/1	-/1	0i-YY	
49	2.1.4	Absences and overtime managed	36							
50	2.1.4.1	Absences managed	37							
	G8.	Global administration of the absence management system supported								
51			41							
52	R49.	Typology of absence types and status managed	42	Absence-Overtime Module	N					
53	R50.	Absences visualizable	42	Absence-Overtime Module	N					
54	R51.	Absence status overwritten by HR officers	42	Absence-Overtime Module	N					
55	R52.	Public holidays calendar managed	42	Absence-Overtime Module	N					

Figure 7. Example of Evaluation Questionnaire

5.5. Generating Effort Estimation Spreadsheets

Our final scenario is to perform some effort estimates and compare different possible options which can be specifying as variants in the RE Model. In the scope of this paper, we will consider that a specific effort field associated with each requirement is available. This can be filled in by some export. It can also be produced by an effort estimation method able to compute estimates from requirements. An example of such a method is the COSMIC Function Points (Dumke, Abran, 2010). The method requires all requirements to be enforcing using Create/Read/Update/Delete (CRUD) operations and is outside of the scope of this paper.

In order to produce the effort estimation, we will combine the following three spreadsheets using a tabular ODS output.

- EFFORT: a simple query to retrieve the requirement.effort
- ALLOCATION: the previously computed traceability matrix of Requirements vs Agents
- COMPUTATION: a specific spreadsheet performing the processing to compute module costs and applying styles on it

```
GENERATE [EFFORT, ALLOCATION, COMPUTATION] AS Tabular
```

The resulting spreadsheet is shown in Figure 8. The process was applied on two different variant of the system which should be specified as additional filters earlier in the process and that we will not detail here.

Req \ Module	Agent5	Agent3	Agent1	Req \ Module	Agent5	Agent3	Agent1
Requirement1	0	10	0	Requirement1	0	10	0
Requirement2	0	0	20	Requirement2	0	0	20
Requirement3	35	0	0	Requirement3	35	0	0
Requirement4	0	40	0	Requirement4	0	40	0
Requirement5	50	0	0	Requirement5	50	0	0
Requirement6	0	0	75	Requirement6	0	0	75
Requirement7	0	0	70	Requirement9	100	0	0
Requirement8	80	0	0				
				COST	185	50	95
COST	165	50	165	TOTAL COST	330		
TOTAL COST	380						

Figure 8. Effort Estimation for two Variants

6. Discussion and Related Work

We point discuss here some strong point emerging from our experiments as well as areas of improvement. We also bring some related work in the picture of this discussion to compare with other approaches and toolsets.

6.1. Strong points

- **Rich Model.** Using an underlying RE model provides rich traceability mechanisms and enables the automatic generation. This can be supported by most tools supporting a model-based approach like DOORS (IBM, 2002), Reqtify (Dassault Systems, 2010), Enterprise Architect (Sparx Systems, 2014) or Eclipse-based tools like TopCased (TopCased, 2011). Like in our case, the majority of tools provide some query mechanisms including predefined reports and finer grained queries (like OCL queries on Eclipse EMF models or SQL queries on the underlying database in Enterprise Architect).

- **Flexibility across artifacts.** The experience with the scenarios revealed a lot of flexibility resulting from the combination of interesting way to combine partial results, resulting in interesting chains of artifacts of different kinds. For example addressing the problem of reporting a concept page in a table revealed easy to support through minor evolution of the tool and relying on all the available transformation processes. In order to enable this, the tool should support a rich set of interconnection across different artifacts. With respect to this, most tools have limitations: DOORS and Reqtify are mainly targeting textual and table report generation but not copying with diagrams. Moreover Reqtify is only "readonly" and is not designed to reprocess its own outputs. Enterprise Architect fully supports the interplay of documents and diagrams. It also supports tabular views (e.g. on diagrams) and for import/export.

- **Ease of Combination** A key point to carry out our experimentations was the support for scenarios combination. In our case the main enabler was a visual report

functionality able to include all the manipulated artifacts and provide a specific behavior depending on the output format (e.g. including a query in a report will generate a formatted table in a document or a spreadsheet tab depending on the output format). Several tools only provide a report generator that are limited to the generation of standard template or require a lot of effort to customise their templates. As an alternative to a visual reporting tool, a good scripting API can also be efficient however it requires more effort to learn and to deploy.

6.2. Areas of Improvement

– **Connection with other models.** The richness of interaction scenarios also depends on the model. In our case it is rather limited to the goal-oriented RE. Other tools like cover larger domains such as design UML (Sparx Systems, 2014) or software product lines (University of Ottawa, 2001). The richer the model is, the greater the number of interesting interactions will be. Of course a tool must stay limited in scope and tool interconnectivity is also important (see later point).

– **Model Maintenance.** As it all model-based approaches corrections have be applied to the model and require regeneration. Tasks like locating and performing a correction in the model should be very easy to perform to avoid user rejection and will result in direct correction in documents that are not any more synchronised with the other artifacts. A simple recommended feature to reduce some corrections is to have a spell checker at the source of text capture.

– **Roundtrip Update.** More powerful roundtrip features could be investigated in order to be able to directly update from change in the documents. Most tools support the tagging of requirements description and attributes and support the synchronisation back to the model. However processing tables and diagrams from the target documents embedding them is more complex and often requires ad-hoc parsing. There is currently no easy solution for handling this at least considering traditional office suites.

– **Online tool.** The toolset we used as well as a number of major RE tools are still often based on a fat client architecture possibly connected to a database server. Nowadays toolsets are quickly evolving moving online in a SaaS mode either explicitly dedicated to requirements like (TraceCloud, 2014) or as evolution of other tools like bugtracker or wikis, e.g. JIRA and Confluence (Atlassian, 2014). Major office suite are already available online e.g. (Google, 2007), as well as diagramming toolsets like (Gliffy, 2009) (also available as Confluence plugin).

– **Interconnection protocol.** Tool integration is best achieved through an open interaction protocol like OSLC (Arwe, 2013). This protocol is now supported by several RE tools like DOORS (Jazz) or Enterprise Architect. However OSLC is based on the linked data principle and only provides the mechanisms to define resources in terms of RDF properties and query mechanisms. Achieving effective integration requires some work that could however be guided by our abstractions.

7. Conclusion and Perspectives

In the paper, we captured and documented a number of interesting interaction scenarios across several of artifacts types found in the requirements engineering process. This task relied on a careful review of a set of large requirements specification projects we carried out over the past ten years. We then looked at how well they were supported in a current state of the art RE tool. We prototyped a number of extensions to be able to cope with limitations and more importantly to experiment with combination and chaining of such scenarios in order to show interesting value added.

Our approach was proved positive, although some threats related to model-based engineering should be addressed. Interesting opportunities were also identified with the development of web-based solutions. We also tried to make our work reusable by others by describing our approach in abstract terms. We believe it is quite easily to transpose in a toolset providing a good extension API either closed (e.g. Enterprise Architect) or Open (e.g. Eclipse tools based on EMF like Papyrus, TopCased).

The developed extensions are part of the new V4 release of the Objectiver tool. Our future work is to support spreadsheets as input and to reference concepts within them. We are also studying more complex quantitative analysis of goal models. Finally we also plan to move to web-based tools and improve our tool interconnectivity using OSLC.

Acknowledgements

This work was funded by the Walloon Region INOGRAMS project (grant number 7171). We warmly thanks our industrial partners for sharing their cases.

References

- Abran A., Bourque P., Dupuis R., Moore J. W. (Eds.). (2001). *Guide to the Software Engineering Body of Knowledge - SWEBOK*. Piscataway, NJ, USA, IEEE Press.
- Arwe J. (2013, May). *Open Services for Lifecycle Collaboration, Core Specification Version 2.0*. <http://open-services.net/bin/view/Main/OslcCoreSpecification>.
- Atlantic Guild. (2014). *Volere Requirements Template, Edition 17*. <http://www.volere.co.uk/template.htm>.
- Atlassian. (2014). *JIRA and Confluence tools*. <https://www.atlassian.com/>.
- AT&T Research. (2014). *Graphviz - Graph Visualization Software*. <http://www.graphviz.org>.
- CETIC, Respect-IT. (2009, January). *Parliamentary Access Control eXchange System - Domain Description, Requirements Specification and Functional Specification*.
- Dassault Systems. (2010). *Reqtify*. <http://www.3ds.com/products-services/catia/capabilities/catia-systems-engineering/requirements-engineering/reqtify>.
- Devos N., Ponsard C., Deprez J.-C., Bauvin R., Moriau B., Anckaerts G. (2012). Efficient reuse of domain-specific test knowledge: An industrial case in the smart card domain. In *ICSE 2012*, p. 1123-1132. IEEE.

- Dumke R., Abran A. (2010). *COSMIC Function Points: Theory and Advanced Practices*. Auerback Publications, Taylor & Francis LLC.
- EMVCo. (2011, November). *EMV Integrated Circuit Card Specifications for Payment Systems - V4.3*. <http://www.emvco.com/specifications.aspx?id=223>.
- Gliffy. (2009). *Diagrams Made Easy*. <http://www.gliffy.com>.
- Google. (2007). *Google Docs*. <http://www.google.com/docs/about>.
- Herrmannsdorfer M., Konrad S., Berenbac B. (2008, March). Tabular Notations for State Machine-Based Specifications. *CrossTalk The Journal of Defense Software Eng.*, No. 8.
- Hummel B., Thyssen J. (2009). Behavioral Specification of Reactive Systems Using Stream-Based I/O Tables. In *Proc. of the 7th IEEE International Conference on Software Engineering and Formal Methods*, pp. 137–146. Washington, DC, USA, IEEE Computer Society.
- IBM. (2002). *Rational DOORS*. <http://www.ibm.com/software/awdtools/doors>.
- IEEE. (1998). *830-1998 - Recommended Practice for Software Requirements Specifications*.
- ISO/IEC/IEEE. (2011). *29148-2011 - Systems and software engineering - Life cycle processes - Requirements engineering*.
- Kelly T., Weaver R. (2004). The Goal Structuring Notation - A Safety Argument Notation. In *Proc. of dependable systems workshop on assurance cases*.
- Lamsweerde A. van. (2009). *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Wiley. Paperback.
- ODMG. (1998). *Object query language user manual v5.0*.
- OMG. (2011a). *Business Process Model and Notation - version 2.0*. <http://www.omg.org/spec/BPMN/2.0>.
- OMG. (2011b). *Unified Modelling Languages - 2.4*. <http://www.omg.org/spec/UML>.
- PaaSage Consortium. (2013). *D6.1.1 Initial Requirements*. <http://www.paasage.eu/images/documents/PaaSage-D6.1.1-Initial-Requirements.pdf>.
- Respect-IT. (2005). *Objectiver Requirements Engineering Tool*. <http://www.respect-it.com>.
- Sommerville I. (2010). *Software engineering* (9th ed.). Harlow, England, Addison-Wesley.
- Sparx Systems. (2014). *Enterprise Architect*. <http://www.sparxsystems.com.au>.
- SUN Microsystems. (2009, April). *OpenOffice.org 3.1 Developer's Guide*. https://wiki.openoffice.org/w/images/d/d9/DevelopersGuide_OOo3.1.0.pdf.
- TopCased. (2011). *Open Source Toolbox for Critical Systems*. <http://www.topcased.org/>.
- TraceCloud. (2014). *Requirements Management - Agile, Waterfall, Change Control*. <https://www.tracecloud.com>.
- University of Ottawa. (2001). *jUCMNav: Juice up your modelling*. <http://jucmnav.softwareengineering.ca/twiki/bin/view/ProjetSEG/WebHome>.
- Yu E. S. K., Mylopoulos J. (1997, April). Enterprise modelling for business redesign: The i* framework. *SIGGROUP Bull.*, Vol. 18, No. 1, pp. 59–63.

Vérification automatique d'exigences pour les politiques d'échange d'information

Exigences de Diffusion et de Non-diffusion d'information

Rémi Delmas, Thomas Polacsek

ONERA, Département Traitement de l'Information et Modélisation
2, avenue Edouard Belin BP74025, 31055 TOULOUSE Cedex 4
{remi.delmas,thomas.polacsek}@onera.fr

RÉSUMÉ. Que cela soit pour la surveillance de la Terre ou même la gestion des relations inter-entreprises, il existe de plus de plus d'organisations interconnectées formant des systèmes d'information décentralisés avec des échanges d'information. Dans de tels cas, nous avons l'exigence que si une information cruciale vient à être connue par l'un des agents, alors les agents concernés par cette information doivent absolument être avertis. Parallèlement, il peut être nécessaire de prévenir tout risque de diffusion d'information n'ayant pas vocation à être divulguée. Dans cet article, nous nous focaliserons plus précisément sur deux types d'exigences antagonistes qui sont, d'une part, la nécessité de partager de l'information et, d'autre part, l'obligation de ne pas diffuser certaines informations. Fort d'une expression formelle de ces deux exigences, nous verrons comment il est possible de les concilier dans une unique politique à l'aide d'une opération de filtrage d'information dont la spécification apparaît naturellement. De plus, nous expliquerons comment des solveurs SMT peuvent être utilisés pour analyser automatiquement des propriétés des politiques.

ABSTRACT. Whether be it for Earth observation, risk management or even companies relations, more and more interconnected organizations form decentralized systems in which the exchange, in terms of diffusion or non-diffusion of information between agents, can have critical dimension. In this paper, we present a formal framework to specify information exchange policies for such kinds of systems and two specific requirements, the need-to-share and the non-diffusion requirements, as well as properties strongly related to them. Wisser from these formal definitions, we see how to reconcile these sometimes two antagonist requirements in a same policy specification with information filtering operations. We also explain how we use state of the art theorem provers to perform automatic analysis of these policies.

MOTS-CLÉS : analyses de politique d'échange, exigences des SI, méthodes formelles, vérification, solveurs SMT.

KEYWORDS: exchange policy analysis, IS requirements, formal methods, verification, SMT solvers.

1. Introduction

Aujourd'hui, de par l'omniprésence des relations numériques, nous avons une interconnexion sans précédent entre les personnes, les entreprises et les organisations. Ces ensembles d'acteurs interconnectés forment des systèmes décentralisés où de l'information est échangée. Prenons pour exemple les systèmes de surveillance des débris spatiaux, Space Situation Awareness (SSA). Dans de tels systèmes, les capacités d'observation spatiales, appartenant à différentes nations et opérateurs privés ou publics, sont mutualisées afin de construire une information complexe et, s'il y a un risque de collision pour un satellite, diffuser une alerte aux agents concernés. La mission d'un tel système est donc de prévenir les opérateurs de satellites lorsqu'une collision potentielle entre des objets en orbite est détectée. Plus précisément, le système doit, en cas de risque de collision, envoyer des informations aux bons agents afin de leur permettre d'éviter la collision, tout en garantissant qu'aucune information sensible sur les objets en orbite, comme leur nature exacte, leurs trajectoires, leurs capacités de manœuvre, etc. ne soit divulguée.

Un autre cas de systèmes décentralisés où des agents s'échangent de l'information est celui des systèmes de surveillance faisant partie du système d'observation globale de la Terre (GEOSS¹). Concernant la prévention des catastrophes naturelles, nous avons un ensemble de partenaires qui partagent des informations dans le but de générer des alertes en cas de risque de catastrophe naturelle. L'exigence ici est que les autorités qualifiées *doivent absolument être averties* dès que des données annonçant une catastrophe sont connues afin que des mesures de protection et de gestion de crise soient prises. Mais cette exigence est contrebalancée, par exemple, par la nécessité de prévenir tout risque de diffusion d'informations confidentielles sur les moyens d'observation de la terre des membres participant à l'effort de surveillance.

Nous avons donc des systèmes dans lesquels l'exigence principale est que, dans certains cas, un agent *doit être absolument averti*. En fait la véritable exigence est un *besoin de connaître* des agents, qui doivent connaître certaines informations pour pouvoir accomplir leurs missions. Cependant, comme nous nous intéressons aux échanges d'information, nous pouvons reformuler cette exigence en exigence de *diffusion* qui signifie que puisqu'un agent a besoin d'une information, quand celle-ci est connue par un autre agent du système, elle doit absolument lui être communiquée. Parallèlement, parce que différentes organisations interviennent dans ces systèmes, il faut prévenir la diffusion d'informations privées ou sensibles entre organisations. Les exigences à concilier dans ces systèmes sont donc antagonistes : d'une part, veiller à ce que des acteurs reçoivent toujours toutes les informations dont ils ont besoin pour accomplir leurs missions respectives ; d'autre part, garantir qu'aucune information sensible ne sera diffusée de manière incontrôlée, ce que nous nommons exigence de *non-diffusion*.

Pour cela, il est nécessaire de spécifier clairement et sans ambiguïté les exigences de diffusion et de non-diffusion, c'est-à-dire les conditions sous lesquelles un

1. Global Earth Observation System of Systems

agent a l'obligation, l'autorisation ou l'interdiction de communiquer des informations aux autres agents du système. Nous appelons une telle spécification une *politique d'échange*. Un des avantages de disposer d'une politique d'échange exprimée formellement est de pouvoir réaliser des vérifications automatiques, via des outils d'aide à la spécification/vérification, utilisables dès les premières étapes du processus de conception.

Dans cet article, la section 2 présente un rappel du langage PEPS² précédemment défini dans (Cholvy *et al.*, 2012) (Delmas et Polacsek, 2013) (Delmas et Polacsek, 2014) et qui permet de spécifier formellement une politique d'échange d'information, ainsi que l'outil PEPS-analyzer associé au langage. Dans la section 3, nous donnons un exemple de politique d'échange qui nous permettra d'illustrer l'ensemble des concepts étudiés dans ce papier. Les sections 4 et 5 sont dédiées respectivement aux exigences de diffusion et de non-diffusion. Malgré l'antagonisme de ces deux types d'exigences, nous verrons dans la section 6 comment un opérateur de filtrage d'information permet de les concilier dans une même politique sans incohérence. Enfin, la section 7 conclut le papier et donne quelques perspectives à ces travaux.

2. Spécification de politiques d'échange

2.1. PEPS

PEPS est un cadre formel outillé pour la spécification et la vérification de politiques de diffusion d'information. Techniquement PEPS est bâti sur la logique du premier ordre multi-sortée³ avec égalité (MSFOL) (Gallier, 1987). Il permet l'utilisation de sortes, de constantes, de fonctions, de prédicats, de variables sortées $x : \mathcal{S}$, de quantificateurs universels (\forall) et existentiels (\exists), et enfin de l'égalité ($=$) et des connecteurs logiques usuels ($\neg, \wedge, \vee, \implies$).

Le langage PEPS est extensible, l'utilisateur peut déclarer ses propres sortes, fonctions et prédicats. Toutefois, il est nativement pourvu de sortes, fonctions et prédicats prédéfinis couvrant les concepts de base des politiques. Ainsi, les sortes \mathcal{A} , \mathcal{I} et \mathcal{T} représentent respectivement les agents, les informations et les sujets sur lesquels les informations portent ; le prédicat $K(a, i)$ signifie que l'agent a connaît l'information i et le prédicat $Topic(i, t)$ signifie que l'information i se rapporte au sujet t (appelés *D-prédicats* pour prédicats de domaine).

Contrairement à la logique déontique standard, notre langage ne dispose pas d'un opérateur d'obligation générique (Castanêda, 1975). En effet, nous nous concentrons uniquement sur la notion d'*obligation d'envoyer une information d'un agent vers un autre* et pas sur la notion d'obligation en général. Par conséquent, nous avons trois prédicats dédiés, appelés *prédicats normatifs (N-prédicats)*, par opposition aux *D-prédicats*) : $O_{Send}(a, b, i)$, $P_{Send}(a, b, i)$ et $F_{Send}(a, b, i)$, qui modélisent respec-

2. PEPS est l'acronyme récursivement défini par : PEPS for exchange policy specification

3. on peut voir une *sorte* comme un *type*

tivement l’obligation, l’autorisation et l’interdiction pour un agent a d’envoyer à un agent b une information i .

Notons que pour exprimer la notion d’obligation et les concepts associés inhérents à une politique, nous aurions pu choisir d’utiliser une logique déontique. Cependant, les outils de preuve pour les logiques modales sont beaucoup moins efficaces que les outils de preuve pour les logiques standard comme les solveurs SAT ou SMT (Sebastiani et Vescovi, 2009). Par conséquent, en n’utilisant pas un opérateur modal, nous perdons en expressivité, mais, dans le cadre d’analyses automatiques, nous gagnons en efficacité.

Dans la logique déontique standard, les modalités d’obligation et de permission sont reliées par l’axiome (D) qui exprime que si p est obligatoire alors p est également permis. Dans PEPS, nous traduisons cet axiome par une propriété de logique, que nous appelons également (D) et qui exprime le fait que s’il est obligatoire pour un agent d’envoyer une information à un autre agent alors il est également permis de l’envoyer.

Définition 1

$$(D) \quad \forall a, \forall b, \forall i, O_{Send}(a, b, i) \implies P_{Send}(a, b, i)$$

De plus, nous définissons les *règles d’échange* comme des formules qui spécifient sous quelles conditions un agent à l’obligation, l’interdiction ou la permission d’envoyer une information à un autre agent. Nous appelons une *politique d’échange (EP)* un ensemble de règles d’échange.

Définition 2 (Règle d’échange)

Une règle d’échange est une formule PEPS fermée de l’une des formes suivantes :

$$\forall x_1, \dots, \forall x_n, (\phi \implies O_{Send}(t_1, t_2, t_3))$$

$$\forall x_1, \dots, \forall x_n, (\phi \implies P_{Send}(t_1, t_2, t_3))$$

$$\forall x_1, \dots, \forall x_n, (\phi \implies F_{Send}(t_1, t_2, t_3))$$

où :

- x_1, \dots, x_n sont toutes les variables présentes dans ϕ , t_1 , t_2 et t_3 ;
- ϕ est une formule sans quantificateur et sans N-prédicats ;
- t_1, t_2 sont des termes de sorte \mathcal{A} ;
- t_3 est un terme de sorte \mathcal{I} .

À cela nous devons ajouter une description formelle du domaine, notée Σ . La déclaration des sortes et des prédicats nécessaires pour décrire le domaine associé à une application particulière est laissée à l’utilisateur. Techniquement, PEPS est extensible uniquement avec de nouveaux types, des fonctions et des D-prédicats, mais il n’est pas possible d’introduire de nouveaux N-prédicats.

Dans la suite, nous appellerons *spécification d'une politique d'échange*, dénotée par EPS , le couple formé par la politique d'échange proprement dite EP et par les contraintes du domaine Σ , auxquelles nous ajoutons la propriété (D).

Définition 3 (Spécification d'une politique d'échange)

$$EPS \equiv \Sigma \wedge \left(\bigwedge_{r \in EP} r \right) \wedge D$$

Enfin, nous utiliserons la notation $P \models Q$ pour dire que Q est une conséquence logique de P , *i.e.* que tout modèle de P est aussi un modèle de Q .

2.2. Vérification automatique de politiques d'échange

Dans cette section nous allons brièvement présenter les mécanismes sous-jacents à notre outil PEPS-analyzer, puis nous présenterons l'ensemble des propriétés génériques, propres aux politiques d'échange, pouvant être automatiquement vérifiées à l'aide de cet outil.

2.2.1. PEPS-analyzer

PEPS-analyzer est un outil d'aide à la spécification de politiques d'échange. Il permet de vérifier des propriétés sur les politiques et, si ces propriétés ne sont pas vérifiées, renvoie des contre-exemples. Ainsi, par interaction avec l'outil, il est possible, à l'aide des contre-exemples renvoyés, de mettre au point incrémentalement une politique qui répond réellement aux exigences souhaitées. Cette tâche devient très difficile voire impossible pour un humain à mesure que la spécification grandit : sur des exemples simples d'une dizaine de règles seulement, la combinatoire des interactions les différentes règles fait qu'il est impossible d'identifier humainement toutes les incohérences entre règles, de garantir la complétude de la politique ou de garantir l'absence de redondance entre règles.

Concrètement, PEPS-analyzer ramène le problème de vérification d'une propriété à un problème de satisfiabilité. Ainsi, vérifier que $P \models Q$ (Q est conséquence logique de P), avec P et Q des formules de la MSFOL, revient à vérifier à l'aide d'un solveur SMT (Satisfiability Modulo Theories) que la formule $P \wedge \neg Q$ est insatisfiable.

Si la première version de PEPS-analyzer implémentait un encodage purement booléen de la MSFOL dans un univers borné et utilisait un solveur SAT (Delmas et Polacsek, 2013), la nouvelle version de l'outil délègue le raisonnement dans la MSFOL à un solveur externe de type SMT, qui gère les quantificateurs nativement.

PEPS-analyzer est développé en Scala et il s'appuie sur le solveur SMT Z3 (de Moura et Bjørner, 2008).

2.2.2. Des propriétés génériques

L'outil PEPS-analyser permet de vérifier automatiquement quatre propriétés génériques sur les politiques : la *consistance*, l'*applicabilité*, la *minimalité* et la *complétude*.

Une politique est *consistante* s'il n'existe pas de cas où il est à la fois interdit et obligatoire (ou permis) pour un agent d'envoyer une information à un autre agent. L'*applicabilité* correspond au fait que la politique ne régit pas des situations qui n'existent pas ou n'arrivent jamais, et une politique est *minimale* s'il n'existe pas de règle qui puisse se déduire des autres (et est donc inutile).

Nous pouvons définir la propriété de complétude comme suit : “dans toutes situations et pour toutes informations, la politique indique si un agent qui connaît cette information à l'obligation, la permission ou l'interdiction de l'envoyer aux autres agents.” Cette définition est assez standard et correspond à celle donnée par (Akl et Denning, 1987) (Denning *et al.*, 1987) dans le contexte des politiques de contrôle d'accès. Cependant, dans les phases préliminaires de conception, il n'est pas strictement nécessaire qu'une politique soit complète. Il peut être utile de n'avoir une politique complète que pour certains sous-domaines de son domaine d'application. En effet, une politique peut être conçue dans le cadre d'une collaboration entre des parties bien distinctes, chacune portant seulement attention au sous-ensemble des sujets qui la concernent. Par exemple, dans le cadre de l'observation de la Terre, les opérateurs militaires peuvent vouloir s'assurer que la politique est complète pour tout ce qui concerne le sujet militaire sans s'intéresser aux autres sujets.

Nous proposons donc une définition paramétrée de la complétude, que nous appelons *T-complétude*. Une politique d'échange est dite *T-complète* si et seulement si pour tout agent qui connaît une information pertinente sur un sujet *T*, la politique spécifie si l'agent a l'obligation, la permission ou interdiction de l'envoyer à un autre agent.

Définition 4 (*T-complétude*)

Soit $EPS = \langle \Sigma, EP \rangle$ la spécification d'une politique d'échange et T une constante de sorte \mathcal{T} . EPS est *T-complète* si et seulement si :

$$EPS \models \forall a, \forall b, \forall i, \\ (K(a, i) \wedge Topic(i, T) \implies (P_{Send}(a, b, i) \vee O_{Send}(a, b, i) \vee F_{Send}(a, b, i)))$$

3. Exemple

Nous allons ici introduire un exemple qui nous permettra d'illustrer l'utilisation de PEPS ainsi que les différents points développés dans les sections suivantes. Considérons des agents qui disposent de moyens d'observation de la Terre et qui décident de

participer à une coalition pour la gestion des risques sismiques. Dans le cadre de cette coalition, il existe un groupe d'agents particulier : le *Groupe de Gestion des Risques Sismiques*, noté *GRS*, dont la mission est de prévenir les fausses alertes, d'organiser les évacuations et de gérer la communication avec le public. La politique de ce système se compose pour le moment de quatre règles :

- r1* “*Tout agent non membre du GRS doit communiquer toute information relative aux risques sismiques à au moins un membre du GRS.*”
- r1b* “*Tout agent non membre du GRS a la permission de communiquer des informations relatives aux risques sismiques à n'importe quel membre du GRS.*”
- r2* “*Il est interdit, pour tout agent non membre du GRS, de communiquer des informations relatives aux risques sismiques à toute personne non membre du GRS.*”
- r3* “*Les membres du GRS ont la permission de communiquer les informations relatives aux risques sismiques à quiconque.*”

La règle *r1* découle clairement de l'exigence de diffusion : il y a une nécessité à ce que les agents communiquent toute information relative à une catastrophe sismique à un membre du *GRS* pour permettre au *GRS* d'accomplir sa mission. La règle *r1b* complète la règle *r1* : tout agent externe au *GRS* sait ainsi s'il a le droit de communiquer une information relative aux risques sismiques aux autres agents du *GRS* en plus de celui avec lequel il a l'obligation de communiquer. La règle *r2* a pour but de prévenir tout risque de panique par la diffusion brutale d'informations au grand public.

Notons que la règle *r3* relève d'un processus d'appréciation qui n'est pas modélisé, et qui n'a pas forcément vocation à l'être au sein de la politique d'échange : le choix réalisé par un membre du *GRS* d'émettre ou pas un avertissement au public. Ici, c'est bien le prédicat de permission d'envoyer une information qui permet de modéliser la politique à ce niveau d'abstraction.

Pour pouvoir modéliser cette politique en PEPS nous devons d'abord déclarer une nouvelle constante *geo*, de sorte \mathcal{T} , qui représente le sujet *risque sismique*. Nous déclarons aussi un nouveau prédicat de domaine *GRS*, qui s'applique sur la sorte \mathcal{A} , et qui modélise l'appartenance au groupe *GRS*. Ainsi, $GRS(a)$ est vrai lorsque l'agent *a* est élément du groupe *GRS*, et faux lorsqu'il ne l'est pas⁴. La politique de notre exemple se modélise donc en PEPS comme suit :

4. Nous avons choisi de modéliser l'appartenance à un groupe de la façon la plus simple possible. Ici, chaque groupe est caractérisé par un prédicat portant sur des agents et indiquant l'appartenance ou pas de l'agent au groupe. Nous aurions tout aussi bien pu introduire une sorte pour les groupes et un prédicat d'appartenance à un groupe de la forme $member(g, a)$.

$$\begin{array}{ll}
r1 : \forall a, \forall i, \exists b, K(a, i) \wedge Topic(i, geo) \wedge \neg GRS(a) \wedge GRS(b) & \implies O_{Send}(a, b, i) \\
r1b : \forall a, \forall i, \forall b, K(a, i) \wedge Topic(i, geo) \wedge \neg GRS(a) \wedge GRS(b) & \implies P_{Send}(a, b, i) \\
r2 : \forall a, \forall i, \forall b, K(a, i) \wedge Topic(i, geo) \wedge \neg GRS(a) \wedge \neg GRS(b) & \implies F_{Send}(a, b, i) \\
r3 : \forall a, \forall i, \forall b, K(a, i) \wedge Topic(i, geo) \wedge GRS(a) & \implies P_{Send}(a, b, i)
\end{array}$$

Nous rajoutons à cela la contrainte de domaine “*Toute information concerne au moins un sujet*” : $d : \forall i, \exists t, Topic(i, t)$.

A l’aide de notre outil PEPS-analyzer, nous pouvons automatiquement vérifier que la spécification $\langle \{d\}, \{r1, r1b, r2, r3\} \rangle$ est *geo-complète*, consistante, applicable et minimale.

4. Exigence diffusion : la propriété de vigilance

Dans les systèmes tels que GEOSS ou le SSA, certains acteurs ont un rôle particulier : pour accomplir leur mission ils doivent absolument connaître toute information relative à un sujet donné. Dans notre exemple, les agents du groupe *GRS* doivent être prévenus de tout ce qui se rapporte au risque sismique. Par conséquent, nous définissons la propriété dite de *T-vigilance* pour un groupe comme le fait qu’un groupe d’agents est toujours correctement informé relativement à un sujet précis *T*.

Définition 5 (*T-vigilance pour un groupe G*)

Soit $EPS = \langle \Sigma, EP \rangle$ la spécification d’une politique d’échange, T une constante de sorte \mathcal{T} et G un prédicat sur la sorte \mathcal{A} caractérisant un groupe d’agents. Alors G est *T-vigilant* dans EPS si et seulement si :

$$EPS \models (\forall a, \forall i, \exists b, K(a, i) \wedge Topic(i, T) \wedge \neg G(a) \wedge G(b) \implies O_{Send}(a, b, i))$$

Remarquons que dans le cas où le groupe se résume à un unique agent et où cet agent est *T-vigilant* sur tous les sujets possibles, alors la propriété se résume au fait que tout autre agent du système doit envoyer tout ce qu’il connaît à cet agent, en d’autres termes, nous sommes face à un agent que nous pouvons qualifier d’*omniscient*.

À l’aide de PEPS-analyzer nous pouvons vérifier que la politique d’exemple $\langle \{d\}, \{r1, R1b, r2, idr3\} \rangle$ vérifie la propriété de *geo-vigilance* pour le groupe *GRS*. Notons qu’ici la règle *r1* est l’instanciation directe de la propriété de vigilance pour le thème *geo*.

5. Exigence de non-diffusion : les propriétés de restriction

Si la propriété *T-vigilance* permet de vérifier qu’un groupe d’agents donné reçoit toujours toute information relevant du sujet *T*, il peut être intéressant, de façon duale,

de vérifier que toute information traitant d'un sujet particulier (comme, par exemple, le sujet *confidentiel*) ne doit pas être envoyée à un groupe d'agents, à un agent unique ou bien tout simplement ne doit pas être envoyée du tout.

Nous définissons la propriété de *restriction de la diffusion d'information concernant un sujet à un groupe donné* selon trois cas : soit il est interdit aux agents hors du groupe de s'échanger une information de ce sujet ; soit il est interdit pour un agent hors du groupe de communiquer une information de ce sujet à un membre du groupe ; soit il est interdit aux membres du groupe de communiquer une information de ce sujet hors du groupe.

Définition 6 (*T*-restriction à un groupe *G*)

Soit $EPS = \langle \Sigma, EP \rangle$ la spécification d'une politique d'échange, T une constante de sorte \mathcal{T} et G un prédicat sur la sorte \mathcal{A} caractérisant un groupe d'agents, nous avons alors :

(a) *T-out-out-restriction pour G dans EPS si et seulement si :*

$$\mathcal{EPS} \models (\forall a, \forall b, \forall i, K(a, i) \wedge Topic(i, T) \wedge \neg G(a) \wedge \neg G(b) \implies F_{Send}(a, b, i))$$

(b) *T-out-in-restriction pour G dans EPS si et seulement si :*

$$\mathcal{EPS} \models (\forall a, \forall b, \forall i, K(a, i) \wedge Topic(i, T) \wedge \neg G(a) \wedge G(b) \implies F_{Send}(a, b, i))$$

(c) *T-in-out-restriction pour G dans EPS si et seulement si :*

$$\mathcal{EPS} \models (\forall a, \forall b, \forall i, K(a, i) \wedge Topic(i, T) \wedge G(a) \wedge \neg G(b) \implies F_{Send}(a, b, i))$$

Pour finir, nous ajoutons aux propriétés de restriction ci-dessus la propriété de *T-restriction stricte* qui signifie que tout échange d'information relative à T est interdit entre agents, quels qu'ils soient. Notons que cette propriété n'est qu'un cas particulier de la définition 6 avec G le groupe vide (modélisé par $\forall a, G(a) \equiv \perp$).

Définition 7 (*T*-restriction stricte)

Soit $EPS = \langle \Sigma, EP \rangle$ la spécification d'une politique d'échange et T une constante de sorte \mathcal{T} , nous avons la *T-restriction stricte* dans EPS si et seulement si :

$$\mathcal{EPS} \models (\forall a, \forall b, \forall i, K(a, i) \wedge Topic(i, T) \implies F_{Send}(a, b, i))$$

Sur notre exemple de surveillance de risques sismiques, nous pouvons vérifier, à l'aide de PEPS-analyzer, que la politique $\langle \{d\}, \{r1, r1b, r2, r3\} \rangle$ satisfait la *geo-out-restriction* pour le groupe GRS . En effet, aucune information relative à *geo* ne

peut être envoyée entre agents hors du groupe *GRS*, par contre, les agents du *GRS* peuvent recevoir ces informations des agents extérieurs aux groupes et ils ont aussi la permission de communiquer ces informations avec les agents hors du groupe.

6. Diffusion et non-diffusion

6.1. Incompatibilité entre la vigilance et la restriction

Enrichissons notre exemple en supposant maintenant qu’il existe dans le système des informations que nous qualifierons de *sensibles*. Pour prévenir la diffusion de telles informations nous rajoutons dans la politique de notre exemple la règle r_4 suivante : “*il est interdit de s’échanger toute information relative au sujet sensible*”. Afin de pouvoir modéliser cette nouvelle règle dans PEPS, nous introduisons une nouvelle constante de sorte \mathcal{T} : *sens*, qui représente le sujet *sensible*.

$$r_4 : \forall a, \forall b, \forall i, K(a, i) \wedge Topic(i, sens) \implies F_{Send}(a, b, i)$$

À l’aide de PEPS-analyzer nous découvrons que, malheureusement, l’ajout de r_4 rend la politique $\{\{d\}, \{r_1, r_{1b}, r_2, r_3, r_4\}\}$ inconsistante.

En effet, considérons le cas où un agent connaît une information relative à la fois à *geo* et *sens*, comme une image prise par un satellite montrant un risque de catastrophe naturelle sur un site civil joutant un site militaire sensible. Dans ce cas là, la règle r_1 oblige l’agent à envoyer cette information à au moins un agent du *GRS*, alors que r_4 interdit à ce même agent d’envoyer l’information à quiconque et donc, à fortiori, à un agent du *GRS*, ce qui viole la propriété de consistance. Remarquons que les règles r_3 et r_4 entraînent également une inconsistance en permettant et en interdisant simultanément la diffusion d’information portant sur *geo* et *sens*.

Le problème d’inconsistance soulevé ici n’est pas spécifiquement lié à l’exemple. Plus généralement, si l’on considère une politique munie d’une propriété de T_1 -vigilance pour un groupe G_1 et d’une propriété de T_2 -restriction pour un groupe G_2 alors, suivant les règles du domaine, il peut être possible de construire des modèles satisfaisant les deux propriétés et où il est à la fois obligatoire et interdit d’envoyer une information. Ces modèles ont tous la même structure : au moins une information concerne les deux sujets, T_1 et T_2 , et les prédicats de groupe G_1 et G_2 sont tels qu’il existe des agents à l’intérieur et à l’extérieur de G_1 et de G_2 qui satisfont les prémisses des propriétés de T_1 -vigilance et T_2 -restriction.

6.2. Une solution ad-hoc

Afin de prévenir les possibles conflits entre les exigences de diffusion et de non-diffusion, nous proposons d’introduire un nouvel opérateur, pour le moment sans si-

gnification particulière, de signature $p(i : \mathcal{I}) : \mathcal{I}$, qui prend en entrée une information et retourne une information.

Dans notre exemple, nous voulons que cet opérateur permette d'*oublier* la partie sensible d'une information. Par conséquent nous posons les deux contraintes de domaine suivantes pour représenter le comportement de notre opérateur abstrait : ($p1$) une information produite par p n'est plus pertinente pour le sujet *sens* et ($p2$) une information relative au sujet *geo* le reste après application de p .

$$\begin{aligned} p1 : \quad & \forall i, \neg \text{Topic}(p(i), \text{sens}) \\ p2 : \quad & \forall i, \text{Topic}(i, \text{geo}) \implies \text{Topic}(p(i), \text{geo}) \end{aligned}$$

Nous devons maintenant adapter la politique d'échange de notre exemple pour préciser dans quels cas l'opérateur abstrait p doit être utilisé.

Premièrement, nous séparons la règle $r1$ en deux nouvelles règles, $r11$ et $r12$, pour exprimer le fait que ($r11$) si une information est liée à des risques sismiques et qu'elle n'a pas de caractère sensible, alors il est obligatoire de l'envoyer à un membre du *GRS* au moins, mais ($r12$) si cette information est également sensible alors il faut envoyer non pas i elle même, mais l'information résultant de l'application de p .

$$\begin{aligned} r11 : \quad & \forall a, \forall i, \exists b, K(a, i) \wedge \text{Topic}(i, \text{geo}) \wedge \neg \text{Topic}(i, \text{sens}) \wedge \neg \text{GRS}(a) \wedge \text{GRS}(b) \\ & \implies O_{\text{Send}}(a, b, i) \\ r12 : \quad & \forall a, \forall i, \exists b, K(a, i) \wedge \text{Topic}(i, \text{geo}) \wedge \text{Topic}(i, \text{sens}) \wedge \neg \text{GRS}(a) \wedge \text{GRS}(b) \\ & \implies O_{\text{Send}}(a, b, p(i)) \end{aligned}$$

Deuxièmement, sur le même modèle que $r1$ et pour les mêmes raisons, nous décomposons la règle $r1b$ en deux nouvelles règles $r1b1$ et $r1b2$:

$$\begin{aligned} r1b1 : \quad & \forall a, \forall i, \forall b, K(a, i) \wedge \text{Topic}(i, \text{geo}) \wedge \text{Topic}(i, \text{sens}) \wedge \neg \text{GRS}(a) \wedge \text{GRS}(b) \\ & \implies P_{\text{Send}}(a, b, p(i)) \\ r1b2 : \quad & \forall a, \forall i, \forall b, K(a, i) \wedge \text{Topic}(i, \text{geo}) \wedge \neg \text{Topic}(i, \text{sens}) \wedge \neg \text{GRS}(a) \wedge \text{GRS}(b) \\ & \implies P_{\text{Send}}(a, b, i) \end{aligned}$$

Troisièmement, nous modifions la règle $r3$ en $r3'$ pour exprimer le fait que tout membre du *GRS* est autorisé à communiquer une information liée à des risques sismiques à tout autre agent, à la condition que cette information ne soit pas sensible.

$$\begin{aligned} r3' : \quad & \forall a, \forall b, \forall i, K(a, i) \wedge \text{Topic}(i, \text{geo}) \wedge \neg \text{Topic}(i, \text{sens}) \wedge \text{GRS}(a) \\ & \implies P_{\text{Send}}(a, b, i) \end{aligned}$$

Nous pouvons à présent vérifier à l'aide de PEPS-analyzer que cette nouvelle politique, $\langle \{d, p1, p2\}, \{r11, r12, r1b1, r1b2, r2, r3', r4\} \rangle$, est *geo-complète*, *sens-*

complète, consistante, applicable, minimale et satisfait la restriction stricte pour le sujet *sens*.

Cependant, cette nouvelle politique ne satisfait pas la propriété de *geo*-vigilance pour le *GRS*. En effet, quand une information concernant un risque sismique est aussi de nature sensible, c'est le résultat de l'application de la fonction p à cette information qui est envoyée au *GRS* alors que la propriété de vigilance définie initialement impose que cela soit l'information elle-même qui soit envoyée. Nous avons introduit la notion de vigilance afin de garantir que le *GRS* puisse réaliser sa mission, ce qui correspondait au fait que toute information relative à *geo* devait être envoyée au *GRS*. Finalement, l'important est qu'un membre du *GRS* soit prévenu et peu importe qu'il reçoive l'information initiale ou l'information après l'application de p , du moment que la partie concernant le sujet *geo* est préservée. Par conséquent, plutôt que de chercher à établir la *geo*-vigilance pour le groupe *GRS*, nous devons vérifier que la politique garantisse que si une information porte sur *geo*, alors doit être envoyée à un membre du *GRS* : soit cette information, soit le résultat de l'application de p , ce qui est résumé par la formule suivante :

$$\begin{aligned} \forall a, \forall i, K(a, i) \wedge Topic(i, geo) \wedge \neg GRS(a) \implies \\ (\exists b, GRS(b) \wedge (O_{Send}(a, b, i) \vee (Topic(p(i), geo) \wedge O_{Send}(a, b, p(i))))) \end{aligned}$$

L'opérateur abstrait p agit comme un filtrage sur les sujets sur lesquels porte une information. Il correspond à une catégorie d'opérations effectuées régulièrement par les organismes qui gèrent des données sensibles, les opérations dites de *déclassification*, de *masquage*, d'*anonymisation*, etc.

6.3. Un opérateur générique de filtrage d'information

Afin de modéliser des opérations telles que la déclassification et ses variantes dans PEPS, nous introduisons un opérateur générique dit de *filtrage*, paramétré par des *modes de filtrage*. Chaque mode spécifie les sujets qui sont *conservés* ou *retirés* par l'opérateur.

Pour cela, nous introduisons : une nouvelle sorte \mathcal{M} , dont les éléments représentent les différents modes de filtrage ; un opérateur de filtrage $filter(m : \mathcal{M}, i : \mathcal{I}) : \mathcal{I}$ ainsi que deux prédicats $preserves(m : \mathcal{M}, t : \mathcal{T})$ et $removes(m : \mathcal{M}, t : \mathcal{T})$. Les axiomes suivants expriment le comportement de l'opérateur de filtrage en fonction du mode :

Définition 8 (*F* axiomes)

$$\begin{aligned} \forall t, \forall m, \quad preserves(m, t) &\implies \neg removes(m, t) \\ \forall i, \forall t, \forall m, \quad Topic(i, t) \wedge preserves(m, t) &\implies Topic(filter(m, i), t) \\ \forall i, \forall t, \forall m, \quad Topic(i, t) \wedge removes(m, t) &\implies \neg Topic(filter(m, i), t) \end{aligned}$$

Le premier axiome garantit la cohérence entre les prédicats de préservation et de suppression : si un mode conserve un sujet, alors il ne le supprime pas. Le deuxième axiome stipule que si une information concerne un sujet qui est préservé par un mode, alors elle concerne toujours ce sujet après filtrage sous ce mode. Le troisième axiome est simplement le dual du deuxième concernant le prédicat de suppression.

L'opérateur ad-hoc p , défini dans la section précédente, est subsumé par l'opérateur filtrage générique. Dans notre exemple, nous introduisons une constante de sorte \mathcal{M} *filterSens*, qui représente le mode de filtrage pour les informations à contenus sensibles et qui préserve les contenus relatifs aux risques sismiques. Concrètement, nous ajoutons la contrainte de domaine suivante qui spécifie les propriétés du filtrage pour ce mode :

$$f : \text{preserves}(\text{filterSens}, \text{geo}) \wedge \text{removes}(\text{filterSens}, \text{sens})$$

Il est important de noter que cette caractérisation des modes de filtrage à l'aide des prédicats *preserves* et *removes* n'est qu'une modélisation partielle des politiques de filtrage du monde réel. En effet, dans la pratique, d'autres conditions doivent être prises en compte pour l'utilisation de l'opérateur, telles que la capacité de l'agent à exécuter effectivement l'opération de filtrage, les caractéristiques du destinataire, *etc.* Toutes les conditions supplémentaires qui caractérisent les modes de filtrage pourraient être regroupées pour former une *politique de filtrage*, comparable à ce qui existe déjà pour les *politiques de déclassification*. Cependant, même si l'extensibilité de PEPS et la puissance expressive de la logique sous-jacente permettent de modéliser ces concepts, nous ne les développerons pas davantage dans cet article.

6.4. Propriétés génériques de vigilance et de restriction

Forts de la définition de l'opérateur de filtrage, nous pouvons redéfinir la propriété de vigilance en une version. Un groupe d'agents est dit T -vigilant si et seulement si tout agent, extérieur au groupe, connaissant une information relative au sujet T , a l'obligation d'envoyer à au moins un agent appartenant au groupe soit l'information soit l'information filtrée avec un mode de filtrage qui préserve les informations concernant le sujet T .

Définition 9 (T -vigilance pour un groupe G)

Soit $EPS = \langle \Sigma, EP \rangle$ la spécification d'une politique d'échange, T une constante de sorte \mathcal{T} et G un prédicat sur la sorte \mathcal{A} caractérisant un groupe d'agents, G est T -vigilant dans EPS si et seulement si :

$$\begin{aligned} \mathcal{E}PS, F \models (\forall a, \forall i, K(a, i) \wedge \text{Topic}(i, T) \wedge G(a) \implies \\ (\exists b, G(b) \wedge (O_{\text{Send}}(a, b, i) \vee \exists m, \text{preserves}(m, T) \wedge O_{\text{Send}}(a, b, \text{filter}(m, i)))))) \end{aligned}$$

Si nous revenons à notre exemple de prévention des risques sismiques, avec cette nouvelle définition de la vigilance, nous pouvons vérifier à l'aide de PEPS-analyser

que la politique $\langle \{d, f\}, \{r11, r12, r1b1, r1b2, r2, r3', r4\} \rangle$, réécrite avec l'opérateur *filter*⁵, satisfait à la fois la nouvelle *geo*-vigilance et la *geo*-out-out-restriction pour le groupe *GRS*, la *sens*-restriction stricte, mais aussi la *geo*-complétude, la *sens*-complétude, la consistance, l'applicabilité et la minimalité.

7. Conclusion

Dans cet article, après avoir donné un bref rappel sur le langage PEPS et sur l'outil PEPS-analyzer, nous avons présenté et formalisé deux nouvelles classes de propriétés : les propriétés liées à l'exigence de diffusion et celles liées à l'exigence de non-diffusion d'information. Ainsi, nous avons montré que dans certains cas il est impossible de satisfaire ces deux exigences simultanément sans introduire un nouvel opérateur qui s'avère correspondre à une opération de filtrage d'information. Pour finir nous avons défini l'opérateur de filtrage de façon générique ainsi que redéfini la propriété de vigilance en fonction de cet opérateur générique.

Jusqu'à présent, nous nous sommes concentrés sur les concepts clés des politiques d'échange d'information. Dans des travaux futurs nous inclurons à ce cadre de spécification la modélisation des organisations qui composent le système. Nous devons déterminer si certains concepts, comme celui des *rôles*, issus des modèles tels que ORBAC (Kalam *et al.*, 2003), peuvent être transposés dans notre cadre, tout en adaptant si besoin les propriétés de base que nous avons définies. Techniquement parlant, la nature extensive de PEPS permet de modéliser aisément de tels nouveaux concepts.

Par ailleurs, il pourrait être intéressant de faire le lien entre la spécification et l'implémentation de ces politiques. Dans le domaine des Architectures Orientées Services, (Barhamgi *et al.*, 2013) montrent qu'il est possible, dans le contexte du contrôle d'accès à l'information, de définir des modèles d'exécution paramétrés par des politiques de confidentialité et d'appliquer ces politiques à l'exécution. Ici, le lien entre la spécification de la politique et sa mise en œuvre par composition des services, est explicite. Nous pourrions étudier si cette approche peut être adaptée dans le cadre du contrôle de la diffusion d'information.

Bibliographie

- Akl S. G., Denning D. E., « Checking Classification Constraints for Consistency and Completeness », *IEEE Symposium on Security and Privacy*, IEEE Computer Society, 1987, p. 196-201.
- Barhamgi M., Benslimane D., Oulmakhzoune S., Cuppens-Boulahia N., Cuppens F., Mrissa M., Taktak H., « Secure and Privacy-Preserving Execution Model for Data Services », Salinesi C., Norrie M. C., Pastor O., Eds., *CAiSE*, vol. 7908 de *Lecture Notes in Computer Science*, Springer, 2013, p. 35-50.

5. toutes les occurrences de $p(i)$ dans les règles de la politique d'échange ont été remplacées par *filter(filterSens, i)*.

- Castañeda H. N., *Thinking and doing*, D. Reidel, Dordrecht, 1975.
- Cholvy L., Delmas R., Polacsek T., « Vers une aide à la spécification d'une politique d'échange d'information dans un SI », *Actes du XXXème Congrès INFORSID, Montpellier, France, 29 - 31 mai 2012*, 2012, p. 352–370.
- de Moura L., Bjørner N., « Z3 : An Efficient SMT Solver », *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS 2008, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2008, Budapest, Hungary, March 29-April 6, 2008. Proceedings*, vol. 4963 de *Lecture Notes in Computer Science*, Springer, 2008, p. 337-340.
- Delmas R., Polacsek T., « Formal Methods for Exchange Policy Specification », Salinesi C., Norrie M. C., Pastor O., Eds., *CAiSE*, vol. 7908 de *Lecture Notes in Computer Science*, Springer, 2013, p. 288-303.
- Delmas R., Polacsek T., « Exigences de confidentialité et de diffusion concernant les politiques d'échanges d'information », *Génie Logiciel*, vol. 111, 2014, p. 49–53, GL & IS.
- Denning D. E., Akl S. G., Heckman M., Lunt T. F., Morgenstern M., Neumann P. G., Schell R. R., « Views for Multilevel Database Security », *IEEE Trans. Software Eng.*, vol. 13, n° 2, 1987, p. 129-140.
- Gallier J. H., « *Logic for Computer Science : Foundations of Automatic Theorem Proving* », chapitre 10, p. 448-476, Wiley, 1987.
- Kalam A. A. E., Benferhat S., Miège A., Baida R. E., Cuppens F., Saurel C., Balbiani P., Deswarte Y., Trouessin G., « Organization based access contro », *POLICY*, IEEE Computer Society, 2003, p. 120-131.
- Sebastiani R., Vescovi M., « Automated Reasoning in Modal and Description Logics via SAT Encoding : the Case Study of K(m)/ALC-Satisfiability », *J. Artif. Intell. Res. (JAIR)*, vol. 35, 2009, p. 343-389.

Une expérience d'un déploiement industriel de LELIE: une relecture intelligente des exigences

Juyeon Kang^{1,2}, Patrick Saint-Dizier²

1. Prometil

42, avenue du Général de Crouette 31100 Toulouse, j.kang@prometil.com

2. IRIT-CNRS

118, route de Narbonne 31062 Toulouse cedex, stdizier@irit.fr

RESUME. Cet article montre à travers de nombreux exemples et de nombreuses situations industrielles comment le prototype universitaire LELIE, un outil intelligent d'aide à la rédaction de documents techniques, a été transféré et valorisé dans un cadre industriel. LELIE est tout d'abord adapté de façon générale au traitement des exigences, suivant les normes et pratiques en vigueur. Nous montrons ensuite comment LELIE pour les exigences est adapté au contexte de chaque industriel, en particulier au niveau des ressources linguistiques et du suivi des pratiques des rédacteurs d'exigences.

ABSTRACT. In this article, we show how the academic prototype LELIE, an intelligent authoring tool for technical texts, developed at IRIT, has been transferred to an industrial context. We first show how it has been customized to the norms and general practices of requirement authoring. We then show how LELIE for Requirements is adapted to the local context of specific companies, in terms of linguistic resources and companies authoring practices of technical writers.

MOTS-CLES : Informatique linguistique, exigences, aide à la rédaction de textes techniques

KEYWORDS: Computational linguistics, requirements, technical text authoring tool

1. Problématique et motivations

Les documents techniques (procédures, exigences, spécifications) forment un genre conceptuel particulier qui suit des contraintes linguistiques fortes en termes de choix lexical (par exemple, quel terme utiliser dans une terminologie de termes métier), de syntaxe (structure des phrases et énoncés), typographique, de style et de contraintes métier variées. Les documents techniques sont conçus de telle façon qu'ils soient faciles à lire, efficaces (rien de superflu et d'un niveau de granularité adéquat), adaptés au lecteur ou à l'opérateur et le moins ambigus possibles. Ces documents doivent être explicites de façon à laisser le moins possible d'espace aux interprétations personnelles des utilisateurs, génératrices de risques. Dans ce but, plusieurs principes et recommandations de rédaction se sont développés : des principes généraux : ceux de la langue contrôlée générale (CNL) et ceux de nombreuses entreprises sous forme de guides ou de recommandations. Le résultat est un ensemble un peu hétérogène de principes qu'il faut adapter à chaque situation.

Les principes de la CNL incluent par exemple : (1) l'interdiction de l'emploi de la négation, du futur ou du passif, de termes passe-partout ou vagues, de termes flous, (2) la limitation de la complexité des énoncés : coordinations et relatives simples et limitées, phrases de longueur modérée, (3) le contrôle des titres, des renvois, du style : régularité des énumérations, position de certaines structures (conditions, expression du but, etc.). Ces principes, dans la réalité de la rédaction technique, sont souvent perçus comme trop contraignants et très complexes à mettre en œuvre. Lorsque l'on ajoute à cette difficulté le fait que les documents techniques sont souvent produits à partir de documents plus anciens via des 'copier-coller', qu'un texte est souvent rédigé par plusieurs rédacteurs et que la relecture d'un document est souvent fastidieuse et couteuse en temps, on comprend qu'il reste dans les textes techniques, même dûment validés, de nombreuses erreurs qui nuisent à leur bonne intelligibilité. Il en résulte des coûts parfois très élevés pour régler a posteriori des problèmes qui auraient pu être évités par une relecture plus soignée.

1.1 Le noyau de LELIE

Ces considérations ont motivé le développement du projet LELIE (projet ANR) (Barcellini *et al.* 2012), (Saint-Dizier, 2014), qui permet la détection de plusieurs catégories d'erreurs dans les documents techniques par le biais d'alertes. LELIE est un système à base de règles qui permet de traiter les contraintes générales de la langue contrôlée aussi bien que les contraintes de rédaction métier. LELIE autorise la détection des erreurs en fonction de la structure de discours ainsi que la définition de degrés de gravité pour chaque type d'erreur, ce qui permet au rédacteur de planifier les corrections d'une façon adaptée à ses priorités. LELIE a été implémenté sous la forme d'un noyau qui couvre les principales erreurs rencontrées en rédaction technique de façon indépendante du domaine. L'implémentation a été réalisée sur la plateforme d'analyse <TextCoop> (Saint-Dizier, 2012). LELIE suit un modèle de

spécification linguistique déclaratif, il est conçu pour être flexible et facile à faire évoluer.

Pour illustrer les possibilités du noyau de LELIE, le tableau ci-dessous résume les principales catégories d'erreurs traitées avec les taux d'erreurs rencontrées. Les documents analysés sont un ensemble de procédures, d'exigences, de spécifications de sécurité et de manuels de produits. Les documents proviennent de trois sociétés S1, S2, S3 de secteurs industriels différents qui restent anonymes à leur demande. Les documents considérés sont des documents validés et opérationnels. 300 pages ont été considérées dans cette analyse. Le tableau ci-dessous propose une moyenne pour chaque erreur pour un document typique de 30 pages :

Tableau 1. Résultats d'analyse du noyau de LELIE

Type d'erreur	Nbre pour 30 pages	S1	S2	S3
Termes flous	66	44	89	49
Déverbaux	29	24	14	42
Modaux dans les instructions	5	0	12	1
Pronoms avec référence incertaine	22	4	48	2
Négation	52	8	109	9
Structures de discours complexes	43	12	65	50
Coordinations complexes ou ambiguës	19	30	10	17
Constructions N+N lourdes ou compléments de noms trop longs	46	58	62	15
Passifs	34	16	72	4
Futurs	2	2	4	1
Phrases trop complexes ou trop longues	108	16	221	24
Enumérations irrégulières	moyen	bas	élevé	moyen
Références incorrectes à d'autres sections ou figures	13	33	22	2
Total erreurs	439			

Notons que dans ce tableau, la société S3 a fait relire ses textes par une société de rédaction spécialisée, il y demeure cependant un bon nombre d'erreurs. On observe que la distribution des erreurs est très diversifiée selon les sociétés, cela tient en particulier à la complexité des textes : ceux de S2 sont par exemple nettement plus complexes que ceux de S1. Enfin, on note qu'il y a en moyenne 15 erreurs par page, ce qui fait environ une alerte toutes les 2 à 3 lignes, hors erreurs liées au métier, ce qui est très considérable et motive largement l'emploi de LELIE.

Le noyau de LELIE est conçu indépendamment d'un domaine avec un lexique général et des règles de détection génériques. Par conséquent, selon les domaines et les textes, on observe un certain nombre d'oublis ou d'alertes inappropriées. Globalement, on observe, selon les types d'erreurs, un niveau de silence qui peut aller jusqu'à 15% et un taux de bruit allant jusqu'à 30%, en particulier pour les termes flous qui sont souvent flous en fonction du contexte métier. Il est clair qu'une adaptation au domaine et aux règles de la société permet d'améliorer grandement la qualité du système et devrait le rendre acceptable dans des contextes opérationnels. Cette communication présente une démarche d'industrialisation du noyau de LELIE appliqué aux exigences.

1.2 La structure du noyau de LELIE

LELIE est implémenté en Dislog, une extension à Prolog qui s'exécute sur <TextCoop>, un méta-interpréteur dédié en premier lieu à l'analyse du discours (Saint-Dizier, 2012 ; 2014). LELIE s'appuie sur Dislog de la façon suivante :

-Les structures incorrectes dans les énoncés d'exigences sont reconnues sur la base de templates qui décrivent ces structures,

-La production des alertes se fait par le biais de la réécriture : le template initial est réécrit avec l'insertion des alertes, par exemple sous forme XML, à l'endroit approprié.

-Les templates proposés par Dislog étendent le pouvoir expressif des expressions régulières : (1) en associant des structures de traits typés aux symboles préterminaux et non-terminaux, (2) en développant des symboles spécifiques, 'gaps', qui permettent d'ignorer des suites finies de mots qui n'ont pas d'intérêt pour ce template, tout en permettant de spécifier des contraintes sur ce qui est ignoré (par exemple les négations ne peuvent pas être ignorées), (3) en introduisant la possibilité de raisonnement, par exemple terminologique ou sur les connaissances du monde, pour affiner un diagnostic d'erreur, (4) enfin, en permettant d'effectuer des calculs pour produire une représentation de l'alerte qui soit précise et satisfaisante.

LELIE a une architecture linguistique modulaire, il est composé :

-d'un ensemble de templates qui décrivent les structures erronées, les autres structures étant par défaut correctes. Les règles sont relativement génériques et sont associées à des filtres pour traiter les cas particuliers (1) soit qui requièrent un traitement spécifique (2) soit qui ne sont pas des erreurs (par exemple le pronom indéfini 'it' ne pose pas de problème de référence pronominale incertaine).

-d'un ensemble de lexiques : langue générale, langue métier, marques spécifiques de certaines constructions.

A ces composants s'ajoutent des procédures de calcul et le moteur TextCoop. Le format général d'une règle de détection d'erreur est :

[patron d'erreur] → [patron de correction] {Contraintes, filtres}.

Un exemple typique de règles de LELIE est la détection de formes passives dans la proposition principale de l'exigence, où apparaît de modal typique de l'exigence (*shall*) :

it shall be demonstrated.... doit être évité:

[pro(*it*), modal, aux(*be*), gap, verb(past participle)] → [pro(*it*), modal, <passif>, aux(*be*), gap, verb(past participle), </passif>], {}.

Cet exemple, un peu simplifié, montre comment les alertes sont insérées dans le résultat. Il n'y a pas de contraintes spécifiques ici, mais on pourrait tolérer certains verbes courants au passif, qui seraient alors filtrés (la règle ne s'applique alors pas) :

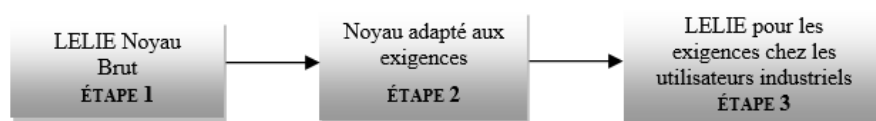
{diff(verb,[liste de verbes non concernés par l'alerte])}.

2. Solutions et techniques pour LELIE Exigences

Le noyau de LELIE a été conçu comme un modèle générique qui doit ensuite être adapté en fonction du sous-genre textuel considéré. C'est en particulier le cas pour l'aide à la rédaction des exigences avec leurs erreurs spécifiques. Nous considérons dans ce document des textes composés uniquement d'exigences ou bien des textes où les exigences sont clairement identifiées. L'analyse des erreurs porte uniquement sur les exigences. Nous nous focalisons sur l'anglais.

L'adaptation de LELIE aux exigences (Etape 2 de la figure 1) puis à un métier ou à une société spécifique (Etape 3 de la figure 1) se fait en deux étapes, afin de préserver une bonne organisation hiérarchique des traitements :

Figure 1. Etapes de développement de LELIE pour les exigences



2.1 La rédaction des exigences : outils et approches

Dans cette section, nous présentons brièvement quelques principes de rédaction définis dans les normes IEEE et les langues contrôlées, généraux tels que spécifiés dans (Weiss, 1991; O'Brien, 2003; Unwalla, 2004; Wyner, 2010; Alred *et al.* 2012; Kuhn, 2013; 2014) ainsi qu'une synthèse des consignes de rédaction développées dans les sociétés avec lesquelles nous avons développé nos expérimentations.

Les exigences se présentent souvent sous forme de listes ou bien de phrases composées identifiées et cadrées par un ID (par exemple, [REQ-SYS_0001]) et structurées en paragraphes. Une exigence se compose d'une ou de plusieurs phrases

qui en définissent par exemple les contraintes ou les différentes facettes (Grady, 2006 ; Hull, 2011). Les règles de la rédaction qui s'appliquent aux exigences ont été identifiées à partir de normes IEEE, de recommandations spécifiques aux exigences et de principes des langues contrôlées (Normes IEEE830-1998, ISO/IEC/IEEE 29148 :2011, ISO/IEC 12207, ARP4754, ASD-STE 100¹, VTT-R-01067-14²). Les contraintes de rédaction précisent la syntaxe et la sémantique ainsi que le style et le lexique que les rédacteurs doivent respecter. Les principales contraintes sont : (1) il faut éviter les termes flous ; (2) les termes utilisés doivent être consistants ; (3) il faut éviter trop de conjonctions de coordination ; (4) il faut utiliser le modal *shall* pour une exigence mandataire et *must* pour une exigence réglementaire; (5) il ne faut exprimer qu'une seule action par exigence, etc. A partir de ces règles, on peut développer un modèle linguistique de rédaction des exigences. Ce modèle s'appuie sur des ressources lexicales et des structures grammaticales typiques.

La rédaction des exigences est une tâche délicate car celles-ci doivent être précises, complètes et non-ambiguës. Le niveau d'imprécision doit être minimisé de façon à ce que le produit réalisé à partir de celles-ci ou les comportements qui en résultent soient le plus possible conforme à la demande ou aux normes. Pour ce faire, une approche à base de templates (Arora *et al.*, 2013) a été mise en place qui propose les structures de langue typiques qui peuvent être utilisées, à l'exclusion d'autres, comme par exemple :

[SUJET, MODAL, VERBE, OBJET(matériel), LIEU, BUT]

Cette approche a fait l'objet de quelques implémentations, où la conformité à ces templates est contrôlée, comme dans RAT-RQA (<http://www.reusecompany.com/>) ou Rubric (<https://sites.google.com/site/rubricnlp/>). Toutefois, dans la pratique quotidienne, en particulier lorsque les exigences sont complexes, ou pour certaines catégories d'exigences générales, ce modèle à base de templates n'est pas suivi ou ne l'est que partiellement. De plus, en termes de traçabilité et de flexibilité, il sera très difficile de suivre et mettre à jour les exigences rédigées à base de ces templates lorsque l'on modifie quelques éléments de ceux-ci. La plupart des rédacteurs souhaitent en effet avoir la liberté de s'exprimer assez librement. LELIE ne propose pas de templates a priori, il intervient a posteriori, en produisant des alertes par rapport aux contraintes de rédaction qui ont été spécifiées ou sélectionnées par le rédacteur.

Le système RAT-RQA offre des fonctionnalités dans un registre proche de notre travail, appliqué aussi aux exigences, mais avec très peu de flexibilité dans la prise en compte de l'opérateur et de la tâche. Ce système est doté d'une interface simple pour son implantation dans une entreprise (où sont essentiellement gérés les aspects lexicaux propres à la société). Ce système permet de définir des règles de rédaction propres à une entreprise (les boilerplates), mais il ne traite pas du style, de la cohésion et ne prend en compte les contraintes discursives dans les textes.

¹ Simplified Technical English, Aerospace and Defense, by Industries Association of Europe, Issue 5, 2010

² Controlled natural language requirements in the design and analysis of safety critical I&C systems, Research Report, SAREMAN Project, 2013

Le système proposé par Attempto (Fuchs, 2012) (<http://attempto.ifi.uzh.ch/site/>) implémente les principes de la rédaction contrôlée de façon générale, appliquée à l'anglais. Un langage formel a été associé (ACE, basé sur la logique et la programmation en logique) qui permet de développer des notations et des méthodes formelles en traitement de la langue naturelle. C'est donc un outil à spectre large en TALN, mais où l'adaptation à un contexte industriel particulier semble laissé ouvert à l'utilisateur, sans support méthodologique particulier.

2.2 LELIE dans le contexte général des exigences

LELIE adapté aux exigences (*LELIE for Requirements*) se base sur le modèle générique du noyau présenté dans la section 1.2 et utilise ses ressources lexicales et règles de base. Ces ressources lexicales contiennent d'une part, non seulement un lexique générique comme les prépositions, les modaux et les verbes d'action mais aussi des ressources davantage orientées vers l'aide à la rédaction telles que les adjectifs et adverbes identifiés comme étant flous a priori, et d'autre part, les règles décrivant des erreurs de rédaction comme, par exemple, le repérage des compléments de noms en cascade à éviter. Ces éléments étant développés indépendamment du domaine et du type de textes (c'est-à-dire, applicable à tout type des documents techniques), il nous fallait une évaluation plus précise des performances sur ces derniers afin de déterminer les points d'amélioration. L'évaluation a été effectuée sur un corpus de 100 pages de texte qui ne contient que des exigences de trois domaines différents : télécommunication, transport, sécurité. A titre indicatif, pour la détection des mots flous, le résultat de cette évaluation a atteint à peu près à 30% de précision et 70% de bruit, et pour les pronoms personnels à référence incertaine, 58% de précision et 42% de bruit (faux positifs).

Cette analyse nous a permis d'organiser les tâches suivantes:

- a) Synthétiser les caractéristiques de la rédaction des exigences dans les standards et les normes
- b) Déterminer les erreurs les plus fréquentes dans la pratique de la rédaction
- c) Déterminer les règles cruciales et prioritaires de la rédaction à implémenter dans notre système
- d) Revoir les vocabulaires pertinents ou non pertinents dans le lexique (termes flous, termes à éviter, ambigus, etc.)

Les règles d'analyse des erreurs ont été réalisées sur la base des standards et normes de la rédaction et ont été renforcées à partir des erreurs vues dans la pratique de la rédaction lors de l'analyse de corpus (méthode manuelle, effectuée par deux rédacteurs techniques et un linguiste), composé de 25 textes d'exigence de différents domaines (télécommunication, transport, banque, écologie).

Nous présentons une partie des règles génériques en fonction de niveaux d'analyse linguistique. Au niveau du lexique, nous vérifions les emplois adaptés des termes :

- 1) Termes flous : par ex. *wherever possible, suitably, adequately*

Ex.: *Intimate contact between dissimilar metals shall be avoided wherever possible.*

Chaque terme a un degré de gravité a priori sur une échelle qualitative de flou.

- 2) Termes inutiles : par ex. *also, then, always*

Inutiles au moins dans certains contextes.

- 3) Expressions interdites ou incomplètes : par ex. « *and/or* », « TBD », « TBC », « etc. »...

Au niveau de la grammaire, nous contrôlons les formes complexes :

- 4) Utilisation de la négation et de la double négation

- 5) Emploi de pronoms à référence incertaine : par ex. *their, them, these,*

- 6) Utilisation de relatives complexes : par ex. *that x...which...*

- 7) Utilisation de conjonctions de coordination : par ex. *and, et/ou or*

Ex.: *FCV shall have a visual mechanical position indicator with the full open and full closed positions clearly and permanently marked into the valve body adjacent to the mechanical position indicator, using the words "OPEN" and "CLOSED".*

- 8) Renvoi difficile : par ex. *below, above, see...*

Ex.: *Testing shall consist of resonance search, endurance dwell and endurance sweep cycling as defined below.*

- 9) Phrases complexes: par ex. la longueur de la phrase, la présence de nombreux acronymes ou de termes techniques non définis, les structures complexes (par ex. [SUJET],[CONDITION],[ACTION]-[OBJET]).

- 10) Construction passive : par ex. imprécision sur qui doit effectuer l'action

Ex.: *it shall be demonstrated as defined in ...*

Au niveau de l'uniformité du style, nous contrôlons la structure si les exigences sont rédigées de façon homogène :

- 11) Coordination des exigences

Ex.: *REQ1 Testing shall consist of resonance search, endurance dwell and endurance sweep cycling as defined below and REQ2 shall be performed in the sequence indicated*

- 12) Régularité des temps : par ex. *shall set or set* au présent

- 13) Position structurelle: par ex. CONDITION-SUJET-ACTION-OBJET

Ex.: [CONDITION] *When signal x is received,* [SUJET] *the system* [ACTION] *shall set* [OBJET] *the signal x received bit* [CONSTRAINT] *within 2 seconds.*

- 14) Enumérations hétérogènes; par ex. contrôle sur les listes énumérées

a. *system interaction*

b. *user interface*

c. *update is not applicable* : non homogène avec les 2 précédents

Ces fonctions sont adaptées à la tradition rédactionnelle des différentes sociétés. Par exemple, la fonction liée à la détection des conjonctions de coordination a six sous-catégories comme deux conjonctions dans une proposition conditionnelle, plus de trois conjonctions dans une proposition subordonnée, etc.

2.3 Prise en compte du métier

Pour que la détection des erreurs soit adéquate à la pratique de la rédaction industrielle, il est primordial de prendre en compte les connaissances des pratiques de rédaction du métier. Par exemple, certaines industries veulent voir la condition dans laquelle une exigence doit être effectuée toujours devant la proposition principale qui exprime l'action (<CONDITION> *Si la température est supérieure à 36°C, <ACTION> le système doit fonctionner...*) et/ou il peut y avoir les termes préférés (par ex. *prise de courant* doit être remplacé par *fiche à N broches*) et les termes métiers. Certains termes généralement flous comme *probable* ou *normale* peuvent être non flous lorsqu'ils sont utilisés dans un contexte industriel précis. Dans ce sens, nous avons développé les filtres qui s'appliquent selon le contexte dans lequel la règle et le lexique sont utilisés. Le mot *normale* sera détecté comme flou dans un contexte et filtré dans un autre contexte. Ces filtrages augmentent le taux de pertinence du système en permettant de minimiser les faux positifs.

Le résultat de cette prise en compte du métier est mesuré par une évaluation qualitative. Le corpus de l'évaluation est composé de 50 pages extraites de six spécifications sur trois domaines : aérospatial, télécommunications, transports. Ce corpus ne comporte que des exigences, il est validé comme étant suffisamment complexe par les rédacteurs techniques. Le corpus a été pré-annoté manuellement par le valideur-rédacteur sans regarder auparavant les annotations faites par LELIE.

Tableau 2. Résultats de l'évaluation après l'adaptation de LELIE aux exigences

Type d'erreurs évaluées	Précision	Rappel	Silence	Bruit (Faux positifs)
Termes flous	75% (295/395)	77% (295/383)	23% (88/383)	25% (100/395)
Trop de conjonctions de coordination	74% (17/23)	93% (17/18)	6% (1/18)	26% (6/23)
Trop de négations	100% (1/1)	100%	0%	0%
Exigences multiples	63% (37/59)	57% (37/65)	43% (28/65)	37% (22/59)
Expressions interdites/incomplète	88% (63/72)	72% (63/88)	28% (25/88)	13% (9/72)
Constructions passives	93% (101/109)	69% (101/146)	31% (45/146)	7% (8/109)

3. L'implantation industrielle

Le noyau de LELIE dans un contexte des exigences étant réalisé, nous nous penchons à présent sur la prise en compte précise des métiers. Ce travail s'appuie sur les rédacteurs de six sociétés. Nous décrivons dans cette partie nos différentes expériences. Pour des raisons de confidentialité, les sociétés restent anonymes et les résultats présentés sont une synthèse de nos évaluations.

3.1 Analyse des spécifications de différents domaines

3.1.1 Approche

Les spécifications techniques envoyées par les sociétés en tant que valideurs-utilisateurs, sous une forme variée (le format .dma exporté de DOORS, Word, Excel), ont été étudiées manuellement pour analyser les structures et les termes de chaque exigence (identifiée par un ID).

Le premier objectif de cette étude a été de déterminer à quel point les règles de l'outil sont en adéquation par rapport à l'observé. Nos observations ont entraîné l'ajout de données lexicales et de filtres à LELIE en complément des premiers ajouts. Ces ajouts plus spécifiques doivent être contrôlés et sont à l'origine d'une méthode de customisation à l'étude.

Le deuxième point important a été le développement d'un lexique spécifique au domaine et au client. Dans cette étape, les termes métiers et termes préférés ont été identifiés et ajoutés dans nos lexiques spécifiques selon le domaine et le client. Le lexique générique reste stable ayant en moyenne 75% de pertinence comme le montre le résultat de l'évaluation (voir le tableau 2), cependant le lexique spécifique des termes métiers et préférés doit évoluer, en particulier, lorsque l'on applique l'outil à un nouveau domaine ou client.

3.1.2 Résultats

Les résultats de l'analyse des erreurs dépendent largement du fait que l'on adapte ou non le lexique et les règles au contexte du projet appliqué. Notamment, la pertinence de la détection des termes flous peut être considérablement augmentée (environ 20 à 30%) par cette adaptation.

A chaque déploiement industriel, nous avons effectué une analyse statistique des résultats en matière de pertinence. Nous présentons dans le tableau suivant le taux des exigences ayant au moins une alerte de LELIE. Il s'agit d'une synthèse pondérée sur les taux obtenus à partir des résultats statistiques de six sociétés. Ces résultats (voir le tableau 3 et la figure 2) sont une indication de la pertinence de l'outil, ils constituent un ensemble d'éléments pour une évaluation indicative.

Tableau 3. Nombres des erreurs par type sur toutes les spécifications testées

Sur un ensemble de 4403 exigences analysées, 2708 exigences (62%) avec au moins une alerte de LELIE	Termes flous	1332
	Trop de conjonctions de coordination	91
	Trop de négations	29
	Expressions à éviter	715
	Construction passive	967
	Exigences multiples	325
	Termes inutiles	21
	Contrôle de la structure de l'exigence	108
	Contrôle de la partie Rationale	79
	Exigence trop longue	306
	Total	3973

Comme le montre le tableau 3, l'outil a détecté un problème en moyenne dans 62% des exigences et les taux minimal et maximal sont respectivement de 52% et 72%, ce qui est très élevé.

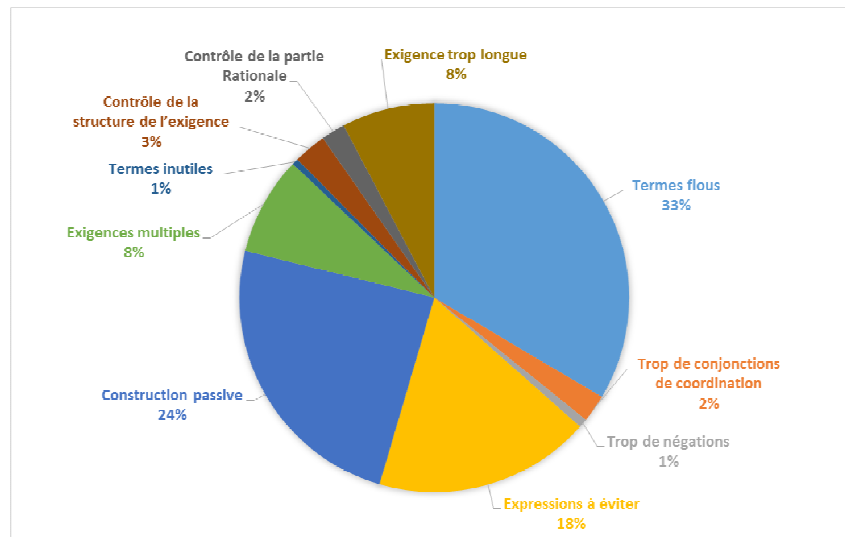


Figure 2. Distribution des erreurs par type

Ce résultat nous donne plusieurs indications par rapport à la pratique rédactionnelle des sociétés concernées : 1) le style de la rédaction (par ex. certaines spécifications majoritairement rédigées sous une forme passive), 2) le niveau de

maturité de la rédaction (par ex. certaines spécifications contenant de nombreux termes indiquant un contenu incomplet comme *To Be Determined*, moins de termes flous dans les spécifications matures que celles de tout début), 3) l'organisation syntaxique de la proposition de l'exigence, 4) le niveau discursif de l'exigence (par ex. la structure du discours comme condition, but, circonstance, illustration...).

Le résultat de certaines sociétés était plus varié que d'autres dans le sens où les spécifications d'une même société ont été rédigées chacune avec un style différent ou pas, par exemple, une spécification avec la forme passive (par ex. *shall be Verb-ed*), une autre avec la forme négative (par ex. *shall not*). Cela montre que la façon dont les rédacteurs rédigent leurs spécifications n'est pas homogène même au sein d'un groupe des rédacteurs ou d'une société en absence de guides spécifiques.

3.1.3 Retours des évaluations

Les évaluations que nous avons menées à l'occasion de différents déploiements industriels sur cet outil et les discussions avec les rédacteurs techniques peuvent être présentées selon deux perspectives complémentaires: l'analyse interne et le retour des clients.

Le premier point correspond à l'analyse des nouveaux besoins suscités par les possibilités de l'outil, cela concerne les trois aspects suivants :

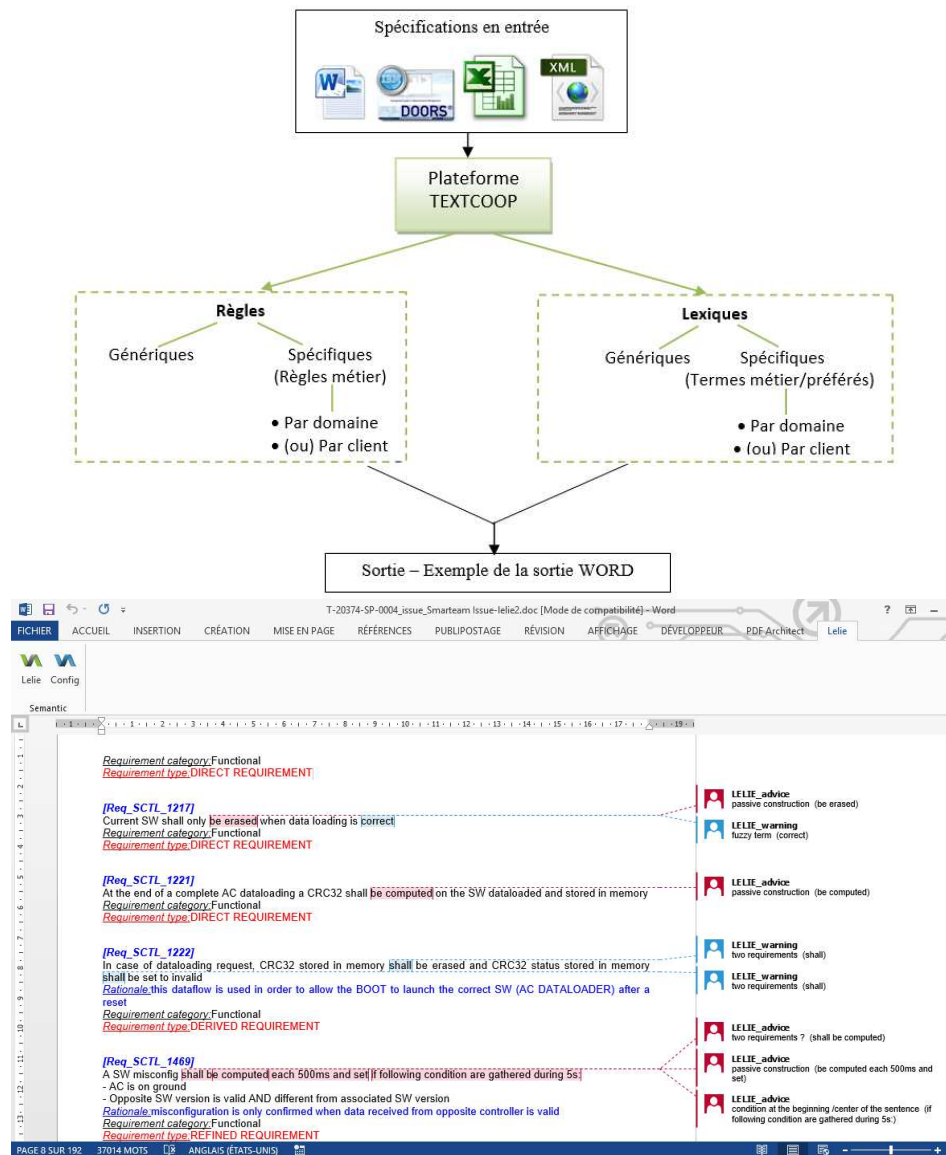
- 1) Aspect fonctionnel : nouvelles fonctions (par ex. analyse de la redondance, de la cohésion ou de la cohérence entre les exigences)
- 2) Aspect technique : amélioration d'interface, modes d'intégration des commentaires aux utilisateurs dans l'outil
- 3) Aspect ergonomie cognitive : adéquation des alertes au fonctionnement du rédacteur.

Le second point, le retour des clients, concerne davantage la possibilité d'activation/désactivation de certaines règles et vocabulaires spécifiques par leur soin. Nous observons que (1) certaines erreurs identifiées comme étant critiques sont acceptées sans problème dans la pratique quotidienne des rédacteurs de certaines sociétés et (2) certains vocabulaires identifiés comme étant problématiques dans la plupart des cas sont acceptables et compréhensibles sans difficulté par les rédacteurs de certaines sociétés.

Ces expériences industrielles nous permettent de constater que l'adaptation d'un domaine à l'autre est plus coûteuse que d'un client à l'autre sur un même domaine. Lorsque l'on aborde un nouveau domaine, le volume du lexique à ajouter peut être de 70% à 100% du lexique existant pour les termes métier. Il s'agit de développer un nouveau lexique spécifique des termes métier à un domaine. Pour les termes génériques, cela correspond à 3% à 5% de l'ajout des nouveaux vocabulaires dans le lexique générique existant, ce qui est très minime. Pour minimiser le travail d'entrées lexicales pour chaque déploiement du domaine, nous développons une plateforme d'acquisition. Nous constatons que cet outil est indispensable pour minimiser le coût de déploiements industriels et sécuriser la mise en place de l'outil chez les utilisateurs finaux. Enfin, les règles de détection existantes restent

relativement stables pour différents domaines car les exigences sont souvent rédigées selon le même schéma, indépendamment du domaine, par contre, des règles spécifiques ou des filtres peuvent être ajoutées lorsqu'une société a ses propres règles de rédaction. Nous représentons dans la figure 3 l'architecture fonctionnelle élaborée et actuellement mise en opération dans le contexte de l'ingénierie des exigences.

Figure 3. Architecture fonctionnelle de LELIE pour les exigences



Le tableau 4 donne le nombre de règles de base et de filtres contextuels ainsi que les entrées lexicales génériques et spécifiques (termes métiers), développées pour les exigences. Ceci donne une mesure des volumes pour le système complet.

Tableau 4. Nombre de règles et volume des lexiques

	Nbre de règles		Lexiques (entrées lexicales)	
	de base	de filtres de contextualisation	génériques	termes métier
Noyau de LELIE	13	1	5681	0
LELIE pour les exigences	37	30	8467	680

3.2 Problèmes et limites : vers une mémoire de correction

Les principes de rédaction de la langue contrôlée ou bien recommandés par les sociétés sont souvent ressentis comme trop rigides par les rédacteurs techniques : elles manquent de flexibilité et de prise en compte du contexte. En effet, dans *‘l’opérateur doit fermer progressivement la vanne’* l’adverbe de manière *progressivement* n’est pas véritablement flou car l’action de fermer est vue comme ponctuelle. Par contre, il l’est dans : *‘le pilote doit réduire progressivement la vitesse à 145 kts’* car cette réduction de vitesse n’est pas une opération ponctuelle. Dans les deux cas, ce n’est pas la manière qui est mise en cause mais plutôt la durée de l’action, donc sa dimension temporelle. Le contexte *vanne* versus *vitesse avion* est crucial dans la détermination d’un emploi flou ou non des adverbes de manière en général. L’énoncé *‘Ce produit ne doit pas être jeté à l’évier’* est incorrect du point de vue de la CNL dans la mesure où la négation est interdite. Cependant, dans ce cas, il est très difficile, si l’on ne connaît pas exactement l’environnement de travail de l’opérateur, de remplacer la négation par le lieu où il faut jeter le produit.

Une analyse trop directe des énoncés, sur la base d’un système de règles, quelque peu rigide, va générer des alertes incorrectes, donc beaucoup de bruit, même si on associe aux règles de détection des erreurs de nombreux filtres. Nous évoluons à présent vers un système hybride qui associe LELIE à un système d’apprentissage qui va observer les corrections faites (ou pas) par les rédacteurs lorsqu’ils examinent les alertes produites par LELIE. Cette observation va permettre d’apprendre quelles alertes sont de réelles alertes et aussi, de façon très précieuse, comment les erreurs sont corrigées. Une première expérimentation a été développée dans (Kang et Saint-Dizier, 2014) sur le suivi des termes flous. Il en est résulté des spécifications de développement d’une mémoire de correction. Très brièvement, cette expérimentation a permis de dégager les éléments suivants :

- A partir d’un certain seuil de situations non corrigées, pour un contexte donné, l’alerte n’est plus produite. La notion de contexte est définie par quatre termes (verbes, noms, adjectifs) à gauche ou à droite du terme flou et de son focus. La suspension de l’alerte à d’autres termes proches est généralisée via la référence à la terminologie du domaine,

- Les corrections effectuées sont mémorisées. Une généralisation ou un ensemble de généralisations (Daelemans *et al.*, 2005) sont réalisées et proposées aux rédacteurs : l'objectif est de faire un choix unique sur l'une de ces généralisations, et de l'appliquer ensuite systématiquement dans un contexte donné, afin de garantir une bonne homogénéité des corrections. La décision est prise par un administrateur ou collectivement par médiation.
- Des propositions de correction issues de ces généralisations sont proposées. Les corrections sont structurées sur deux niveaux : un patron général de correction qui contient des éléments sous-spécifiés : les recommandations de correction contextuelles, qui sont instanciées en fonction du contexte.

Par exemple, pour un déterminant flou tel que '*a few*' appliqué à un nom *Noun* :
[a few Noun] → [less than X Noun].

Le patron de correction propose de remplacer *a few* par une expression non floue: moins de N, où N dépend du contexte. Par exemple, 'N=12' dans *take-off a few knots above V1* → *take-off less than 12 knots above V1*, avec

Contexte = [focus: knots, autres: take-off, above V1].

De même pour le cas de *progressivement*, un patron général de correction est induit: [*progressively* Verb(durative)] → [*progressively* Verb(durative) in X(time)].

Où une variable contextuelle indiquera la durée, par exemple *3 minutes* pour la réduction de vitesse de l'avion dans l'exemple donné en 3.2.

Nous allons poursuivre cette expérimentation avec d'autres catégories d'erreurs complexes à régler telles que la négation ou les phrases trop complexes. Ce système de mémoire de correction interviendra alors à la suite des règles de LELIE :

- (1) pour filtrer les alertes qui n'en sont pas, améliorant en cela le confort du rédacteur, ainsi, pour les termes flous, nous sommes passés de 30% de bruit à environ 12%,
- (2) pour proposer des éléments de correction qui restent toujours à l'appréciation des rédacteurs. Ceux-ci apprécient les alertes, mais sont en fait très demandeurs de solutions de correction.

4. Conclusion et perspectives

Nous avons présenté dans ce document les résultats d'une expérience de déploiement industriel à partir du noyau de LELIE, prototype universitaire. Un outil de relecture automatisée des exigences a été développé via différentes étapes d'études en adaptant le modèle générique du noyau de LELIE au contexte des exigences puis à celui des utilisateurs. La dernière version de LELIE pour les exigences est toujours en cours d'évolution dans le cadre du déploiement industriel. Dans l'étape suivante de notre travail, cette version plus finalisée sera évaluée et comparée aux résultats de la première évaluation présentés dans le tableau 2.

Outre ces résultats nous réfléchissons au développement de plusieurs éléments complémentaires: nouveaux types d'erreurs suscités par le dialogue avec les

utilisateurs, par exemple les structures N+N en anglais (Garnier, 2011), interface IHM plus avancée, acquisition lexicale automatisée et éditeur de règles, outil de mise au point pour le linguiste et intégration dans d'autres services (suites éditoriales). D'un point de vue plus fondamental, la notion de mémoire de correction nous paraît essentielle. Nous prévoyons aussi de développer des outils d'analyse de la cohérence entre exigences, mais ceci demeure très complexe.

Références

- Alfred G.J. *et al.* (2012). *Handbook of Technical Writing*. St Martin's Press, NewYork.
- Arora C. *et al.* (2013). Automatic Checking of Conformance to Requirement Boilerplates via Text Chunking: An Industrial Case Study, *7th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM2013)*. Baltimore, MD, USA.
- Barcellini F *et al.* (2012). Risk Analysis and Prevention: LELIE, a Tool dedicated to Procedure and Requirement Authoring, *LREC*, Istanbul.
- Daelemans W. et van Der Bosch A.(2005). *Memory-Based Language Processing*, Cambridge.
- Fuchs N.E. (2012). First-Order Reasoning for Attempto Controlled English, In *Proceedings of the 2nd International Workshop on Controlled Natural Language (CNL2010)*, Springer.
- Garnier, M., (2011) Correcting errors in N+N structures in the production of French users of English, *EuroCall*, Nottingham.
- Grady J.O. (2006). *System Requirements Analysis*, Academic Press, USA.
- Hull E. *et al.* (2011). *Requirements Engineering*, Springer.
- Kang J. et Saint-Dizier P. (2014). Towards an Error Correction memory to Enhance Technical Texts Authoring in LELIE, *CNL'14*, Galway, Ireland, August.
- Kuhn T. (2013). A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language and Information*, 22(1).
- Kuhn T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1).
- O'Brien S. (2003). *Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets*. Dublin City University report.
- Saint-Dizier P. (2012). Processing Natural Language Arguments with the <TextCoop> Platform. *Journal of Argumentation and Computation* 3(2).
- Saint-Dizier P. (2014). *Challenges of Discourse Processing: the case of technical documents*. Cambridge Scholars, UK.
- Unwalla M. (2004). *AECMA Simplified English*, <http://www.techscribe.co.uk/ta/aecma-simplified-english.pdf>.
- Weiss E. H. (2000). *Writing remedies. Practical exercises for technical writing*. Oryx Press.
- Wyner A. *et al.* (2010). *On Controlled Natural Languages: Properties and Prospects*.

Session Recommendation

Approche hybride de recommandation des nouvelles collaborations

Nesrine BEN YAHIA¹, Wala REBHI², Narjès BELLAMINE³, Henda BEN GHEZELA⁴

¹ RIADI-ENSI, Université de la Manouba, nesrine.benyahia@ensi.rnu.tn

² RIADI-ENSI, Université de la Manouba, wala.rebhi@ensi.rnu.tn

³ RIADI-ENSI, Université de la Manouba, narjes.belamine@ensi.rnu.tn

⁴ RIADI-ENSI, Université de la Manouba, henda.benghazala@ensi.rnu.tn

RÉSUMÉ. Suite à l'émergence des organisations et en particulier celles distribuées, les groupes sont très dynamiques et les gens sont constamment à la recherche de nouveaux collaborateurs. Dans ce contexte, nous proposons une nouvelle approche de recommandation des collaborations qui cherche les collaborateurs potentiels et pertinents. Cette approche se base sur un nouvel algorithme de détection de communautés pour composer dynamiquement les groupes en considérant simultanément les interactions sociales entre les membres de l'organisation et leurs homophilies. L'approche proposée a été expérimentée sur un graphe de terrain représentant un laboratoire de recherche en le considérant comme une organisation où les membres représentent les chercheurs et les interactions représentent leurs relations de co-publication et de co-encadrement.

ABSTRACT. With the spread of organizations and especially distributed ones, teams are highly dynamic and people are continually seeking for new collaborators. In this context, we propose an approach of collaboration recommendation that relies on community detection technique to find potential collaborators that can help in problem solving. Our contribution consists on proposing a combined metric that considers both social connections and homophily of participants. The proposed approach has been tested on a real graph representing a research laboratory where members represent researchers and interactions represent their relationships of co-publishing and co-supervision.

MOTS-CLÉS : recommandation de collaborations, détection de communautés, interaction sociale, homophilie.

KEYWORDS: collaboration recommendation, community detection, social connections, homophily.

1. Introduction

Le travail des organisations est en grande partie un travail de prise de décisions et de résolution de problèmes. En outre, avec l'émergence des organisations et en particulier celles distribuées, les équipes sont de plus en plus très dynamiques et les gens sont constamment à la recherche de nouveaux collaborateurs pour les aider dans la résolution de leurs problèmes (Matthews *et al.*, 2012). Aujourd'hui, le concept de communauté de pratiques est devenu un objectif d'analyse utile pour identifier et caractériser « les collectivités » (groupes de personnes) pertinentes qui participent à la création des bonnes pratiques et la génération des solutions innovantes via le partage des expériences et la socialisation (Ackerman *et al.*, 2013).

Ainsi, la collaboration est importante pour les organisations qui se veulent compétitives et comment les collaborateurs sont sélectionnés est une chose cruciale pour prendre les bonnes décisions (Matthews *et al.*, 2012). Dans cette direction, des efforts techniques ont été élaborés pour soulever le problème d'identification des collaborateurs les plus pertinents mettant en place des systèmes de recommandation des collaborations (Ackerman *et al.*, 2013).

Dans cet article, nous nous intéressons à la détection et la décomposition dynamique des communautés au sein des organisations pour la recommandation des nouvelles collaborations pour un membre cherchant les personnes pertinentes qui peuvent l'aider dans la résolution de son problème. Pour ce faire, nous proposons une approche hybride de détection de communautés qui prend en considération simultanément les interactions sociales et l'homophilie des individus. Cette recommandation ne cherche pas à identifier les collaborateurs les plus pertinents de façon indépendante mais plutôt elle cherche à identifier tout un groupe de collaborateurs (une communauté).

Cet article est organisé comme suit : dans la section 2, nous présentons un tour d'horizon de différentes techniques utilisées pour la recommandation de collaborations. La section 3 présente l'approche hybride proposée qui combine les interactions sociales et l'homogénéité des individus. Dans la section 4, une expérimentation et une évaluation de cette proposition seront présentées.

2. Une vue globale des techniques de recommandation des collaborations

Au sein d'une organisation, (Lu et Zhou, 2011) affirment que les personnes «similaires» sont susceptibles d'être connectées et font la distinction entre deux définitions du mot «similaire» : une similarité basée attributs et une similarité structurelle. La première considère le partage de nombreux traits en commun, quant à la deuxième considère exclusivement la structure du réseau. Dans cette section, nous présentons une revue des principales techniques qui ont été utilisées pour recommander des nouvelles collaborations selon ces deux points de vue.

2.1. Techniques basées sur la similarité structurelle

S'appuyant sur le principe de la similarité structurelle, il y a une tendance à intégrer les réseaux sociaux dans les systèmes dans le but de faciliter la recherche des nouvelles collaborations. Dans ce contexte, certaines solutions utilisent la visualisation des réseaux sociaux comme une indication et un moyen de trouver une personne avec qui collaborer (Smith *et al.*, 2000). Par exemple, (Kautz *et al.*, 1997) utilisent les relations de copublication et de co-citation pour créer un réseau social et sa visualisation est utilisée pour trouver un expert possible. (Nardi *et al.*, 2002) proposent une application de gestion de communication qui utilise une visualisation de réseau social pour présenter des contacts et favoriser la collaboration entre les individus. D'autres solutions tentent d'utiliser les informations des réseaux sociaux pour recommander des personnes compétentes avec qui collaborer. Par exemple, (Ogata *et al.*, 2001) proposent de trouver des collaborateurs par le biais des réseaux sociaux et des échanges électroniques dans une organisation en considérant un collaborateur potentiel en tant que candidat qui échange fréquemment, récemment et réciproquement des e-mails avec le demandeur de collaboration.

2.2. Techniques basées sur la similarité des attributs

Dans ce contexte, un phénomène appelé homophilie indique que les gens ont tendance à collaborer avec d'autres qui ont les mêmes intérêts et les caractéristiques démographiques similaires (McPherson *et al.*, 2001). Différents travaux se sont basés sur ce principe d'homophilie pour sélectionner des collaborateurs potentiels, mais ils ont utilisé différentes techniques pour trouver ces personnes. (Liben-Nowell et Kleinberg, 2003) ont examiné les différentes méthodes pour la prédiction des liens et ont affirmé que la méthode des voisins communs (la similarité entre deux noeuds x et y est le nombre de voisins que x et y ont en commun) et le coefficient de Jaccard (métrique très connue dans le domaine de recherche d'information qui permet de mesurer la probabilité que les deux noeuds x et y ont une même caractéristique pour une caractéristique choisie au hasard) donnent les meilleurs résultats dans la plupart des cas. (Dong *et al.*, 2012) ont utilisé également la technique de prédiction des liens (link prediction) en se basant sur des mesures de similarité pour analyser la proximité des noeuds (représentant les personnes) dans un réseau. Et, ils ont conclu que les noeuds très similaires sont plus susceptibles d'être reliés dans le futur réseau. Avec le même principe d'homophilie, d'autres travaux ont utilisé la technique de filtrage collaboratif classique utilisée dans les systèmes de recommandation des items (tels que les films) en se basant sur les opinions (aimer ou non) des utilisateurs afin d'aider les personnes dans le choix des éléments. (Cai *et al.*, 2010) ont réutilisé ce principe en se basant sur deux aspects de similarité : deux personnes sont similaires (1) si elles sont sélectionnées par un ensemble d'utilisateurs en commun ou (2) si elles sélectionnent à la fois un ensemble d'utilisateurs en commun.

2.3. Synthèse des techniques de recommandation des collaborations

La combinaison des deux points de vue cités précédemment assume une approche de recommandation basée sur une similarité des attributs et une similarité structurale. En outre, et comme les communautés sont considérées selon (Wenger et Snyder, 2000) comme l'un des moyens les plus rapides pour la résolution de problèmes dans les organisations, nous proposerons dans la section suivante une approche hybride de recommandation des nouvelles collaborations en nous basant sur la détection de communautés et en considérant simultanément les interactions sociales (la structure de l'organisation) et l'homophilie des individus (la similarité des attributs).

3. Proposition d'une approche hybride de recommandation des collaborations

L'approche proposée se base sur la combinaison d'une technique de fouille de réseaux sociaux i.e. la détection de communautés et d'une méta-heuristique i.e. l'optimisation par essaim particulière pour l'optimisation de la qualité de la communauté détectée. Nous allons commencer par la présentation de ces deux techniques. Ensuite, nous allons présenter notre contribution qui réside dans la proposition d'un nouvel algorithme d'optimisation de détection de communautés pour la recommandation des nouvelles collaborations (estimées les plus pertinentes).

3.1. Techniques utilisées

3.1.1. La détection de communautés (DC)

La détection de communauté est considérée comme une technique de classification non supervisée qui est multidisciplinaire où l'utilisation du partitionnement de graphe, l'informatique, et la sociologie sont combinés (Newman et Girvan, 2004). Elle vise à diviser les noeuds du réseau (représentant l'organisation) en groupes (communautés) avec un faible couplage (relations extérieures entre les communautés) et une forte cohésion (relations internes au sein des communautés). Afin d'évaluer la qualité de la détection de communautés, (Newman et Girvan, 2004) ont proposé une métrique, appelée la modularité qui doit être maximisée pour résulter un bon partitionnement. Elle représente, pour un partitionnement donné, la différence entre la fraction d'arêtes dans le réseau considéré et la probabilité (la valeur attendue) de la même quantité dans un réseau avec le même partitionnement mais avec une répartition au hasard des arcs en respectant les degrés des noeuds.

La formule de cette métrique est donnée par (1) où n représente le nombre des sommets du réseau, A_{ij} est le nombre d'arêtes entre les sommets i et j (qui est normalement 0 ou 1), donc A_{ij} sont les éléments de la matrice d'adjacence. A_{ij} représentent la proportion d'arêtes avec un sommet dans la classe (communauté) C_i et l'autre dans la classe C_j . Ensuite, le nombre prévu d'arêtes entre les sommets i et j si les arêtes sont placées aléatoirement est $\frac{k_i k_j}{2m}$, où k_i et k_j représentent respectivement les degrés de sommets i et j , m est le nombre total d'arêtes dans le réseau et $\delta(C_i, C_j)$ est le

symbole de Kronecker qui vaut 1 si les arguments C_i et C_j sont égaux et 0 sinon (C_i représente la communauté qui contient les sommets i). Pour les graphes pondérés, qui possèdent des poids sur les arêtes, la même formule de modularité est utilisée et A_{ij} devient le poids de l'arête entre i et j , k_i est la somme des poids des arêtes attachées aux sommets i et m est égal à la somme des poids des arêtes (Blondel *et al.*, 2008).

$$M(P) = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(C_i, C_j) \quad (1)$$

Dans le cadre de cet article, nous avons sélectionné cette technique parce que la modularité permet de combiner la technique de la similarité structurelle (en considérant les interactions sociales) et la technique de la similarité basée attributs (en considérant l'homophile entre les sommets). Toutefois, dans notre contexte, notre objectif est de déterminer pour un membre donné (représenté par un sommet dans le réseau) le groupe (la communauté) le plus pertinent qui peut l'aider dans la résolution de son problème. Donc, nous avons besoin d'une nouvelle métrique qui prend en considération deux types de poids : les poids des arcs et les poids des noeuds. Le poids d'un arc représente le degré de similarité entre les sommets reliés par cet arc. Quant au poids d'un noeud, il représente le degré de similarité entre ce noeud et le noeud représentant le membre demandant l'aide même si ces deux noeuds ne sont pas reliés par un arc. Nous avons présenté dans cette section la première technique que nous allons utiliser pour la recommandation des nouvelles collaborations. Dans la section suivante, nous allons présenter la seconde technique.

3.1.2. L'optimisation par essaim particulière (OEP)

Comme l'optimisation de la modularité est typiquement une optimisation NP-difficile qui nécessite un temps d'exécution exponentiel pour trouver la solution optimale (Leskovec *et al.*, 2010), plusieurs travaux se sont basés sur l'utilisation des (méta) heuristiques pour trouver rapidement une solution réalisable et approchée, pas nécessairement la solution optimale exacte, pour un problème d'optimisation difficile. Nous citons par exemple l'utilisation du recuit simulé dans (Guimera *et al.*, 2004), l'algorithme glouton (Newman et Girvan, 2004) et (Blondel *et al.*, 2008), les colonies de fourmis dans (Sadi *et al.*, 2009) et (He *et al.*, 2011) et l'algorithme génétique dans (Pizzuti, 2008) et (Mazur *et al.*, 2010). L'enjeu de toutes ces techniques d'optimisation est de minimiser le temps de convergence de façon qu'elles fournissent rapidement une solution. Cependant, des études empiriques, réalisées dans (Parsopoulos et Vrahatis, 2010) et (El dor, 2012), ont confirmé que, l'optimisation par essaim particulière (OEP), qui consiste en une méta-heuristique pour l'optimisation NP-difficile (Clerc et Siarry, 2004), est plus rapide que les techniques citées précédemment et elle donne de résultats meilleurs. Dans ce qui suit, nous allons expliquer le principe et l'algorithme d'OEP.

L'optimisation par essaim particulière (OEP) est introduite dans (Eberhart et Kennedy, 1995) comme une technique évolutive de computation inspirée de la simulation du comportement social des oiseaux ou de poisson. La population dans l'OEP est appelée un essaim où les individus, les particules, sont candidats à des solutions du pro-

blème d'optimisation dans l'espace de recherche multidimensionnel (D dimensions). Chaque particule propose une solution et évalue sa proposition en calculant la fonction objective (fitness) qui doit être optimisée.

L'algorithme d'OEP est itératif et le nombre d'itérations n'est pas fixe, il peut être obtenu de façon empirique. Pour chaque itération t, chaque particule i est caractérisée par sa position $\vec{x}_i(t)$ et sa vitesse $\vec{v}_i(t)$ qui sont mises à jour de manière synchrone dans chaque itération de l'algorithme. Une particule ajuste sa vitesse en fonction de sa propre expérience de vol (la solution optimale dans son historique) et l'expérience de vol d'autres particules (la solution optimale pour l'essaim entier) de telle manière qu'elle accélère à des positions qui permettent la convergence rapide vers la solution optimale. Par conséquent, il existe deux types de position vers laquelle une particule est accélérée. La première est la meilleure position personnelle d'une particule obtenue jusqu'à l'itération courante, appelée $Pbest_i$. L'autre est la meilleure position globale obtenue à l'itération courante par toutes les particules, appelée Gbest. La position $\vec{x}_i(t+1)$ et la vitesse $\vec{v}_i(t+1)$ à l'itération t+1 pour une particule i peuvent être calculées en utilisant les formules suivantes :

$$\vec{x}_i(t) = \vec{x}_i(t-1) + \vec{v}_i(t) \quad (2)$$

$$\vec{v}_i(t+1) = \omega \vec{v}_i(t) + c_1 r_1 [x\vec{p}_i(t) - \vec{x}_i(t)] + c_2 r_2 [g(\vec{t}) - \vec{x}_i(t)] \quad (3)$$

Les paramètres ω , c_1 , et c_2 sont des coefficients constants fixés par l'utilisateur, r_1 et r_2 sont des nombres aléatoires tirés à chaque itération, $g(\vec{t})$ est la meilleure solution trouvée jusqu'à l'instant t et $x\vec{p}_i(t)$ est la meilleure solution trouvée par le particule i. C'est le vecteur vitesse qui dirige le processus de recherche et reflète la "sociabilité" des particules.

L'algorithme d'OEP peut se résumer dans le pseudo-code suivant :

- 1) Les paramètres de l'algorithme sont initialisés,
- 2) Les positions et les vitesses initiales des particules sont générées en utilisant des valeurs aléatoires,
- 3) Chaque particule évalue sa position (sa solution) en calculant la fonction objective,
- 4) La position et la vitesse de chaque particule sont mises à jour en utilisant(2) et(3),
- 5) Si le nombre maximal d'itérations est atteint, les solutions finales sont données et Gbest est considérée comme la solution optimale détectée. Sinon, l'algorithme boucle à 3 et ainsi de suite.

3.2. Proposition d'un nouvel algorithme combinant la détection de communautés et l'optimisation par essaim particulière

Nous allons présenter dans cette section notre contribution qui concerne la proposition d'un nouvel algorithme de recommandation des collaborations basé sur la combinaison de la détection de communautés et l'OEP. Cet algorithme cherche à optimiser une nouvelle fonction bi-objective qui tient compte à la fois les interactions sociales (les poids des arcs) et l'homophilie des membres (les poids des noeuds).

3.2.1. Définition d'une nouvelle métrique bi-objective

Rappelons que notre objectif est de déterminer le groupe le plus pertinent qui peut participer dans la résolution de problèmes. Pour ce faire et étant donné un membre qui a un problème et qui cherche des collaborateurs pertinents, nous allons utiliser la détection de communautés pour identifier la communauté à laquelle appartient ce membre et ensuite lui recommander les différents membres de cette communauté.

Dans ce contexte, le graphe qui représente l'organisation, est pondéré par des poids des arcs et des noeuds. Le poids d'un arc représente le degré de similarité entre les sommets déjà reliés par cet arc. Quant au poids d'un noeud, il représente le degré de similarité entre ce noeud et le noeud représentant le membre demandant l'aide même si ces deux noeuds ne sont pas reliés par un arc.

Maintenant, pour prendre en considération ces deux types de poids dans une même métrique, nous proposons une nouvelle fonction bi-objective qui doit être optimisée pour obtenir un partitionnement de bonne qualité. Cette fonction se base sur deux composantes qui doivent être maximisées simultanément :

La première composante concerne la fréquence des interactions sociales entre les individus en se basant sur l'hypothèse que les personnes qui socialisent fréquemment (ont des interactions entre elles) sont plus susceptibles d'être des collaborateurs efficaces. Elle s'appuie sur la qualité structurelle (en termes de liens sociaux) pour évaluer le partitionnement. Pour cela, nous proposons de réutiliser la modularité de Newman pour les graphes pondérés.

La deuxième composante concerne la fréquence d'homophilie entre l'individu cherchant les collaborateurs et les autres individus en se basant sur l'hypothèse que les personnes qui ont des caractéristiques en commun (par exemple les centres d'intérêts ou les préférences) sont plus susceptibles d'être des collaborateurs efficaces. Elle s'appuie sur la qualité basée attributs (en termes de similarité des attributs) pour évaluer le partitionnement. Pour cela, nous proposons de réutiliser la notion de l'inertie (Lebart *et al.*, 1982).

L'inertie est une métrique qui permet de mesurer la dispersion d'un nuage pondéré (un ensemble de noeuds où chaque noeud possède un poids). Si nous considérons I_{total} l'inertie totale d'un nuage alors selon la formule de Huygens (4), I_{total} est composé de deux quantités : l'inertie intra-classes I_{intra} et l'inertie inter-classes I_{inter} .

$$I_{total} = I_{inter} + I_{intra} \quad (4)$$

L'inertie intra-classes I_{intra} , donnée par la formule (5), permet de mesurer le degré d'homogénéité entre les noeuds appartenant à la même classe. Elle calcule leurs distances par rapport au centre de gravité de la classe (le point représentant la moyenne de la classe). Dans la formule (3), p_i représente le poids de la classe C_i (qui est la somme des poids de tous les noeuds appartenant à la même classe C_i). Ensuite, I_i représente l'inertie de chaque classe C_i (qui est la somme des carrés des distances entre chaque sommet dans la classe C_i et le centre de gravité g_i de C_i divisées par la cardinalité de C_i). Enfin, $d(x,y)$ est la distance entre x et y (qui est généralement de la distance euclidienne).

$$I_{intra} = \sum_i p_i I_i = \sum_i p_i \sum_{x_j \in C_i} \frac{d^2(x_j, g_i)}{|C_i|} \quad (5)$$

L'inertie inter-classes I_{inter} , donnée par la formule (6), permet de mesurer le degré d'hétérogénéité entre les classes. Elle calcule les distances entre les centres de gravité g_i de chaque classe C_i de la partition par rapport au centre de gravité global g de tout le nuage (g est le poids du nuage divisé par sa cardinalité).

$$I_{inter} = \sum_i p_i d^2(g_i, g) \quad (6)$$

Ensuite, une inertie inter-classes forte reflète l'hétérogénéité entre les classes et une inertie inter-classes faible reflète l'homogénéité au sein des classes. Pour obtenir une bonne partition du nuage, nous considérons le rapport I entre l'inertie inter-classes I_{inter} et l'inertie totale I_{total} qui doit être maximisé.

$$I = \frac{I_{inter}}{I_{total}} \quad (7)$$

Par conséquent, la nouvelle métrique Q que nous allons utiliser pour évaluer la qualité de la détection de communautés, donnée par (8), représente une fonction bi-objective. En effet, son objectif est de maximiser simultanément la modularité et l'inertie inter-classes de façon à obtenir des communautés denses et homogènes en termes d'interactions sociales fortes et en termes d'homophilie importante. Cette nouvelle métrique est basée sur une combinaison pondérée de M et I et doit être maximisée pour obtenir une bonne division du réseau (α est un facteur de pondération dynamique où $0 < \alpha < 1$, si nous voulons obtenir l'équité entre la modularité et la proportion de l'inertie inter-classes, nous pouvons mettre α à 0,5).

$$Q = \alpha M + (1 - \alpha) I = \alpha \frac{1}{2m} \sum_{i,j \in V} [A_{ij} - \frac{k_i k_j}{2m}] \delta(C_i, C_j) + (1 - \alpha) \frac{I_{inter}}{I_{total}} \quad (8)$$

Dans ce qui suit, nous allons montrer comment cette métrique va être exploitée et maximisée pour assister la recommandation des collaborations.

3.2.2. L'algorithme d'optimisation proposé

Comme nous l'avons déjà mentionné et justifié, nous proposons d'utiliser l'optimisation par essaim particulière pour l'optimisation de la nouvelle fonction bi-objective proposée. Ainsi, l'algorithme de détection de communautés que nous proposons est un algorithme d'optimisation qui se base sur l'OEP.

L'objectif de cet algorithme consiste à partitionner l'organisation en deux clusters disjoints, l'un représente la communauté qui englobe le membre cherchant les collaborateurs et les membres qui lui sont similaires et l'autre représente le groupe qui englobe le reste de l'organisation. Pour déterminer les membres similaires (à recommander), nous avons besoin de : (1) identifier la fonction de similarité entre les membres pour calculer leurs degrés de similarité et (2) identifier un seuil au-dessus duquel un membre sera considéré comme similaire au membre demandeur.

En ce qui concerne la fonction de similarité, nous allons modéliser l'organisation et ses membres par un patron conceptuel ontologique (PCO) que nous avons proposé dans (Ben Yahia *et al.*, 2013). Ce PCO permet de représenter les caractéristiques de l'organisation (ses objectifs, sa structure et ses groupes) ainsi que les profils de ses membres. Donc, la fonction de similarité représente la similarité sémantique entre les concepts des profils des membres.

En ce qui concerne le seuil, nous allons utiliser l'OEP pour déterminer la meilleure valeur qui permet de maximiser la fonction bi-objective. Donc, pour le présent problème à optimiser, dans chaque itération, chaque particule propose un seuil puis évalue la fonction en se basant sur ce seuil.

Pour l'initialisation des paramètres de l'algorithme proposé, nous utilisons les valeurs suivantes :

- La population initiale est composée de 10 particules, le nombre d'itérations est de 1000, $c_1 = 2,8$ et $c_2 = 1,3$ en justifiant ces choix sur l'étude empirique réalisée dans (Carlisle et Dozier, 2001) qui confirme que ces valeurs sont les meilleures.

- Les positions et les vitesses initiales des particules sont générées en utilisant des valeurs aléatoires dans (W_{nmin} , W_{nmax}) où W_{nmin} et W_{nmax} sont respectivement les valeurs minimales et maximales des poids des noeuds.

- La valeur de $Pbest_i$ est initialisée à la position initiale de la particule i et $Gbest$ est initialisée à la meilleure valeur de tous les $Pbest_i$.

L'algorithme proposé peut se résumer comme suit. Après l'initialisation des paramètres :

- 1) Chaque particule prend une position.
- 2) Ensuite, elle évalue cette position en calculant la modularité M et la proportion d'inertie inter-classes I et elle déduit la valeur de Q .
- 3) Enfin, elle met à jour la valeur de $Pbest_i$ (si elle a amélioré sa position sinon elle garde la valeur ancienne) et même chose pour $Gbest$ (si une particule a donné une valeur $Pbest_i$ mieux que $Gbest$ donc celle-ci va être écrasée par $Pbest_i$ sinon $Gbest$ garde la valeur ancienne).

4) Si le nombre maximal d'itérations est atteint, Gbest représente la meilleure solution obtenue par tout l'essaim. Sinon, chaque particule va refaire le même travail.

3.3. Présentation de l'approche de recommandation proposée

L'approche de recommandation des collaborations, illustrée par la figure 1, que nous proposons se base sur trois phases :

1) La première phase concerne la modélisation de l'organisation. En fait, cette dernière est modélisée graphiquement par un graphe ou les noeuds représentent les membres et les arcs représentent les interactions sociales entre ces membres. Elle est modélisée conceptuellement par l'instanciation du PCO organisationnel proposé dans (Ben Yahia *et al.*, 2013).

2) La deuxième phase concerne l'application de l'algorithme proposé pour détecter la communauté qui peut aider un membre dans la recherche des collaborateurs pertinents. Cet algorithme calcule les poids des arcs et des noeuds et ensuite optimise la qualité bi-objective pour détecter la bonne communauté.

3) La troisième phase recommande les membres de la communauté détectée (considérés comme des collaborateurs pertinents et efficaces) au membre demandant l'aide.

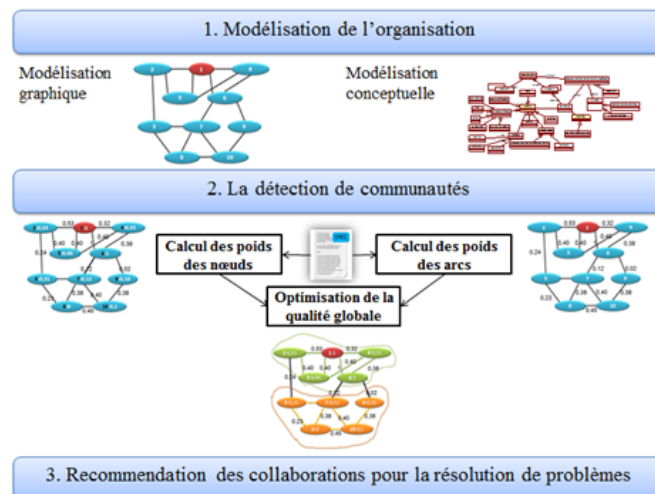


Figure 1. L'approche proposée de recommandation des collaborations.

4. Expérimentation de l'approche proposée sur un graphe de terrain : un laboratoire de recherche

Afin d'évaluer la performance de l'algorithme de recommandation proposé, nous proposons de comparer les résultats de cet algorithme par les résultats donnés par l'application de l'algorithme de Louvain (Blondel *et al.*, 2008) sur un graphe de terrain qui représente le laboratoire RIADI¹. En effet, nous avons testé les deux approches en considérant le laboratoire RIADI comme une organisation où les membres (les noeuds) représentent les chercheurs et les interactions (les arcs) représentent les relations de co-publication et de co-encadrement. Le graphe ainsi que le PCO organisationnel ont été construit progressivement (en ajoutant à chaque fois des nouveaux chercheurs et leurs copublications) en utilisant les fiches-chercheurs de différents membres. Le tableau 1 expose les différents résultats obtenus lors de l'application de notre approche et ceux obtenus lors de l'application de l'approche de Louvain (un bloc de chercheurs est à chaque fois ajouté jusqu'à l'obtention d'un graphe englobant 100 noeuds). Dans ce tableau, nous comparons les résultats obtenus par les deux approches en nous basant sur la qualité de détection de communautés et le temps d'exécution. Pour notre approche, nous donnons séparément les résultats liés à la modularité et ceux liés à la proportion d'inertie inter-classes. Ensuite, nous supposons le facteur de pondération α est égal à 0,5 (nous obtenons ainsi l'équité entre les interactions sociales et l'homophilie des participants) et nous donnons les résultats liés à la qualité globale.

Pour mieux visualiser ces mesures, nous avons exploité quelques valeurs pour tracer les deux courbes suivantes.

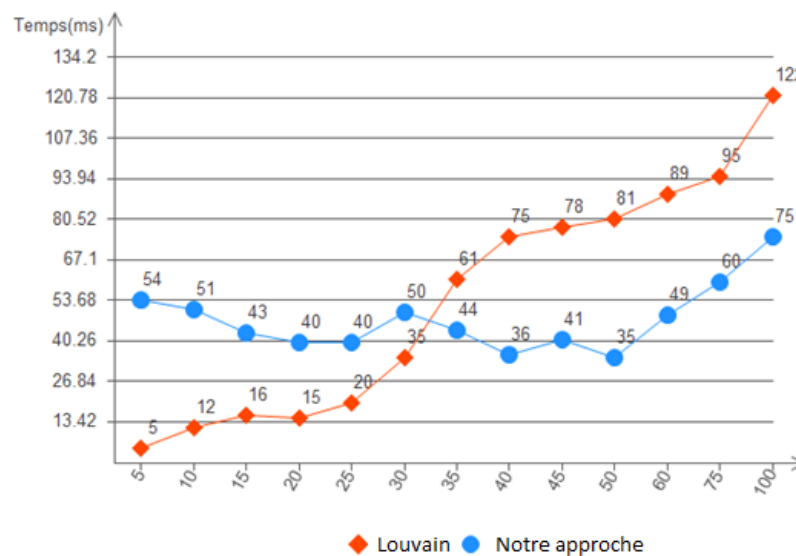
La première courbe donnée par la figure 2 représente une comparaison du temps d'exécution des deux algorithmes. D'après cette courbe, nous remarquons que l'algorithme de Louvain est plus rapide uniquement pour des graphes de petites tailles. Cependant l'algorithme utilisé dans notre application a un temps d'exécution qui est presque stable. Cela signifie que ce dernier permet de converger indépendamment de la taille du réseau exploité, ce qui n'est pas le cas pour l'algorithme de Louvain.

La deuxième courbe donnée par la figure 3 représente une comparaison de la modularité de la communauté détectée. D'après cette courbe, nous pouvons affirmer que la recommandation est plus performante et plus précise dans le cas de l'application du nouvel algorithme que dans le cas où l'algorithme de Louvain est appliqué pour détecter la communauté la plus proche d'un noeud donné. De plus l'algorithme que nous avons choisi tient compte de la similarité entre les attributs des noeuds, ce qui n'est pas le cas pour l'algorithme de Louvain qui considère uniquement les liens entre les noeuds.

1. www.riadi.rnu.tn

Table 1. Tableau des mesures.

Réseau social		Algorithme de Louvain		Algorithme de recommandation proposé			
Nombre de noeuds	Nombre de liens	Temps d'exécution (ms)	Modularité de la communauté détectée	Nombre de communautés détectées	Temps d'exécution (ms)	Modularité de la communauté détectée	Nombre de communautés détectées
5	7	5	0.17	1	54	0.29	2
10	16	12	0.24	2	51	0.24	2
15	23	16	0.23	3	43	0.23	2
20	30	15	0.20	4	40	0.20	2
25	38	20	0.16	5	40	0.25	2
30	42	35	0.26	2	50	0.26	2
35	47	61	0.14	7	44	0.25	2
40	53	75	0.06	6	36	0.24	2
45	59	78	0.12	8	41	0.25	2
50	63	81	0.11	9	35	0.24	2
60	73	89	0.11	10	49	0.23	2
75	92	95	0.12	12	60	0.21	2
100	114	122	0.14	16	75	0.19	2

**Figure 2.** Comparaison du temps d'exécution des deux algorithmes.

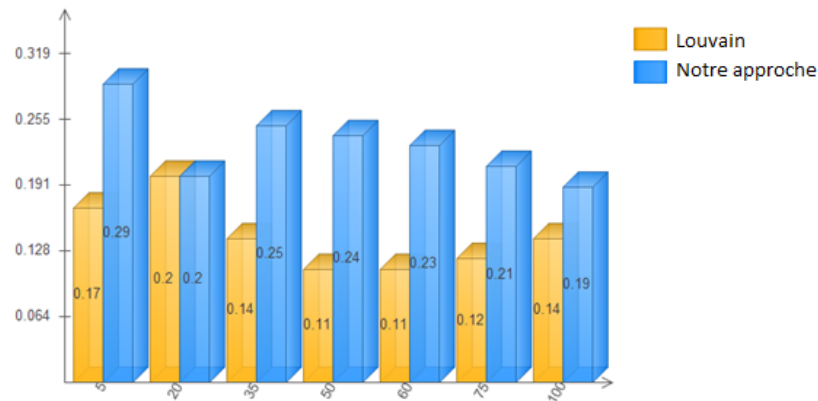


Figure 3. Comparaison de la modularité de la communauté détectée par les deux algorithmes.

5. Conclusion et perspectives

Nous avons proposé dans cet article une approche hybride de recommandation des collaborations en combinant la détection de communautés (une technique de partitionnement de graphe en considérant l'organisation comme un réseau social) et l'optimisation par essaim particulaire (une méta heuristique). Cette approche se base sur une fonction bi-objective qui considère la modularité (les interactions sociales entre les individus) et l'homophilie (l'homogénéité des individus) pour évaluer la qualité de la communauté détectée. Ensuite, nous avons eu recours à l'optimisation par essaim particulaire pour maximiser cette qualité bi-objective. A ce stade de réflexion, nous proposons les deux perspectives suivantes : i) vérifier la scalabilité de l'approche proposée pour des graphes de terrains de taille importante, ii) étudier la résilience des communautés en se basant sur leurs productivités et leurs rendements.

Bibliographie

- Ackerman M. S., Dachtera J., Pipek V., Wulf V., « Sharing Knowledge and Expertise : The CSCW View of Knowledge Management », *J Computer Supported Cooperative Work*, vol. 22(4-6), p. 531-573, 2013.
- Ben Yahia N., Bellamine N., Ben Ghezala H., « Vers une architecture multicouche d'ontologies dédiée à la résolution mixte de problèmes », In : *Int Proc of Extraction et gestion des connaissances, Toulouse*. 263-268, 2013.
- Blondel V., Guillaume J. L., Lambiotte R., Lefebvre E., « Fast unfolding of communities in large networks », *J Statistical Mechanics : Theory and Experiment*, vol. 10, p. 10008-10020, 2008.

- Cai X., Bain M., Krzywicki A. and Wobcke W., Kim Y., Compton P., Mahidadia A., « Collaborative Filtering for People to People Recommendation in Social Networks », *In : IntProc of Australasian Conference on Artificial Intelligence, South Australia*. 476-485, 2010.
- Carlisle G., Dozier G., « An Off-The-Shelf PSO », *In : IntProc of Particle Swarm Optimization Workshop, USA*. 1-6, 2001.
- Clerc M., Siarry P., « Une nouvelle méta-heuristique pour l'optimisation difficile : la méthode des essais particuliers », *J l'enseignement des sciences et technologies de l'information et des systèmes*, 2004.
- Dong Y., Tang J., Wu S., Tian J., Chawla N., Rao J., Cao H., « Link Prediction and Recommendation across Heterogeneous Social Networks », *In : IntProc of IEEE 12th International Conference on Data Mining, Belgium*. 181-190, 2012.
- Eberhart R. C., Kennedy J., « New optimizer using particle swarm theory », *In Proceedings of the 6th International Symposium on Micro Machine and Human Science, Nagoya, Japan*. 39-43, 1995.
- El dor A., Perfectionnement des algorithmes d'Optimisation par Essaim Particulaire. Applications en segmentation d'images et en électronique, Thèse de doctorat, Paris-Est University, 2012.
- Guimera R., Sales-Pardo M., Amaral L., « Modularity from fluctuations in random graphs and complex networks », *Physical Review Ep*. 70, 2004.
- He D., Liu J., Liu D., Jin D., Jia Z., « Ant colony optimization for community detection in large-scale complex networks », *In : IntProc of International Conference on Natural Computation (ICNC), Chinap*. 1151-1155, 2011.
- Kautz H., Selman B., Shah M., « Referral Web : Combining Social Networks and Collaborative Filtering », *Magazine Communications of the ACM*, vol. 40(3), p. 63-65, 1997.
- Lebart L., Maurineau A., Piron M., *Traitement des données statistiques*, Dunod Paris, 1982.
- Leskovec J., Lang K., Mahoney M., « Empirical Comparison of Algorithms for Network Community Detection », *In : IntProc of International Conference on World Wide Web, USA*. 631-640, 2010.
- Liben-Nowell D., Kleinberg J., « The link prediction problem for social networks », *In : IntProc of Conference on Information and Knowledge Management, USA*. 556-559, 2003.
- Lu L., Zhou T., « Link prediction in complex networks : a survey », *J Physica A : Statistical Mechanics and its Applications*, vol. 390(6), p. 1150-1170, 2011.
- Matthews T., Whittaker S., Moran T., Yuen S., Judge T., « Productive interrelationships between collaborative groups ease the challenges of dynamic and multi-teaming », *J Computer Supported Cooperative Work*, vol. 21(4-5), p. 371-396, 2012.
- Mazur P., Zmarzkowski K., Orkowski A J., « Genetic Algorithms Approach to Community Detection », *J Act a Physica Polonica A*, vol. 117(4), p. 703-705, 2010.
- McPherson M., Smith-Lovin L., Cook J., « Birds of a feather : Homophily in social networks », *Annual Review of Sociology*, vol. 27(1), p. 415-444, 2001.
- Nardi B., Whittaker S., Isaacs E., Creech M., Johnson J., Hainsworth J., « Integrating Communication and Information Through ContactMap », *Communications of the ACM*, vol. 45(4), p. 89-95, 2002.
- Newman M., Girvan M., « Finding and evaluating community structure in networks », *J Physical Review*, vol. 69(2), p. 26-113, 2004.

- Ogata H., Yano Y., Furugori N., Jin Q., « Computer Supported Social Networking For Augmenting Cooperation », *J Computer Supported Cooperative Work*, vol. 10(2), p. 189-209, 2001.
- Parsopoulos K., Vrahatis M., « Particle Swarm Optimization and Intelligence : Advances and Applications », *Information Science Reference (IGI Global), USA*, 2010.
- Pizzuti C., « Community detection in social networks with genetic algorithms », *In : Int Proc of 10th annual conference on Genetic and evolutionary computation, ACM New York, USA*, p. 1137-1138, 2008.
- Sadi S., Etaner-Uyar S., Gunduz S., « Community detection using ant colony optimization techniques », *In : IntProc of 15th International Conference on Soft Computing*, p. 206-213, 2009.
- Smith M. A., Cadiz J., Burkharter B., « Conversation Trees and Threaded Chats », *In : IntProc of ACM Conference on Computer-Supported Cooperative Work ,Philadelphia*, vol. , p. 97-105, 2000.
- Wenger E., Snyder W., « Communities of practice : the organizational frontier », *J Harvard Business Review*, vol. 78(1), p. 139-146, 2000.

Une approche de recommandation proactive dans un environnement mobile

Imen Akermi^{1,2} — Mohand Boughanem¹ — Rim Faiz²

¹ *Institut de Recherche en Informatique de Toulouse, SIG
118, route de Narbonne, F-31062 Toulouse Cedex 9, France
{imen.akermi,mohand.boughanem}@irit.fr*

² *Laboratoire de Recherche Opérationnelle, de Décision et de Contrôle de Processus.
IHEC, Université de Carthage, 2016 Carthage Présidence, Tunisie
rim.faiz@ihec.rnu.tn*

RÉSUMÉ. Les systèmes de recommandation contextuelle visent à combiner un ensemble de technologies et de connaissances sur le contexte de l'utilisateur pour lui fournir une information pertinente au moment où il en a le plus besoin, c'est ce qu'on appelle la recommandation proactive. Dans cet article nous proposons une approche de recommandation contextuelle et proactive dans un environnement mobile qui apprend implicitement les préférences de l'utilisateur. Nous avons évalué notre approche dans le cadre de la tâche "Contextual Suggestion Track" de TREC 2014. Les résultats que nous avons obtenus sont prometteurs.

ABSTRACT. The Context Aware Recommender Systems aim to combine a set of technologies and knowledge about the user context in order to deliver the most appropriate information to the user need at just the right time. It is called Proactive Recommendation. In this paper, we propose an approach for context-aware proactive recommendation in a mobile environment that aims at providing users with relevant information tailored to their own needs and preferences. The user preferences are implicitly identified and learned. We evaluated our approach through the 2014 TREC Task "Contextual Suggestion Track". The results that we obtained are promising.

MOTS-CLÉS : modélisation du contexte, système de recommandation contextuelle, recommandation proactive.

KEYWORDS: context modeling, context-aware recommender systems, user modeling, proactive recommendation.

1. Introduction

L'accès aux informations pertinentes, adaptées aux besoins et au profil de l'utilisateur est un enjeu majeur dans le cadre actuel caractérisé par une prolifération massive des ressources d'information hétérogènes. Le développement d'appareils mobiles équipés de plusieurs fonctionnalités telles que la caméra, le WIFI, la géo-localisation et bien d'autres permettent aux systèmes mobiles actuels de recherche d'information d'être hautement contextualisés, pouvant ainsi fournir à l'utilisateur des informations pertinentes au bon moment quand il en a le plus besoin, sans attendre qu'il établisse une interaction avec son appareil. Un système proactif peut prendre la forme d'un guide touristique personnalisé qui se base sur la localisation et les préférences de l'utilisateur pour suggérer à ce dernier des endroits intéressants sans qu'il fournisse sa préférence ou une requête explicite. Cela permettra de réduire les efforts, le temps et l'interaction de l'utilisateur avec son appareil mobile et de présenter les informations pertinentes au bon moment et au bon endroit.

L'efficacité de ce type de système requiert une bonne modélisation des dimensions du contexte et plus particulièrement la modélisation du profil de l'utilisateur. En effet, comme mentionné par Mizzaro et Vassena (2011), plusieurs dimensions de contexte, tels que la localisation, le temps, les activités des utilisateurs, leurs besoins, la lumière, le bruit, le mouvement, etc., doivent être pris en compte pour bien capturer tous les facteurs pouvant influencer la recommandation. La véracité et la multiplicité de ces facteurs rend les modèles de contexte trop complexes à implémenter. D'autre part, l'utilisation de peu de dimensions mène à des modèles de contexte incapables de comprendre l'environnement de l'utilisateur (Mizzaro et Vassena, 2011).

Dans cet article, nous proposons une approche de recommandation contextuelle et proactive dans un environnement mobile qui intègre les technologies mobiles et les dimensions spatio-temporelles pour la modélisation d'un profil situationnel de l'utilisateur. En effet, on vise à identifier la bonne information à recommander en fonction de la situation dans laquelle se trouve l'utilisateur.

Cet article est organisé comme suit. Nous présentons dans la section 2 un état de l'art se focalisant sur la recommandation contextuelle. La section 3 présente l'approche proposée. Dans la section 4, nous présentons et discutons les résultats des expérimentations de notre approche sur la collection TREC Contextual Suggestion Track 2014. Nous concluons et énonçons quelques perspectives en section 5.

2. Etat de l'art

Les systèmes de recommandation proactifs (SRP) comme décrit par Melguizo *et al.* (2007), récupèrent de grandes quantités de documents, décident quelle information disponible est probablement en rapport avec les besoins des utilisateurs, et suggèrent ces informations sans une demande explicite de l'utilisateur.

Récemment, plusieurs approches ont été développées pour soutenir la recommandation proactive. On distingue les approches qui se basent sur l'historique de comportement de l'utilisateur pour déterminer ses intérêts. Le comportement de l'utilisateur englobe l'historique de navigation Web/clics (Mizzaro et Vassena, 2011); les comportements antérieurs de visite de lieux pour les systèmes basés sur la localisation (Li *et al.*, 2012; Pu *et al.*, 2012) et l'historique des articles lus pour les systèmes de recommandation d'articles d'actualité (IJntema *et al.*, 2010; Arora et Shah, 2011; Athalye, 2013; Dumitrescu et Santini, 2012).

D'autres approches de recommandation traitent le contexte de l'utilisateur d'un point de vue centré sur l'activité. Les activités communes utilisées pour construire le profil de l'utilisateur dans ces systèmes pourraient prendre la forme de pages Web ou des documents ouverts (Budzik et Hammond, 2000; Finkelstein *et al.*, 2001; Rhodes et Maes, 2000; Prekop et Burnett, 2002; Dumais *et al.*, 2004; Karkali *et al.*, 2013); de conversation ou l'activité en cours, telles que les messages texte, les appels téléphoniques (Popescu-Belis *et al.*, 2011) et de l'activité de l'utilisateur à travers les réseaux sociaux tels que le contenu des tweets de l'utilisateur sur Twitter (Phelan *et al.*, 2011; De Francisci *et al.*, 2012; O'Banion *et al.*, 2012).

Toutes les approches mentionnées ci-dessus et beaucoup d'autres ont essayé de relever le défi d'inférer le contexte de l'utilisateur et ses préférences, au bon moment et sans son intervention. Cependant, la plupart de ces systèmes rencontrent certaines limites. En effet, diverses approches (Bedi et Agarwal, 2012; Liu *et al.*, 2010; Saranya et Sadhasivam, 2012) reposent sur l'historique de comportement passé ou actuel de l'utilisateur tel que son historique de navigation sur internet ou son historique de visite de lieux, pour déterminer ses intérêts. Cependant, on ne peut pas se référer seulement au comportement passé de l'utilisateur car il peut ne pas contenir suffisamment d'informations sur ses intérêts. Un utilisateur peut naviguer sur une page web par hasard ou visiter un endroit particulier sans pour autant porter un intérêt à cet endroit.

Egalement, de nombreuses approches de recommandation (Budzik et Hammond, 2000; Finkelstein *et al.*, 2001; Rhodes et Maes, 2000; Prekop et Burnett, 2002; Dumais *et al.*, 2004; Karkali *et al.*, 2013) traitent le contexte de l'utilisateur d'un point de vue centré sur l'activité qui peut prendre la forme d'une conversation en cours ou des documents ouverts. Néanmoins, on ne peut pas réduire le profil de l'utilisateur à ces activités. On peut ouvrir un document pour travailler dessus ou avoir une conversation sur un sujet sans pour autant vouloir recevoir des recommandations qui y sont liées. De plus, l'activité sociale d'un utilisateur fournie par les commentaires exprimés à travers ses comptes sociaux ne peut pas fournir suffisamment d'informations pour construire un profil équilibré, car elle pourrait être limitée, voire absente. Ce genre de profilage des utilisateurs contient beaucoup de bruit à gérer et pas assez d'informations pour établir un profil assez complet.

Il existe aussi d'autres limites communes entre les systèmes de recommandation contextuels qui doivent être abordées, telle que, la dépendance à un domaine. En

effet, de nombreux systèmes (Pu et al., 2012 ; Yeung and Yang, 2010 ; Quercia et al., 2010) dépendent d'un seul domaine (tourisme, cinéma, nouvelles ...) et ont des dimensions de contexte spécifiques à appliquer en fonction du domaine. Cependant, la plupart de ces systèmes reposent presque sur la même combinaison de dimensions qui comprend essentiellement les préférences de l'utilisateur, la localisation, et le temps avec une légère différence sur la façon d'aborder ces dimensions. Par conséquent, de nombreux services liés à des domaines différents peuvent profiter des mêmes informations de contexte sans pour autant encombrer le mobile de l'utilisateur. Notre approche traite ces limites et contribue aux travaux existants en impliquant et en combinant les caractéristiques suivantes:

- La non-dépendance à un domaine particulier: nous couvrons divers domaines dans le processus de recommandation.
- Un système proactif: nous proposons un système de recommandation contextuel et proactif qui peut aider l'utilisateur à faire face aux problèmes de la surcharge d'information en recommandant une information pertinente qui correspond à ses intérêts personnels au moment où il en a le plus besoin sans attendre que l'utilisateur pose une requête.

Dans notre approche, on vise à identifier l'information adéquate à recommander selon la situation de l'utilisateur.

3. Une approche de recommandation contextuelle et proactive dans un environnement mobile

L'idée principale que nous proposons est l'intégration des informations relatives à un utilisateur, sa tendance de navigation et les technologies mobiles (GPS, accéléromètre, etc.) afin de recommander de manière proactive des informations pertinentes à l'utilisateur sans avoir à attendre que ce dernier initie une interaction avec son appareil. Nous décrivons dans ce qui suit le modèle de contexte proposé ainsi que le processus de recommandation.

3.1. La modélisation du contexte

Nous considérons le contexte comme une représentation multidimensionnelle *Contexte* = (profil, localisation, temps, activité), composée de quatre dimensions :

- Le profil de l'utilisateur est défini par les préférences de l'utilisateur et ses données démographiques
- Localisation: La position de l'utilisateur définie par des coordonnées GPS
- Temps: défini par des étiquettes numériques ou temporelles (matin, soir, ...)
- L'activité de l'utilisateur: l'activité de l'utilisateur en cours

Les instanciations de ces dimensions à un moment donné t forment une situation S où la recommandation sera jugée comme nécessaire ou pas. Nous extrayons implicitement les informations relatives à chaque dimension sans l'intervention de l'utilisateur. Nous définissons dans les sections suivantes les différents éléments du processus de modélisation de contexte.

3.1.1. *Le profil de l'utilisateur*

Nous utilisons un modèle de profil classique (Daoud et al. 2009) défini par deux attributs, les données démographiques; des informations relatives à l'utilisateur telles que le nom, l'âge, etc. ; et les préférences de l'utilisateur par rapport à des centres d'intérêt spécifiques (C). Un centre d'intérêt est représenté par une catégorie et son poids, désignant l'intérêt que porte l'utilisateur à la catégorie en question. Les catégories sont prédéfinies en utilisant « The Open Directory Project Dmoz¹ ».

$profil = \{C_i, w_i\} ; i=1..n$

Une catégorie est représentée par des termes pondérés extraits à partir de son historique de navigation et les pages qu'il suit sur les réseaux sociaux :

$C_i = \{t_j^{(i)}, w_j^{(i)}\} ; j=1..m$

3.1.2. *Le temps*

Selon le dictionnaire Merriam-Webster², le temps est défini comme "*un continuum non-spatial qui est mesuré en termes d'événements qui se succèdent du passé par le biais du présent au futur*". Le temps peut aussi être exprimé comme la période mesurée ou mesurable au cours de laquelle une action, un processus ou une situation existe ou continue (Whitrow 1988). Selon la deuxième définition, plusieurs unités de temps sont définies, par exemple, heures, jours, mois et années.

La flexibilité de la conception du temps et de sa mesure implique différentes représentations. En effet, le temps peut être représenté comme une variable continue dont les valeurs déterminent les moments précis au cours desquels des items sont évalués par un utilisateur donné. Exemple : *un utilisateur A a donné un score à un objet O à un temps $t = 1$ juin 2010 à 18:05:00*

Une autre manière de modéliser le temps est d'identifier des valeurs catégorielles, pour des périodes de temps spécifiques. Par exemple, dans le domaine du tourisme, la variable "saison" peut être exprimée ainsi: *saison = {saison chaude, saison froide}*.

Le temps peut également être modélisé de façon hiérarchique qui permet de définir le degré de granularité de la dimension temporelle.

Exemple:

$Jour_semaine = \{lundi, mardi, \dots, Dimanche\} ; temps = \{matin, midi, \dots, nuit\}$

¹ <http://www.dmoz.org/docs/en/about.html>

² <http://www.merriam-webster.com/dictionary/time>

Nous adoptons le modèle hiérarchique. Dans notre approche, nous divisons une journée en intervalles de temps (time slots) d'une certaine longueur. C'est selon ces intervalles que nous choisissons entre autre le type d'information à recommander.

On représente le temps en 2 niveaux :

1. Temps de la journée (D_t): la journée est divisée en cinq périodes qui sont le matin, le midi, l'après-midi, le soir et la nuit. Ces périodes sont encadrées selon des intervalles de temps répartis sur 24 heures.

$$D_t = \{ \text{morning } [07:00, 12:00], \text{ midday } [12:00, 14:00], \text{ afternoon } [14:00, 18:00], \text{ evening } [18:00, 22:00], \text{ night } [22:00, 00:00] \cup [00:00, 07:00] \}$$

2. Jour de la semaine (D_w): on distingue deux classes qui sont les jours de travail (du lundi au vendredi) et le reste des jours (weekend, vacances et jours fériés)

3.1.3. *La localisation*

Le système de positionnement global (GPS) intégré ou installé dans le dispositif mobile permet de définir l'emplacement de l'utilisateur. Cet emplacement est défini en termes de latitude et de longitude. Ces coordonnées GPS ne sont pas les seules caractéristiques que nous pouvons considérer lors de la définition d'un emplacement. En effet, comme indiqué par Dobson (2005), il existe différentes façons de caractériser l'emplacement de l'utilisateur: *Position absolue* ; *Position Relative* (à côté de, ...) ; *Un nom de lieu* ; *Une classe nommée* qui représente le type du lieu, par exemple, musée, école, etc.

Dans notre approche, nous considérons la position absolue et la représentation de classe nommée pour caractériser la dimension spatiale de l'utilisateur. Nous définissons deux niveaux pour cette dimension utilisés dans le processus d'extraction de l'information à recommander:

- L'emplacement actuel: il s'agit de l'emplacement actuel de l'utilisateur à un moment donné.
- Les endroits liés à l'utilisateur: Ce sont les endroits liés à la vie quotidienne de l'utilisateur (travail, domicile, ...) et qui pourraient être déterminés en étudiant les endroits dans lesquels se trouve l'utilisateur pendant un certain laps de temps.

Le type de l'emplacement actuel peut être déterminé à l'aide de plusieurs outils tels que geonames³ ou le réseau social Foursquare qui attribuent une catégorie à une combinaison de coordonnées GPS donnés. Nous utilisons Foursquare pour déduire les informations de localisation: La catégorie du lieu et son nom.

3.1.4. *L'activité de l'utilisateur*

L'activité de l'utilisateur actuelle se résume en ce que l'utilisateur fait à un moment donné et peut être déterminée à travers son agenda ou l'application ouverte sur son téléphone portable. Cette dimension est importante dans notre approche et est utilisée pour évaluer si nous pouvons initier le processus de recommandation ou

³ <http://www.geonames.org/>

pas. En effet, nous ne pouvons pas recommander des informations, d'une manière imposante, alors que l'utilisateur est toujours à une réunion par exemple.

3.2. *Le processus de recommandation*

3.2.1. *Inférence de la catégorie d'information à recommander*

Nous considérons que le processus de recommandation implique un jeu de situations qui reflètent une catégorie d'intérêt spécifique caractérisée par les instanciations des dimensions spatio-temporelles et qui sont organisées dans une base de connaissances. Une situation est alors représentée par deux dimensions spécifiques: le temps, plus précisément le jour de la semaine (D_w), et le temps de la journée (D_t); et la catégorie du lieu (D_l). Une situation est formellement représentée comme $S = (D_l, D_w, D_t)$.

En effet nous estimons que le besoin de l'utilisateur en information change selon le moment de la journée. Une étude⁴ réalisée en 2013 révèle que 58% de personnes lisent avant toute chose leurs emails et les actualités le matin. Au déjeuner, les gens préfèrent naturellement trouver le restaurant le plus proche. Par exemple, la situation « déjeuner » est décrite par :

- D_l : n'importe quel lieu
- D_w : n'importe quel jour
- D_t : $t \in [12,14]$

La catégorie de l'information à recommander est inférée de la situation actuelle (Exemple : « Restaurant » pour la situation « déjeuner »).

3.2.2. *Extraction de l'information à recommander*

Afin de récupérer les informations à recommander, une requête q est formulée comme suit : $q = (\text{latitude}, \text{longitude}, \text{catégorie_d'intérêt})$

La requête est ensuite envoyée à un service basé sur la géo-localisation et choisi en fonction de la catégorie d'intérêt. Par exemple si la situation implique la recommandation d'actualités, on utilise le réseau social Feedly⁵ pour en extraire les actualités. Si la situation implique la recommandation d'un film, on interroge le réseau social Flixster⁶. L'intérêt de l'utilisation des réseaux sociaux pour l'extraction de l'information à recommander réside dans le fait que le filtrage collaboratif y est implicite ce qui assure non seulement la recommandation d'une information pertinente par rapport aux préférences de l'utilisateur mais aussi une information que d'autres utilisateurs recommande.

Le résultat de la requête envoyée est exprimé par un ensemble d'items I :

$$I = \{i_1, \dots, i_n\}.$$

⁴ <http://blog.marketo.com/2013/07/email-wanted-dead-or-alive-infographic.html>

⁵ <http://feedly.com/index.html#discover>

⁶ <https://fr.flixster.com/>

Si on prend l'exemple de la recommandation d'actualités, l'ensemble I sera représenté par un ensemble d'articles d'actualités :

$$I = \{A_1 (\text{titre}, \text{contenu}), \dots, A_k (\text{titre}, \text{contenu})\} ; k=1..p$$

Ensuite, nous filtrons à partir de I , les items convenant le mieux aux préférences de l'utilisateur en calculant un score d'utilité de l'item. Ce score combine un score de pertinence thématique de l'item vis-à-vis des préférences thématiques de l'utilisateur et un score géo-pertinence, évaluant l'accessibilité de l'item vis-à-vis de la position actuelle de l'utilisateur. Plus précisément, concernant la pertinence thématique, on considère qu'un item est représenté par un vecteur de mots pondérés (mots existants à partir des éléments textuels décrivant l'item).

La pertinence du contenu est calculée par le score de similarité cosin:

$$\text{Topic}_{\text{rel}}(VC_i, It) = \frac{\sum_{j=1}^n VC_i^{(j)} \times It_j}{\sqrt{\sum_{j=1}^n (VC_i^{(j)})^2} \times \sqrt{\sum_{j=1}^n (It_j)^2}} \quad (2)$$

Où:

VC_i : Le vecteur termes préférence relatif à la catégorie C_i

It : Le vecteur termes de l'item i

La géo-pertinence n'est utilisée que dans le cas où l'utilisateur doit se déplacer pour se rendre à l'item proposé. Elle est exprimée par un score qui mesure l'accessibilité à l'emplacement de l'item suggéré. L'accessibilité est mesurée comme la distance entre deux coordonnées GPS correspondant à l'emplacement actuel de l'utilisateur et l'emplacement de l'item proposé: (Point₁(lat1,long1) et Point₂(lat2,long2)) :

$$\text{accessibility} = R \times c \quad (3)$$

Où :

R : The earth radius=6,371Km

$c = 2 \times \text{atan2}(\sqrt{a}, \sqrt{1-a})$

$a = \sin^2((\text{lat}2 - \text{lat}1)/2) + \cos(\text{lat}1) \times \cos(\text{lat}2) \times \sin^2((\text{long}2 - \text{long}1)/2)$

La pertinence globale d'un élément à recommander est calculée comme suit :

$$\text{Rel} = \alpha \times \text{Topic}_{\text{rel}}(VC_i, It) + (1 - \alpha) \times \text{accessibility} \quad (4)$$

Les items sont classés selon leurs scores de pertinence globale.

4. Expérimentations

Nous avons évalué notre approche dans le cadre de la tâche « *Contextual Suggestion Track*⁷ » de TREC 2014. Nous présentons, dans cette section, une description générale de cette tâche, ensuite nous exposerons les résultats que nous avons obtenus.

⁷ <https://sites.google.com/site/trecontext/trec-2014>

4.1. Description de la tâche « Contextual Suggestion Track - CST » de TREC 2014

Cette tâche étudie les techniques de recherche pour les besoins d'information complexes qui dépendent fortement du contexte et des intérêts des utilisateurs. Le but de la tâche est de recommander à des utilisateurs se trouvant dans un contexte donné des lieux qui pourraient les intéresser. Un profil est associé à chaque utilisateur et indique la préférence de l'utilisateur par rapport à un ensemble de suggestions de lieux. Chaque suggestion comprend un titre, une description et une URL associée. Par exemple, une suggestion pourrait être une recommandation d'avoir une bière à la Dogfish Head Alehouse. Un profil est défini par un ensemble de préférences négatives ou positives par rapport à l'ensemble des suggestions de lieux. Chaque contexte correspond à un lieu géographique particulier (une ville). Par exemple, le contexte pourrait être Gaithersburg, Maryland. Le but de la tâche est de retourner pour chaque pair profil/contexte, une liste de lieux intéressants allant jusqu'à 50 lieux. Chaque suggestion doit être adaptée au profil (en fonction des préférences de l'utilisateur) et au contexte (selon l'emplacement).

4.1.1. Construction des profils

Les profils sont construits à partir de la liste de suggestions jugées par l'utilisateur. Chaque suggestion est évaluée selon deux notes, une note pour le titre et la description du lieu et une note pour son site web. Les notes sont données sur une échelle de cinq points sur la base du degré d'intérêt que peut porter cet utilisateur pour le lieu en question:

4 (très intéressé), 3 (intéressé), 2 (Neutre), 1 (Pas intéressé), 0 (Fortement pas intéressé), -1 (Site web ne s'est pas chargé ou pas de score)

Pour construire le profil thématique de l'utilisateur, on identifie, pour chaque suggestion, sa catégorie dans l'API Google Places⁸ en utilisant l'url qui est associée à la suggestion en question. A l'issue de cette étape, un profil sera alors modélisé comme un ensemble de catégories pondérés qui comportent les termes associés aux suggestions que l'utilisateur préfère : $profil = \{C_i(w_i)\} ; i=1..n$

Le calcul du poids assigné à chaque catégorie, pour chaque profil, prend en considération les scores que ce profil a attribués aux suggestions relevant de cette catégorie d'intérêt :

$$weight(category) = \frac{\sum_{s \in C} R_{td} + R_w}{N_p} \quad (5)$$

Où :

$\forall s \in C$: pour chaque suggestion s appartenant à cette catégorie d'intérêt C

R_{td} : le score donné au titre et à la description du lieu d'attraction

R_w : le score du site web du lieu d'attraction

N_p : le nombre de suggestions appartenant à cette catégorie d'intérêt

⁸ <https://developers.google.com/places/documentation/?hl=fr>

Les catégories sont représentées par des termes pondérés qu'on extrait à partir des descriptions des suggestions associées à chaque catégorie.

4.1.2. Sélection des suggestions

La tâche TREC définit le contexte en fonction des coordonnées GPS fournies dans cette tâche : $Contexte = \{latitude, longitude\}$. Pour notre part, on instancie le contexte en exploitant directement les données fournies par la tâche. Pour chaque contexte, on extrait les lieux intéressants en soumettant une requête aux géo-services Google Places et Foursquare⁹, formulé comme suit :

$q = \{latitude, longitude, categorie\}$

Suite à cette requête, on obtient un ensemble de lieux. Un lieu, est modélisé par un objet ayant des attributs spécifiques et qui appartient à une catégorie d'intérêt donné : $lieu = \{nom, url, description, accessibilité, categorie\}$

L'*accessibilité* : représente la distance qui sépare le lieu du contexte représentés par leurs coordonnées GPS ($P_1(lat1, long1)$ et $P_2(lat2, long2)$) et calculée par la formule (3) . Pour le total des 50 contextes donnés, on a pu extraire à peu près 15 000 lieux d'attraction ce qui fait en moyenne 300 lieux pour chaque contexte qui sont répertoriés sous des catégories.

4.1.3. Appariement Profils/Suggestions

Le processus de sélection des lieux intéressants pour chaque profil-contexte est résumé comme suit :

Pour chaque $p_i \in P$

 Pour chaque $c_j \in C$

- Calculer pour chaque suggestion $S \in C_{j_suggestions}$ la pertinence globale (selon la formule (4))
- Normaliser les scores de pertinence globale des suggestions entre 0 et 1
- Extraire les suggestions dont le score de pertinence globale est $\geq 0,5$ **Et** qui appartiennent aux catégories d'intérêt appréciées par le profil : $category_{weight} normalisé > 0.5$

Où:

P : l'ensemble des profils

C : l'ensemble de contextes

$C_{j_suggestions}$: L'ensemble des suggestions relatives au contexte C_j

⁹ <https://fr.foursquare.com/>

4.2. Résultats:

L'évaluation des résultats que nous avons obtenus comporte deux volets : l'évaluation de la pertinence géographique et la pertinence profil. Pour mesurer la géo-pertinence des lieux qu'on a suggérés, on extrait un échantillon V_g qui consiste en l'intersection de notre collection de lieux avec les lieux qui ont été évalués géographiquement dans la tâche TREC pour chaque contexte. Nous avons obtenu $|V_g| = 4802$ lieux en commun, tous contextes confondus. Parmi ces lieux, 4644 ont été évalués géographiquement pertinents. Ce qui implique une précision géographique totale égale à 0,97 et est calculée comme suit :

$$geo_relevance = \frac{Nb_Geo_Relevant_Venues}{|V_g|} \quad (6)$$

Le tableau 1 présente en détail les résultats.

Tableau 1. La géo-pertinence (Nb_Total est le nombre total des lieux suggérés pour chaque contexte ; Nb_GP est le nombre des lieux géographiquement pertinents ; Nb_NGP est le nombre des lieux identifiés comme géographiquement non pertinents)

Context ID	Nb_Total	Nb_GP	Nb_NGP	Context ID	Nb_Total	Nb_GP	Nb_NGP
101	120	118	2	127	86	80	6
102	74	71	3	128	118	117	1
103	116	113	3	129	106	106	0
104	103	100	3	130	35	29	6
105	58	56	2	131	132	128	4
106	116	114	2	132	115	111	4
107	137	132	5	133	49	46	3
108	80	77	3	134	213	210	3
109	113	109	4	135	82	75	7
110	80	76	4	136	60	56	4
111	63	61	2	137	48	47	1
112	24	22	2	138	155	152	3
113	171	165	6	139	35	33	2
114	68	66	2	140	94	91	3
115	103	92	11	141	69	67	2
116	24	23	1	142	114	111	3
118	79	73	6	143	52	43	9
119	111	107	4	144	203	202	1
120	129	127	2	146	99	96	3
121	111	108	3	147	105	103	2
122	114	111	3	148	26	26	0
123	180	180	0	149	204	201	3
124	198	195	3	150	69	67	2
125	80	73	7				

Pour mesurer la pertinence profil, il y avait deux alternatives à considérer. Une première alternative consiste à considérer pour chaque contexte et pour chaque run¹⁰, l'intersection de notre collection de lieux avec les lieux proposés par le run en question, cependant cette intersection donnait quasiment l'ensemble vide vu que nos méthodes d'extractions de lieux sont différentes. Nous avons donc opté pour une solution intermédiaire qui consiste à considérer l'intersection de notre collection de lieux avec l'union des lieux que chaque run a proposés, tous profils confondus. La cardinalité de cette intersection est $|V_p| = 889$ lieux. Ensuite nous avons mesuré pour chaque lieu le niveau d'intérêt qu'il a sollicité de la part des profils en se basant sur le nombre de profils qui ont évalué ce lieu comme pertinent par rapport au nombre d'évaluations totale du lieu en question. Par rapport à la collection obtenue, nous avons obtenu une précision moyenne de 0,56 calculée comme suit:

$$profile_relevance = \frac{\sum_{l=1}^{|V_p|} (nb_p_l / nb_p_tot_l)}{|V_p|} \quad (6)$$

Où :

nb_p_l : le nombre de profils qui ont évalué le lieu (l) comme pertinent

$nb_p_tot_l$: le nombre total de profils qui ont évalué le lieu (l)

Pour se situer approximativement par rapport aux autres participants, nous avons appliqué cette démarche d'évaluations sur chaque run. On a mesuré pour chaque run et pour chaque lieu que ce run propose, le degré d'intérêt sollicité par les profils en appliquant la formule (6). Le tableau 2 illustre les résultats obtenus pour chaque run.

Tableau 2. Résultats des autres participants

id	run	Nb to	Nb in	Préc	id	run	Nb to	Nb in	Préc
1	BJUTa	1495	352	0,24	20	simpleScore	1496	244	0,16
2	BJUTb	1495	327	0,22	21	simpleScoreImp	1496	254	0,17
3	BUPT 01	671	81	0,12	22	tueNet	1497	129	0,09
4	BUPT 02	704	78	0,11	23	tueRforest	1497	138	0,09
5	cat	1496	318	0,21	24	UDInfoCS2014	1495	201	0,13
6	choqrun	1465	196	0,13	25	UDInfoCS2014	1495	360	0,24
9	dixlticmu	1496	367	0,25	26	uogTrBunSumF	1495	272	0,18
10	gwl	1466	58	0,04	27	uogTrCsLtrF	1496	199	0,13
11	lda	1496	163	0,11	28	waterlooA	1497	260	0,17
14	RAMARUN	1497	260	0,17	29	waterlooB	1497	286	0,19
15	RUN1	1494	229	0,15	30	webis 1	1477	269	0,18
16	run_DwD	1496	181	0,12	31	webis 2	1474	230	0,16
17	run_FDwD	1496	262	0,18					

¹⁰ Un run est une application proposé par une équipe pour traiter la tâche « Contextual Suggestion » dans TREC 2014

D'après le tableau 2, on note que la précision de la pertinence profil de chaque run par rapport aux lieux qu'ils ont proposés ne dépasse pas **0,25**. Cela est expliqué par le fait, que pour chaque run, il y a un grand écart entre le nombre de profils qui ont jugé un lieu (appartenant à un seul contexte) comme pertinent et le nombre de jugements totaux. Généralement, les résultats que nous avons obtenus en appliquant notre approche sont prometteurs et montrent que l'utilisation de la classification par catégories des préférences d'un utilisateur implique une meilleure précision thématique par rapport à ses intérêts. Ces résultats montrent aussi que les choix des paramètres tels que le rayon que nous avons fixé pour l'extraction des lieux par rapport aux contextes sont efficaces. Cependant, cette évaluation n'a mis en examen qu'une partie de notre approche. En effet, la notion du temps n'est pas intégrée alors que cette dimension est considérée dans notre travail.

5. Conclusion

Le but fondamental des systèmes de recommandations contextuels consiste à combiner le contexte et l'environnement de l'utilisateur dans une même infrastructure pour mieux caractériser les informations dont l'utilisateur a besoin afin d'améliorer le processus de recommandation. Nous avons proposé une approche de recommandation contextuelle et proactive dans un environnement mobile qui permet la recommandation d'éléments pertinents qui répondent aux préférences de l'utilisateur, juste au bon moment sans attendre que l'utilisateur initie une requête. Plus précisément, notre contribution est divisée en deux principales parties: La modélisation d'un profil situationnel de l'utilisateur et la définition d'un cadre d'agrégation pour la combinaison des dimensions contextuelles et sociales. Nous planifions actuellement d'appliquer notre approche au sein de l'entreprise Tunav¹¹ qui développe des produits et des systèmes de navigation et de tracking par GPS, des applications cartographiques et des géo-portails qui offrent à l'utilisateur l'accès à une masse importante d'information géographique. Ces systèmes se diversifient dans différents domaines d'activité dont le transport routier (public, marchandise, taxi, etc.), les travaux, les services, la distribution, la santé, le tourisme. Nous avons comme objectif d'apporter le « just-in-time information » et la dimension contextuelle aux systèmes qui en ont besoin afin de fournir aux utilisateurs des produits personnalisés adaptés à leurs propres besoins et préférences.

Remerciements

Ce projet est effectué dans le cadre d'une thèse de doctorat MOBIDOC du programme PASRI¹² financé par l'UE et administré par l'ANPR¹³.

¹¹ <http://www.tunav.com/>

¹² <http://www.pasri.tn/> - Le Projet d'Appui au Système de Recherche et de l'Innovation (PASRI), un projet financé par l'Union européenne.

¹³ <http://www.anpr.tn/> - Agence Nationale de la Promotion de la Recherche Scientifique en Tunisie.

Références

- Arora A., Shah P. (2011). Personalized news prediction and recommendation. Unpublished doctoral dissertation, Stanford University.
- Athalye S. (2013). Recommendation system for news reader. Unpublished doctoral dissertation, San Jose State University.
- Bedi P., Agarwal S. (2012, Dec). A situation-aware proactive recommender system. In Hybrid intelligent systems (his), 2012 12th international conference on, p. 85-89.
- Budzik J., Hammond K. J. (2000). User interactions with everyday applications as context for just-in-time information access. In Proceedings of the 5th international conference on intelligent user interfaces, pp. 44–51. New York, NY, USA, ACM.
- Daoud M., Tamine L., Boughanem M. (2011). A personalized search using a semantic distance measure in a graph-based ranking model. Information Science, SAGE Publications, Vol. 37, No. 6, pp. 614-636.
- De Francisci Morales G., Gionis A., Lucchese C. (2012). From chatter to headlines: Harnessing the real-time web for personalized news recommendation. In Proceedings of the fifth acm international conference on web search and data mining, pp. 153–162. New York, NY, USA, ACM.
- DOBSON, S. (2005). Leveraging the subtleties of location. In Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies. ACM, pp. 189-193.
- Dumais S., Cutrell E., Sarin R., Horvitz E. (2004). Implicit queries (iq) for contextualized search. In Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval, pp. 594–594. New York, NY, USA, ACM.
- Dumitrescu D. A., Santini S. (2012). Improving novelty in streaming recommendation using a context model. In Cars 2012 : Acn recsys workshop on context-aware recommender systems.
- Finkelstein L., Evgenly G., Yossi M., Ehud R., Zach S., Gadi W. et al. (2001). Placing search in context: the concept revisited. In Proceedings of the Tenth International World Wide Web Conference.
- Gershman A., Wolfe T., Fink E., Carbonell J. (2011). News personalization using support vector machines. Proceedings of the Workshop on Enriching Information Retrieval.
- IJntema W., Goossen F., Frasinca F., Hogenboom F. (2010). Ontology-based news recommendation. In Proceedings of the 2010 edbt/icdt workshops, pp. 16:1–16:6. New York, NY, USA, ACM.
- Karkali M., Pontikis D., Vazirgiannis M. (2013). Match the news: A firefox extension for realtime news recommendation. In Proceedings of the 36th international acm sigir conference on research and development in information retrieval, pp. 1117–1118. New York, NY, USA, ACM.
- Lee H., Park S. J. (2007). Moners: A news recommender for the mobile web. Expert Systems with Applications, Vol. 32, No. 1, pp. 143 - 150.

- Li W., Eickhoff C., Vries A. P. de. (2012). Want a coffee?: Predicting users' trails. In Proceedings of the 35th international acm sigir conference on research and development in information retrieval, pp. 1171–1172. New York, NY, USA, ACM.
- Lin C., Xie R., Li L., Huang Z., Li T. (2012). Premise: Personalized news recommendation via implicit social experts. In Proceedings of the 21st acm international conference on information and knowledge management, pp. 1607–1611. New York, NY, USA, ACM.
- Liu J., Dolan P., Pedersen E. R. (2010). Personalized news recommendation based on click behavior. In Proceedings of the 15th international conference on intelligent user interfaces, pp. 31–40. New York, NY, USA, ACM.
- Melguizo M. C. P., Bogers T., Deshpande A., Boves L., Bosch A. van den. (2007). What a proactive recommendation system needs - relevance, non-intrusiveness, and a new longterm memory. In J. Cardoso, J. Cordeiro, J. Filipe (Eds.), *Iceis* (5), p. 86-91.
- Mizzaro, S., Vassena, L. (2011). A social approach to context-aware retrieval. *World Wide Web*, Vol. 14, No. 4, pp. 377–405.
- Nguyen T., Riedl J. (2013). Predicting users preference from tag relevance. In S. Carberry, S. Weibelzahl, A. Micarelli, G. Semeraro (Eds.), *User modeling, adaptation, and personalization*, Vol. 7899, p. 274-280. Springer Berlin Heidelberg.
- O'Banion S., Birnbaum L., Hammond K. (2012). Social media-driven news personalization. In Proceedings of the 4th acm recsys workshop on recommender systems and the social web, pp. 45–52. New York, NY, USA, ACM.
- Phelan O., McCarthy K., Bennett M., Smyth B. (2011). On using the real-time web for news recommendation and discovery. In Proceedings of the 20th international conference companion on world wide web, pp. 103–104. New York, NY, USA, ACM.
- Popescu-Belis A., Yazdani M., Nanchen A., Garner P. N. (2011). A speech-based just-in-time retrieval system using semantic search. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Systems demonstrations, pp. 80–85. Stroudsburg, PA, USA, Association for Computational Linguistics.
- Prekop P., Burnett M. (2003, July). Activities, context and ubiquitous computing. *Comput. Commun.*, Vol. 26, No. 11, pp. 1168–1176.
- Pu Q., Lbath A., He D. (2012). Location based recommendation for mobile users using language model and skyline query. *International Journal of Information Technology & Computer Science (IJITCS)*, Vol. 4, No. 10, pp. 19–28.
- Quercia D., Ellis J., Capra L. (2010). Nurturing social networks using mobile phones. *IEEE Pervasive Computing*, Vol.9, No. 3.
- Rhodes B., Maes P. (2000). Just-in-time information retrieval agents. *IBM Systems Journal*, Vol. 39, No. 3.4, pp. 685-704.
- Saranya.k.g, Sadhasivam G. S. (2012, November). Article: A personalized online news recommendation system. *International Journal of Computer Applications*, Vol. 57, No. 18, pp. 6-14.
- Tchunte, D., Canut, M.F., Jessel, N., Peninou, A., El Haddadi, A. (2010). Visualizing the Evolution of Users' Profiles from Online Social Networks. In Proceedings of the

International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp.370-374.

Yeung K. F., Yang Y. (2010, Sept). A proactive personalized mobile news recommendation system. In Developments in e-systems engineering (dese), 2010, p. 207-212.

Whitrow, G. (1989). Time in History: Views of Time from Prehistory to the Present Day. Oxford paperbacks, Oxford University Press.

Un modèle de langue pour l'estimation de la fraîcheur et la pertinence des documents Web

Mariam Bambia ^{*,} — Mohand Boughanem ^{**} — Rim Faiz ^{*}**

** LARODEC, Université de Tunis, Le Bardo, Tunisie
mariembambia@yahoo.fr*

Rim.Faiz@ihec.rnu.tn

*** IRIT, 118 Route de Narbonne, Toulouse, France
bougha@irit.fr*

RÉSUMÉ. Un système de recherche d'information performant doit satisfaire les différents types de besoins des utilisateurs visant une variété de catégories de requêtes. Ces catégories comprennent les requêtes sensibles au temps où le contenu récent est l'exigence principale de l'utilisateur. Cependant, l'utilisation des caractéristiques temporelles des documents pour mesurer leur fraîcheur reste une tâche difficile étant donné que ces caractéristiques ne sont pas représentées avec précision dans les documents récents. Dans cet article, nous proposons un modèle de langue qui estime la pertinence et la fraîcheur des documents vis-à-vis des requêtes sensibles au temps réel. Notre approche modélise la fraîcheur en exploitant des sources d'information fraîches, plus précisément des termes extraits à partir des tweets récents et thématiquement pertinents par rapport à la requête de l'utilisateur. Nos expérimentations montrent que l'extraction des termes frais à partir de Twitter améliore les résultats de recherche requis par l'utilisateur.

ABSTRACT. An effective information retrieval system must satisfy different users search intentions expecting a variety of query categories, comprising recency sensitive queries where fresh content is the major user's requirement. However, using temporal features of documents to measure their freshness remains a hard task since these features may not be accurately represented in recent documents. In this paper, we propose a language model which estimates the topical relevance and freshness of documents with respect to real-time sensitive queries. In order to improve freshness ranking, our approach models freshness by exploiting terms extracted from recently posted tweets topically relevant to each real-time sensitive queries. Our experiments show that there is a clear advantage of using twitter to extract fresh keywords.

MOTS-CLÉS : Requêtes sensibles au temps, Modèles de langue, Mots-clés frais.

KEYWORDS: Real-time sensitive queries, Language models, Fresh keywords.

1. Introduction

Un système de recherche d'information performant doit être en mesure de fournir des résultats pertinents pour différentes catégories de requêtes (par exemple, les requêtes non-sensibles au temps, les requêtes sensibles à la localisation, les requêtes sensibles au temps réel) lancées par différents utilisateurs ayant des attentes différentes. Cependant, il existe des catégories de requêtes dont le contenu et la pertinence des documents peuvent varier au fil du temps. Cela est particulièrement important pour les requêtes sensibles au temps, où les résultats requis ne doivent pas être uniquement pertinents au sujet de la requête, mais doivent contenir également des informations fraîches répondant aux exigences de l'utilisateur [(Dai *et al.*, 2011), (Moon *et al.*, 2010)]. En effet, la fraîcheur est classée en quatrième position parmi les neuf raisons principales motivant les internautes à préférer un document Web par rapport aux autres documents répondant à leur requête ¹. Par exemple, un utilisateur qui a soumis une requête sur une émission sportive cherche généralement le contenu télévisuel le plus frais, comme les matchs les plus récents et les dernières actualités du sport.

Toutefois, un événement qui a été diffusé la semaine dernière et qui a été classé comme étant pertinent pour une requête donnée, peut sans doute devenir moins important par rapport aux nouveaux événements frais et relatifs à cette thématique. Pour cette famille de requêtes, on peut proposer aux utilisateurs des options avancées pour filtrer explicitement les résultats de recherche à une période spécifique. Cette alternative présente deux limites principales : 1) plusieurs informations sont nécessairement spécifiées par l'utilisateur. Celui-ci est amené à investir l'effort au moyen de reformulation, bien que son besoin en contenu frais est immédiat, 2) seule une partie limitée du Web comme les flash infos peut répondre à des requêtes sensibles au temps réel, même si l'utilisateur spécifie explicitement la dimension temporelle, 3) extraire les caractéristiques temporelles des pages Web reste une tâche difficile, car elles peuvent ne pas être représentées correctement dans les documents frais. Par exemple, la requête "Débats présidentiels" où les documents pertinents sont sensibles aux jours, voire même aux heures, requiert la mesure de l'âge véritable des documents pertinents pour l'estimation de leur fraîcheur. L'exécution de ce type de requêtes complique la procédure de récupération du contenu le plus frais et l'estimation de cette notion de la fraîcheur des informations.

Le besoin en contenu frais est exprimé de plus en plus par les les internautes interagissant à travers les microblogs, les actualités et les réseaux sociaux. Par conséquent, les médias sociaux semblent être des sources d'informations intéressantes pour estimer la fraîcheur des documents Web (Damak *et al.*, 2012), (Ben Jabeur *et al.*, 2012), (Massoudi *et al.*, 2011), (Huo et J. Tsotras, 2010) and (Pasca, 2008). Dans cet article, nous estimons que la fraîcheur d'un document se réfère à un critère

1. <http://consumersunion.org/wp-content/uploads/2013/05/a-matter-of-trust.pdf>

qui dépend des occurrences des mots-clés frais dans ce document. Les mots-clés frais sont des termes qui font le Buzz et sont communément partagés dans le Web social et pertinents par rapport à la thématique spécifiée dans la requête. Dans notre cas, nous supposons que les mots-clés frais sont les termes qui apparaissent dans des ressources publiées récemment. Nous considérons que les microblogs, les dernières nouvelles et les tendances d'actualités sont les meilleures sources qui peuvent fournir ces mots-clés.

Les questions de recherche principales abordées dans cet article sont :

- Comment exploiter l'information sociale afin d'estimer la fraîcheur des documents afin de répondre aux requêtes sensibles au temps réel ?
- Comment agréger les deux critères de fraîcheur et de pertinence ?

Nous proposons un modèle de langue unifié pour l'estimation de la fraîcheur et la pertinence des documents Web. Cet article est structuré comme suit. La section 2 dresse un état de l'art sur les modèles proposés en Recherche d'Information (RI) pour l'estimation de la fraîcheur. La section 3 détaille le modèle de langue proposé. Section 4 présente la méthodologie expérimentale mise en œuvre et analyse les résultats obtenus. Finalement, nous concluons cet article et délimitons les perspectives.

2. Etat de l'art

Il existe plusieurs définitions de la fraîcheur des documents Web dans la littérature. Ces définitions diffèrent selon les mesures et les moyens utilisés pour caractériser les documents Web. La plupart des travaux antérieurs de classement de fraîcheur exploitent une variété de caractéristiques de la fraîcheur (par exemple le TimeStamp) pour le classement des documents. Ces caractéristiques se réfèrent à des signaux temporels et à des propriétés relatives aux requêtes et aux documents.

Selon (Dong *et al.*, 2010a), la notion de la fraîcheur introduit l'idée de l'extraction de l'âge d'un document. Ils ont proposé des caractéristiques de fraîcheur pour représenter la fraîcheur (par exemple Timestamp, Linktime, Webbuzz). Cependant, ils ont intégré les URLs fraîches extraites à partir de Twitter dans un système Web général. Ils ont utilisé ainsi un algorithme d'apprentissage automatique afin de prédire le classement approprié des résultats de recherche pour les requêtes, en se basant sur l'étiquette <requête, url>. Ils ont rajouté ensuite des caractéristiques extraites de Twitter, telles que les caractéristiques textuelles représentant l'ensemble des tweets contenant une URL donnée, afin d'identifier les URL fraîche (Dong *et al.*, 2010b).

(Karkali *et al.*, 2012) déclare un document comme étant frais s'il n'a pas été découvert auparavant par l'utilisateur et s'il est assez frais à l'égard de ses attentes et son historique de navigation. Ils ont proposé une variante de BM25 afin d'extraire

les mots-clés pertinents pour les intérêts des utilisateurs en tenant compte de son historique de navigation. Ils ont recueilli un ensemble de mots-clés des pages Web récemment parcourues par l'utilisateur, introduit la fréquence de chaque terme sur une période donnée et attribué un poids à chaque mot-clé afin d'en extraire les plus frais.

(Wang *et al.*, 2012) ont défini un ensemble de caractéristiques temporelles (par exemple le TimeStamp et le StoryAge de l'URL) permettant d'extraire toutes les dates mentionnées dans l'URL afin d'estimer sa fraîcheur et sa pertinence. (Inagaki *et al.*, 2010) ont déterminé un ensemble de caractéristiques temporelles, telles que « Webbuzz » qui représente le taux de mise à jour de l'URL et qui reflète la popularité des pages Web. Ils ont eu recours au taux de clics pondéré en fonction du temps afin de prélever la pertinence de l'URL pour une requête en calculant le rapport entre le nombre de sessions et celui des clics.

Dans le but d'estimer la variation temporelle de la pertinence des documents pour les requêtes sensibles au temps, (Moon *et al.*, 2012) ont analysé le retour des clics qui dépend du classement des résultats observés par l'utilisateur. (Dai *et al.*, 2011) ont introduit un framework supervisé qui exploite le profil temporel des requêtes en intégrant des caractéristiques de classement pour améliorer la fraîcheur des résultats de recherche. (Lee et Kim, 2013) ont conçu un modèle de clics pour détecter un point de temps, à partir duquel l'intention de recherche pour une requête donnée change.

En outre, plusieurs approches ont été proposées pour identifier et détecter les requêtes sensibles au temps [(Dakka *et al.*, 2012), (Dong *et al.*, 2010a) and (Diaz, 2009)]. Cependant, extraire les caractéristiques temporelles des pages Web constitue une tâche difficile vu qu'elles ne peuvent pas être représentées avec précision dans les documents récents.

Dans notre approche, nous supposons que la fraîcheur, en particulier les thématiques "fraîches", peut être fournie par des sources d'informations en temps réel telles que les actualités, les microblogs et les réseaux sociaux. Ces sources sont susceptibles d'englober des contenus potentiellement utiles à exploiter pour mesurer la fraîcheur des documents.

3. Un modèle de langue estimateur de fraîcheur et de pertinence

La pertinence est une mesure abstraite de la façon dont un document répond aux besoins de l'utilisateur représentés par une requête. Plusieurs travaux de recherche ont exploité les modèles de langue pour modéliser cette pertinence. Ces modèles se basent principalement sur le contenu des documents et exploitent la distribution des

mots dans les documents.

Sous sa version de base, un modèle de langue traite une requête comme étant un échantillon textuel et il estime la probabilité que la requête soit générée par le modèle du document. Plus précisément, la pertinence d'un document d vis-à-vis d'une requête Q est donnée par :

$$P(Q, d) = \prod_{t \in Q} P(t|d) \quad [1]$$

Afin d'éviter le problème de fréquences nulles qui se produit lorsqu'un document ne contient pas un terme de la requête, nous avons utilisé dans l'Équation 1 l'estimateur de lissage Dirichlet (Zhai et Lafferty, 2004), μ est la taille de l'échantillon sur $[0, 1[$:

$$P(t|d) = \frac{tf(t,d) + \mu P(t|c)}{|d| + \mu} \quad [2]$$

Les deux probabilités $P(t|d)$ et $P(t|c)$ sont estimées en utilisant l'estimateur du maximum de vraisemblance. Équation 2 permet d'estimer seulement la pertinence du document d à la requête Q .

Afin de favoriser les documents frais, nous supposons que la fraîcheur est décrite par un ensemble de termes frais connus a priori $F = \{f_1, f_2, \dots, f_n\}$. On tente ensuite d'évaluer la proximité des termes des documents de ceux qui sont frais.

Sur le plan théorique, la probabilité d'un terme peut être réécrite comme suit :

$$P(Q|d) = \prod_{t \in Q} \sum_{f \in F} P(t|f) * P(f|d) \quad [3]$$

Nous optons à mesurer $\hat{P}(t|f)$, qui permet d'estimer le degré d'indépendance entre le terme t et le terme frais f et de définir comment un terme t est proche de f (Equation 4). Cette probabilité peut être estimée à l'aide de la co-occurrence ou la probabilité conditionnelle classique (Bai *et al.*, 2006). Dans notre cas, nous l'estimons tout en prenant en considération le nombre de documents dans les quels t et f apparaissent simultanément. La formulation proposée n'est pas une probabilité pure. Nous avons utilisé le coefficient de Dice, mais on suppose que ce facteur permet d'approcher cette probabilité.

$$\hat{P}(t|f) = \frac{NbDoc(t,f)}{NbDoc(t)+NbDoc(f)} \quad [4]$$

Avec :

$Nbdoc(t)$: Nombre de documents contenant le terme t ,

$NbDoc(f)$: Nombre de documents contenant le terme frais f ,

$NbDoc(t, f)$: Nombre de documents contenant t et f .

Concernant $P(f|d)$, il est estimé en utilisant l'équation 5.

Nous obtenons :

$$P(f|d) = \frac{freq(f,d)}{|d|} \quad [5]$$

Avec :

$freq(f, d)$: La fréquence du terme frais f dans le document d ,

$|d|$: La somme des fréquences de tous les termes dans le document.

$$P(t|C) = \frac{tf(t)}{|C|} \quad [6]$$

Avec :

$tf(t)$: la fréquence du terme t dans tous les documents de la collection,

$|C|$: la somme des fréquences de tous les termes dans la collection,

$P(t|C)$ est la probabilité estimée d'observer le terme t dans la collection. Il constitue le rapport de la fréquence du terme t au le nombre des termes dans toute la collection des documents (Equation6).

4. Expérimentations

Nous menons une série d'expérimentations afin de comparer le modèle de langue proposé avec deux méthodes de référence : le modèle HBM25 proposé par (Karkali *et al.*, 2012) et le modèle surpondéré introduit par (Dong *et al.*, 2010b). Notre objectif consiste à réordonner une liste de documents tout en tenant compte des deux critères de classement : la fraîcheur et de la pertinence.

4.1. Données

Pour construire notre ensemble de données, nous avons collecté 1000 requêtes à partir des grands titres des actualités à partir de l'API «BBC»². En d'autres termes, chaque requête représente un grand titre extrait à partir de « BBC News ». Ces requêtes sont recueillies pendant quatre jours à partir de 08/03/2014 au 11/03/2014, 250 requêtes par jour. Ces requêtes sont susceptibles d'être sensibles aux jours, voire même aux heures. Chaque requête est soumise au moteur de recherche Google. Nous avons sélectionné ensuite les 40 premiers documents (pages Web). L'ensemble des pages Web sont recueillies pour chaque requête afin de former notre collection de documents (notée C). Nous visons à réordonner ces documents en fonction de leur fraîcheur et de leur pertinence. Afin de former notre collection de termes frais, nous avons utilisé l'API « Twitter 4j »³ pour extraire les tweets les plus récents en soumettant des hashtags représentant chaque requête. Nous avons sélectionné les tweets publiés une heure avant l'événement (l'actualité) jusqu'à 12 heures après l'événement. Ensuite, nous avons extrait les X premiers termes les plus fréquents à partir de ces tweets. Nous avons considéré différentes valeurs de X={10, 20, 40, 100} en fonction de leurs fréquences dans les tweets.

Dans la pratique, nous avons employé le moteur de recherche Lucene⁴ intégrant le lemmatiseur Porter et éliminant une liste de mots vides. Afin d'évaluer l'efficacité des résultats, nous avons proposé à 50 personnes de notre université une série de requêtes à évaluer. Chaque évaluateur a jugé 20 requêtes. Ils ont été invités à évaluer les documents en fonction de leur fraîcheur et leur pertinence. Chaque paire requête-document a été évalué le même jour que l'échantillonnage de la requête. Nous définissons deux qualités de pertinence : pertinent et non pertinent. Ensuite, nous leur demandons de juger la fraîcheur de chaque paire requête-document : frais ou non-frais.

Le tableau 1 montre que nous avons exprimé les annotations de pertinence et de fraîcheur séparées dans une seule note représentée par une valeur définie. Nous agrégons les qualités de pertinence et de fraîcheur, de telle sorte qu'un document soit étiqueté pertinent s'il est frais par rapport à la thématique abordée.

Degré de Pertinence/ Degré de Fraîcheur	Frais	Non Frais
Pertinent	2	1
Non Pertinent	0	0

Tableau 1. *Étiquettes agrégées selon le degré de pertinence et de fraîcheur*

2. <https://developer.bbc.co.uk/>

3. <http://www.twitter4j.org/>

4. <http://www.lucene.apache.org/>

Nous avons utilisé la mesure d'évaluation Pertinence et Fraîcheur Cumulées et Actualisées (DCRF) qui permet de pénaliser les erreurs de classement tout en s'appuyant sur les jugements de pertinence et de fraîcheur effectués par les évaluateurs :

$$DCRF_n = \sum_{i=1}^n \frac{RF_i}{\log_2(i+1)} \quad [7]$$

Où i est la position dans la liste des documents réordonnés, n est le nombre de documents et RF_i est la note de la fraîcheur et de la pertinence selon les étiquettes agrégées et éditées par les évaluateurs (Tableau 1). Nous utilisons la MAP calculée en fonction des valeurs RF_i .

Vu que la précision dépend de l'accord entre les évaluateurs, nous avons eu recours au coefficient « Fleiss Kappa ». Cette mesure est envisagée afin de fournir une mesure quantitative de l'ampleur de l'accord entre les évaluateurs (J. Viera et M. Garrett, 2005).

4.2. Modèles de référence

Nous comparons notre modèle avec deux modèles de référence à savoir le modèle HBM25 et le modèle pondéré. Les trois approches tentent de réordonner les résultats de recherche initiale (les premiers documents retournés par Google).

Le modèle HBM25 est une variante de la fonction de pondération des termes BM25 présentée par (Karkali *et al.*, 2012). Il affecte à chaque terme d'un document un poids en se basant sur sa fréquence dans l'historique des pages parcourues par l'utilisateur sur une période donnée. Afin de mettre en œuvre cette méthode, nous avons défini un corpus de tweets envoyés par chaque utilisateur sur une période de temps, puis, nous avons introduit le modèle temporel de chaque terme figurant dans la collection des documents. Ensuite, nous avons employé la fonction de pondération des termes afin d'en extraire les mots-clés les plus frais.

Le second modèle de référence (TWM) est le modèle surpondérée proposé par (Dong *et al.*, 2010a). Ils ont utilisé des caractéristiques de fraîcheur représentées dans les documents récents, tels que les dates, pour fournir leurs évidences temporelles. Pour chaque requête, nous recueillons les paires requête sensible au temps-documents et les paires requête non sensible au temps-document. Ensuite, nous avons eu recours à la fonction de perte utilisée par (Dong *et al.*, 2010a) afin de générer des poids pour chaque paire requête-document en se basant sur le temps de soumission des requêtes et sur les dates figurant dans la collection des documents.

5. Résultats

Le tableau 2 énumère les *MeanDCRF* (MDCRF) et les *MAP* des trois approches pour les quatre jours d'évaluation. On remarque que plus l'événement est ancien, plus la *MAP* et la *MDCRF* des modèles de référence diminuent, tandis que le modèle *FREL* est toujours plus puissant en terme de *MAP* et *DCRF*. Cependant, un document peut devenir éventé quelques heures après, ce qui affecte sa pertinence par rapport à une requête sensible au temps. La *MAP* obtenue pour notre modèle *FRel* surpasse celles calculées pour les modèles HBM25 et TWM de plus de 20% chaque jour. Cela montre que l'extraction de mots-clés frais à partir des tweets récents est potentiellement indispensable pour améliorer le classement de fraîcheur pour les requêtes sensibles au temps réel.

Afin d'évaluer l'accord entre les différents évaluateurs, nous avons utilisé le coefficient Fleiss Kappa. Nous obtenons un Fleiss Kappa de 73,60% sur la fraîcheur des documents et de 78% sur la pertinence des documents, ce qui indique l'existence d'un accord substantiel entre les évaluateurs sur les deux critères.

Le tableau 2 montre que notre méthode obtient les *MDCRF* les plus élevées, soit 60% au cours des 4 jours, alors que les *MDCRF* des autres méthodes sont à moins de 50%.

Notre méthode surpasse le modèle surpondéré (Dong *et al.*, 2010a). Ceci peut être expliqué par l'hypothèse définie par ce modèle : un document est étiqueté « très frais » que si son contenu est créé le même jour de la requête (événement). Toutefois, les caractéristiques de fraîcheur dont les valeurs ne sont pas enregistrées en temps opportun, peuvent être inexactes et peuvent dégrader la pertinence de certains documents récents.

En ce qui concerne la méthode HBM25, certains termes sont pénalisés parce qu'ils apparaissent fréquemment dans la plupart des pages vues par l'utilisateur au cours d'une période donnée, alors qu'ils peuvent être frais quelques jours après.

Le Tableau 3 illustre un exemple des cinq documents récupérés à partir de la liste de recherche retournée par Google et réordonnés par notre modèle pour la requête 'Malaysian MH370 FLIGHT'. Selon les jugements de fraîcheur, nous soulignons que notre modèle favorise le contenu le plus pertinent et le plus frais en haut de la liste des documents réordonnés.

		HBM25	TWM	FRel
Jour 1	MAP	41.2	50.21	78
	MDCRF	50.01	39.30	64.20
Jour 2	MAP	41	50.10	69.22
	MDCRF	42.12	43.79	87.31
Jour 3	MAP	40.15	46.50	68.74
	MDCRF	36.30	42	68.92
Jour 4	MAP	41.22	35	75.60
	MDCRF	38.44	34.27	75.30

Tableau 2. *Comparaison de FRel avec les modèles de référence selon la MAP et la DCRF*

	rang	Document	Pertinence	Fraîcheur
(a) Résultats de recherche initiale retournés par le moteur de recherche Google				
rang dans (b)	1(4)	http://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370	pertinent	non-frais
	2(5)	http://www.malaysiaairlines.com/mh370	non pertinent	frais
	3(12)	http://www.independent.co.uk/news/world/australasia/\mh370-airline-boss-claims-missing-flight-did-not-crash-into-indian-ocean-\9790455.html http://www.redbullstratos.com/about-pilot-biography	non pertinent	non-frais
	4(2)	http://www.bbc.com/news/world-asia-26503141	pertinent	frais
	5(8)	http://www.atlantico.fr/dossier/disparition-\vol-mh370-malaysia-airlines-1015555.html	non pertinent	non-frais
(b) Reclassement des résultats par FRel				
rang dans (a)	1(9)	http://timesofindia.indiatimes.com/the-mystery-of-Malaysia-Airlines-flight-MH370/specialcoverage/32010069.cms	pertinent	frais
	2(4)	http://www.bbc.com/news/world-asia-26503141	pertinent	frais
	3(6)	http://www.redbullstratos.com/the-team/felix-baumgartner/	pertinent	frais
	4(1)	http://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370	pertinent	non-frais
	5(2)	http://www.malaysiaairlines.com/mh370	non pertinent	frais

Tableau 3. Un exemple d'amélioration de classement selon la fraîcheur en se comparant aux résultats de recherche retournés par Google. La requête est "Malysian Flight", la durée de lancement de la requête est de 16 :31 à 20 :59

Nous avons également évalué l'impact du nombre de mots-clés frais. La Figure 1 montre que plus le nombre des mots-clés frais existants dans un document est élevé et varie au fil du temps, plus le *DCRF* et la *MAP* augmentent.

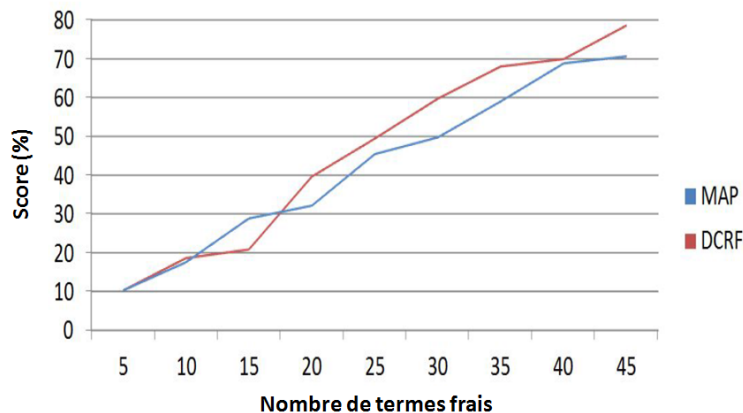


Figure 1. L'effet du nombre de mots-clés frais sur la performance des scores

6. Conclusion

Nous proposons un modèle de langue qui permet d'estimer la pertinence et la fraîcheur des documents Web à l'égard des requêtes sensibles au temps en se basant sur les mots-clés frais et pertinents aux thématiques extraits à partir des tweets les plus récents. Les résultats obtenus soulignent que notre approche surpasse les modèles de référence. Les expérimentations réalisées montrent qu'il est indispensable d'utiliser l'information sociale pour mesurer la fraîcheur des documents afin de répondre aux requêtes sensibles au temps. Nous avons également montré l'importance d'utiliser Twitter comme étant une source pour l'extraction des termes frais récurrents dans les microblogs et décrivant les événements les plus récents. Nos résultats indiquent ainsi l'importance des modèles de langue pour améliorer le classement de fraîcheur des documents Web. Les résultats encourageants de notre modèle ouvrent plusieurs orientations futures : Nous envisageons dans un premier temps d'intégrer notre modèle dans les systèmes de recommandation afin de proposer des programmes de télévision les plus pertinents et les plus frais répondant aux différentes attentes des utilisateurs. Un autre axe de recherche sera d'étendre la personnalisation par d'autres caractéristiques de classement, tel que le comportement social de l'utilisateur, et d'exploiter les annotations pour traiter la fraîcheur des émissions TV.

Bibliographie

Bai J., yun Nie J., Cao G., « Context-dependent term relations for information retrieval », *Proc. Empirical Methods in Natural Language Processing (EMNLP '06)*, p. 551-559, 2006.

- Ben Jabeur L., Tamine L., Boughanem M., « Featured tweet search : Modeling time and social influence for microblog retrieval », *Proceedings of International Conference on Web Intelligence*, China, p. 166-173, 2012.
- Dai N., Shokouhi M., Davison B., « Learning to Rank for Freshness and Relevance », *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 95-104, 2011.
- Dakka W., Gravano L., G. Ipeirotis P., « Answering General Time-Sensitive Queries », *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*, vol. 24, p. 220-235, 2012.
- Damak F., Pinel-Sauvagnat K., Guillaume C. G., « Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? », *Conférence francophone en Recherche d'Information et Applications (CORIA)*, Bordeaux, France, LABRI, p. 317-328, mars, 2012.
- Diaz F., « Integration of news content into web results », *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ACM, New York, NY, USA, p. 182-191, 2009.
- Dong A., Chang Y., Zheng Z., Mishne G., Bai J., Zhang R., Buchner K., Liao C., Diaz F., « Towards recency ranking in web search », *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, ACM, New York, NY, USA, p. 11-20, 2010a.
- Dong A., Zhang R., Kolari P., Bai J., Diaz F., Chang Y., Zheng Z., Zha H., « Time is of the essence : improving recency ranking using Twitter data », *Proceedings of the 19th international conference on World Wide Web*, WWW '10, ACM, New York, NY, USA, p. 331-340, 2010b.
- Huo W., J. Tsotras V., « Temporal top-k search in social tagging sites using multiple social networks », *Proceedings of the 15th International Conference on Database Systems for Advanced Applications*, p. 498-504, 2010.
- Inagaki Y., Sadagopan N., Dupret G., Dong A., Liao C., Chang Y., , Zheng Z., « Session Based Click Features for Recency Ranking », *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010.
- J. Viera A., M. Garrett J., « Understanding Interobserver Agreement : The Kappa Statistic », *Family Medecine Research Series*, vol. 37, n° 5, p. 360-363, 2005.
- Karkali M., Plachouras V., Vazirgiannis M., Stefanatos C., « Keeping Keywords Fresh : A BM25 Variation for Personalized Keyword Extraction », *Proceedings of the 2nd Temporal Web Analytics Workshop*, p. 17-24, 2012.
- Lee S., Kim D., « A click model for time-sensitive queries », *Proceedings of the 22nd International World Wide Web Conference*, WWW '13, p. 147-148, 2013.
- Massoudi K., Tsagkias M., Rijke M., Weerkamp M., « Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts », *Proceedings of the 33rd European Conference on IR Research*, Dublin, Ireland, p. 362-367, 2011.
- Moon T., Chu W., Lihong L., Zheng Z., Chang Y., « Online learning for recency search ranking using real-time user feedback », *ACM Transactions on Information Systems*, vol. 30, n° 4, p. 20, 2010.
- Moon T., Li L., Liao C., Zheng Z., Chang Y., « An online learning framework for refining recency search results with click feedback », *ACM Transactions on Information Systems*, vol. 30, n° 4, p. 20, 2012.

- Pasca M., « Towards temporal web search », *Proceedings of the 2008 ACM symposium on Applied computing*, ACM, New York, NY, USA, p. 1117-1121, 2008.
- Wang H., Dong A., Li L., Chang Y., « Joint Relevance and Freshness Learning From Click-throughs for News Search », *Proceedings of the 21st International World Wide Web Conference Committee*, p. 579-588, 2012.
- Zhai C., Lafferty J., « A Study of Smoothing Methods for Language Models applied to Information Retrieval », *ACM Transactions on Information Systems*, vol. 2, n^o 2, p. 179-214, 2004.

Ateliers

L'innovation par les SI dans l'écosystème

Porteurs : Abdelkader Achi (CRI, Université Paris 1 Panthéon-Sorbonne), Charlotte Hug (CRI, Université Paris 1 Panthéon-Sorbonne), Camille Salinesi (CRI, Université Paris 1 Panthéon-Sorbonne)

Résumé :

Alors qu'il est très largement reconnu que le développement des entreprises repose en grande partie sur l'innovation et que l'informatique s'y est durablement installée, notamment par le biais des Systèmes d'Information, on dispose de peu de connaissances concernant l'innovation par les Systèmes d'Innovation. L'atelier « innovation par les SI dans l'écosystème », organisé pour la première fois dans le cadre du congrès INFORSID 2015 a pour ambition de créer une communauté de chercheurs autour du thème de l'innovation par les SI. L'ère du numérique se caractérise par le besoin d'innover, la collaboration et le partage des nouveaux savoirs (Gartner, 2014). Bien entendu, l'innovation est consubstantielle au développement de l'informatique observé lors de la seconde moitié du 20ème siècle et en ce début de 21ème siècle. Néanmoins, ce qu'attendent les entreprises ne peut être résumé à de simples développements technologiques – dont elles seraient, pour la plupart, simples consommatrices. Il s'agit plutôt d'utiliser les SI comme levier. Cela peut être parce que les changements entraînés par l'introduction ou la rénovation des SI amènent à s'interroger sur l'entreprise elle-même : ses processus, ses services, les leviers de sa performance, ses partenariats et marchés, etc. Une autre approche, plus conceptuelle consiste à s'approprier les méthodes, techniques et outils initialement pensés pour l'ingénierie des SI, et à les réutiliser pour l'entreprise elle-même. Enfin il peut s'agir de nouvelles modalités sociales, cognitives, organisationnelles, ou culturelles amenées au cours des vagues de révolutions numériques. L'enjeu pour le monde des SI est important : le rôle des SI au sein des organisations balance entre centre de coûts et centre de valeurs. Une tendance forte de ces dernières années est donc d'utiliser des SI pour co-créditer de la valeur avec les parties prenantes et innover. On assiste ainsi à la prolifération de nouveaux espaces de création de connaissances et d'innovation : crowdsourcing, communautés, médias sociaux, etc. et de nouveaux modèles d'innovation collaboratifs, inclusifs, ouverts à tous les acteurs de l'écosystème. Cependant, hormis les cas de succès flagrants de « digital natives » tels que Google, Facebook et Amazon, on ne dispose encore que de très peu de témoignages, et encore moins d'études et de données chiffrées, de modèles, de bonnes pratiques, ou de référentiels qui nous permettraient de bien comprendre les leviers de l'innovation par les SI, les facteurs critiques de succès, l'efficacité des différentes méthodes, techniques et outils, etc. Or une telle connaissance est cruciale si l'on souhaite permettre aux organisations d'adopter une démarche d'amélioration fondée sur les bonnes pratiques, rationnelle et systématique. Comme dans tous les autres domaines d'ingénierie des SI, mener une démarche systématique d'innovation qui s'appuie de manière efficace sur les SI est un problème difficile qui requiert des savoir-faire spécifiques assis sur une connaissance établie scientifiquement. Il est indéniable que certaines entreprises arrivent à actionner les bons leviers pour relever les défis de l'innovation par les SI, mais les expériences malheureuses montrent des degrés de maturités très divers, et l'on peut s'interroger sur la réelle utilisation de savoirs scientifiques attestés et d'outils d'ingénieurs éprouvés dans les démarches d'innovation. L'atelier « innovation par les SI dans l'écosystème », cherche à répondre à cette problématique par une réflexion commune entre industriels et chercheurs sur les bonnes pratiques, facteurs de succès et freins, processus ou dispositifs à mettre en œuvre dans une démarche d'innovation par les SI. Les thèmes abordés concernent aussi bien des sujets tels que les méthodes de créativité, que le crowdfunding, ou la gestion des communautés et les modèles participatifs d'innovation, que les écosystèmes d'innovation, ou l'impact des technologies nouvelles telles que le Cloud sur l'innovation dans les entreprises.

Comité de programme :

Carole BOUCHARD (Ecole Nationale Supérieure d'Arts et Métiers)

Karim HEDEOUD-PERROT (Culturibles)

Martin MAHAUX (Université de Namur, Belgique)

Dominique RIEU (Université de Grenoble)

Nicolas VERNEY (GDF SUEZ)

Autour du concept d’Observatoire en environnement

Porteurs : André Miralles (Irstea, UMR Tetis), Thérèse Libourel (UM, UMR Espace Dev), François Pinet (Irstea, UR TCFC)

Résumé :

Face aux divers enjeux environnementaux et sociétaux, le concept d’observatoire longtemps dévolu à la stricte observation de l’univers évolue et diverses acceptions voient le jour. Quoi qu’il en soit, à l’heure actuelle, ce concept recouvre le domaine des Systèmes d’observation et d’information jusqu’à celui des systèmes d’aide à la décision et appui à la gouvernance. L’atelier souhaite confronter des points de vue différents sur :

- les verrous essentiels en termes de modélisation (notamment au niveau des échelles de temps et d’espace, du choix des indicateurs) et de mises en œuvre innovantes (réseaux de capteurs, collaboration, interaction, etc.)
- les définitions et formalisation de ces observatoires, leur rôle et fonctionnalités (observation, diffusion, suivi, simulation), leur périmètre, leur architecture ainsi que sur des expérimentations relatives à divers secteurs :
- observatoire de phénomènes naturels,
- observatoires de phénomènes sociétaux (santé, trafic, agriculture, etc.)
- et leur apport en termes de difficultés rencontrées, de malentendus entre scientifiques et autres acteurs, etc.

Comité de programme :

Nabil Ben Khatra (OSS Tunisie)
Françoise Gourmelon, CNRS
KhiariHabiba, OSS Tunisie
Frédéric Hubert, Université Laval
Olivier Lobry, CNRS OSU OREME
Philippe Lopistéguy, Université de Pau
Mir Mostafavi, Université Laval
Isabelle Mougenot, UM2 – UMR EspaceDEv
Franck Ravat, IRIT
Ronan Tournier, IRIT

Modélisation des cartographies : principes, démarches, et techniques

Porteurs : Yemna Sayeb (RIADI, Université de la Manouba), Henda Hajjami (RIADI, Université de la Manouba), Wahiba Bahsoun (IRIT, Université Paul Sabatier)

Résumé :

Aujourd'hui les initiatives, démarches, méthodes et outils destinés à faire évoluer le SI se multiplient. Ces offres proposent une vision globale du SI, de la vue métier jusqu'aux couches techniques. Des cadres de références ont existé et les éditeurs de logiciels proposent des solutions allant de la description des processus métiers, la définition des services fonctionnels du SI, ..., jusqu'à l'implémentation des services techniques. Dans ce contexte nous observons un recours à la modélisation à double objectif : un objectif de compréhension décrivant l'existant du SI, à travers des modèles contemplatifs, et un objectif d'intervention définissant une cible pour le SI et mettant en évidence des modèles productifs. C'est dans ce contexte que les cartographies se présentent comme des moyens de modélisation de plus en plus recommandés. Elles permettent de fournir une image la plus exacte et la plus exhaustive du système d'information qui évolue dans le temps et constituent la colonne vertébrale de toute démarche d'urbanisation à travers la représentation du SI à différents niveaux d'abstraction. En fait, la carte est un vecteur d'information qui permet d'accéder facilement aux différents types d'information selon les différentes visions d'étude du système d'information. L'élaboration et la modélisation des cartographies sont donc devenues un moyen stratégique pour la compréhension du système d'information et surtout pour anticiper les changements.

Le principal objectif de cet atelier est de rassembler la communauté de chercheurs afin de débattre ensemble les problèmes actuels rencontrés dans la modélisation des cartographies et surtout de mettre au clair l'apport de la communauté des professionnels dans cette thématique. L'atelier se veut un espace d'échange entre les deux mondes de chercheurs et professionnels afin d'alimenter une réflexion commune sur le rôle stratégique de la modélisation des cartographies.

Comité de programme :

Bahsoun Wahiba, IRIT, Université Paul Sabatier Toulouse 3
 Ben Ghezala Henda, RIADI, ENSI, Université de la Manouba
 Boughzala Imed, Département Systèmes d'Information (DSI) Telecom Ecole de Management – Institut Mines-Télécom 9
 Boujlida Naceur, LORIA, Université de Lorraine
 GaaloulKhaled, LIST, The Luxembourg Institute of Science and Technology
 El Abed Lamia, RIADI, ISG, Université de Tunis
 Ghannouchi Sonia, RIADI, ISG, Université de Sousse
 Ghenima Malek, RIADI, ESC, Université de la Manouba
 Hbacha Anja, RIADI, ENSI, Université de la Manouba
 Jammoussi Yassine, RIADI, ENSI, Université de la Manouba
 Kraiem Naoufel, RIADI, ISAMM, Université de la Manouba
 Salinesi Camille, CRI, Université Paris 1 Panthéon Sorbonne
 Sayeb Yemna, RIADI, ISAMM, Université de la Manouba
 Simonin Jacques, STICC, Institut Télécom / Télécom Bretagne – Technopôle Brest Iroise

SI, Connaissance, Homme numérique et décision

Porteurs : Pierre Emmanuel Arduin (PSL, Université Paris-Dauphine), Linda Atif (PSL, Université Paris-Dauphine), Camille Rosenthal-Sabroux (PSL, Université Paris-Dauphine)

Résumé :

« One day, computers, bytes, and bits will be a part of almost everything we do. ¹ » (Negroponte, 1995, p. 43)

Les récentes évolutions des technologies de l'information et de la communication reposent pour l'essentiel sur l'augmentation, l'accélération et parfois même l'omniprésence d'une pratique devenue indispensable dans les organisations : la numérisation. Le numérique change nos habitudes nous rendant ainsi des Hommes numériques comme l'avait prédit Negroponte (1995) lorsqu'il parlait de « Digital Beings ». L'Homme numérique est un hyperactif numérique dépendant d'outils numériques qui ont la particularité d'être au cœur de ses activités, ils se dressent entre lui et ce qui l'entoure, ils peuvent l'aider dans son quotidien mais il changent aussi sa façon d'être, le rendant « hybride » (Calmart, 2015). L'Homme numérique entretient des rapports spécifiques avec ce qui l'entoure : l'espace, le temps, la mémoire et la connaissance. Ainsi, il a des besoins nouveaux : il aimerait effectuer plusieurs tâches simultanément, réduire son temps d'accès à l'information, prendre de meilleures décisions, acquérir et partager des connaissances. De cette évolution des pratiques résulte une évolution des entreprises qui tendent à devenir des « entreprises numériques » (Ménard, 2010). Ainsi, une nouvelle façon de penser les organisations voit le jour : les processus, la recherche et l'innovation, la gouvernance, la communication, la relation client, la gestion des carrières et des compétences doivent être pensés différemment pour prendre en compte les spécificités de l'Homme numérique.

Il est admis qu'un système d'information peut être vu comme un ensemble de ressources numériques et humaines organisées afin de traiter, diffuser et stocker des informations (Reix, 2002). Des observations que nous avons pu mener sur le terrain dans des organisations montrent qu'un système d'information, ensemble de ressources numériques et humaines, doit être managé : ses potentialités et ses limites doivent être connues et maîtrisées, non seulement d'un point de vue technologique, mais aussi d'un point de vue managérial, organisationnel et humain. Dans les organisations, les individus ne sont plus de simples utilisateurs du système d'information, mais ils en sont des composants à part entière. Leurs vies sociales sont par exemple supportées par des réseaux sociaux en ligne et leur santé est entretenue au travers de montres connectées. Le nombre des artefacts numériques augmente et cette augmentation va de pair avec l'évolution des usages qui en sont faits (Canohoto et al., 2015) : là où hier un terminal reliait l'utilisateur à l'ordinateur, aujourd'hui les points d'entrée dans le système d'information sont multiples, universels, interopérants et de plus en plus discrets. La numérisation de masses de données et leur traitement au travers de concepts tels que le big data tendent à faire de l'Homme un « Homme numérique ».

La prise de décision est une construction psychologique. On peut en effet déduire d'un comportement observable qu'une décision a été prise. Dans une organisation, un individu prend une décision à partir des connaissances dont il est porteur et qu'il a contextualisées au regard des informations qui lui sont parvenues notamment au travers du système d'information. Les spécificités de l'Homme numérique sont susceptibles d'impacter ses processus décisionnels : son interactivité exacerbée avec des outils numériques ou encore les quantités des données qu'il injecte et récupère dans un réseau sans vraiment en avoir conscience.

Dans ce contexte tout à la fois nouveau pour les organisations mais acquis pour les individus qui

1. « Un jour, les ordinateurs, les octets et les bits feront partie de presque tout ce que nous faisons »

sont bien souvent déjà des Hommes numériques, quel serait le rôle et la place des systèmes d'information ? Qu'est-ce qu'un « Homme numérique » et comment répondre à ses besoins ? Quel impact l'analyse des données numériques a-t-elle sur les processus décisionnels et sur les organisations ?

L'objectif de cet atelier est d'être un espace d'échange ouvert entre chercheurs et professionnels de plusieurs domaines au travers d'une table ronde et de débats afin d'alimenter une réflexion commune sur ce qu'est « le numérique », sur les spécificités de l'Homme numérique et son impact sur les systèmes d'information et de décision.

Canohoto, A.I., Dibb, S., Simkin, L., Quinn, L., and Analogbei, M. (2015). Preparing for the future – how managers perceive, interpret and assess the impact of digital technologies for business, In : Proceedings of the 48th Hawaii International Conference on System Sciences.

Calmart P. (2015). L'homme à venir, Edition SW Télémarque.

Ménard, B. (2010). L'entreprise numérique, Nuvis.

Negroponte N. (1995). Being Digital, Laffront.

Reix, R. (2002). Système d'information et management des organisations, Vuibert, 4ème édition, Paris.

Comité de programme :

Pierre-Emmanuel Arduin, PSL, Université Paris-Dauphine, DRM UMR CNRS 7088

Linda Atif, PSL, Université Paris-Dauphine, LAMSADE UMR CNRS 7243

Yacine Atif, United Arab Emirates University, Émirats arabes unis

Cédric Campo-Paysaa, ON-X Groupe, Puteaux

Michel Grundstein, PSL, Université Paris-Dauphine, LAMSADE UMR CNRS 7243

Isabelle Linden, Université de Namur, Belgique

Elsa Negre, PSL, Université Paris-Dauphine, LAMSADE UMR CNRS 7243

Káthia Oliveira, Université de Valenciennes, LAMIH UMR CNRS 8201

Alexandre Pachulski, Talentsoft, Boulogne-Billancourt

Camille Rosenthal-Sabroux, PSL, Université Paris-Dauphine, LAMSADE UMR CNRS 7243

Ines Saad, ESC Amiens, Laboratoire MIS, Université de Picardie Jules Verne

Thierno Tounkara, Institut Mines Telecom, Télécom Ecole de Management, Évry

Dragos Vieru, École des sciences de l'administration, TÉLUQ, Université du Québec, Canada

Journée bilatérale AFIA / INFORSID SI&IA

Les utilisateurs au centre du SI : des modèles aux raisonnements

Porteurs : Catherine Faron Zucker (AFIA), Mireille Blay Fornarino (INFORSID), Célia da Costa Pereira (AFIA)

Résumé :

Par définition même du système d'information (SI), l'utilisateur est au cœur de celui-ci. Depuis plusieurs années, en particulier dans la communauté INFORSID, des compétences scientifiques et technologiques ont été développées pour renforcer la prise en compte de l'utilisateur que ce soit dans la conception du SI, dans l'analyse de son usage ou dans les raisonnements liés par exemple à l'aide à la décision. Des thématiques de recherche analogues sont en plein essor dans la communauté de l'Intelligence Artificielle : en plongeant l'utilisateur au cœur des systèmes à base de connaissances, l'objectif est d'assister l'utilisateur dans le traitement de la masse des données disponibles ; ceci afin de les visualiser, contextualiser, expliquer, recommander, classer, prédire, approcher, sécuriser, fiabiliser, pour pallier leur imprécision, leur incohérence. L'objectif de cet atelier est d'être un espace de discussion, de regards croisés et de confrontation entre les problèmes rencontrés sur le terrain, les solutions pratiques mises en œuvre et les avancées scientifiques dans la communauté INFORSID et celle de l'Intelligence Artificielle.

Comité de programme :

Nathalie Aussenac Gilles, IRIT
Annie Cavarero, UNS, I3S
Salem Benferhat, CRIL Artois
Elise Bonzon, LIPADE
Laurence Cholvy, ONERA
Sylvie Despres, Université Paris 13 - LIMICS
Hubert Dubois, CEA
Sophie Dupuy-Chessa IMAG
Jérôme Euzenat, INRIA
Sébastien Ferré, IRISA
Marie-Pierre Gervais, LIP6
Souhila Kaci, LIRMM
Sébastien Laborie, Université de Pau, LIUPPA
Frédérique Laforest, Telecom Saint Etienne

RegineLaleau Université Paris Est Créteil, LACL
Jérôme Lang, LAMSADE
Marie-Christine Lagasquie-Schiex, IRIT
Xavier Lepallec Université de Lille, Cristal
Jean Lieber, LORIA
Pierre Marret, Université Jean Monnet, Laboratoire Hubert Curien
Guy Melançon, LABRI
Franck Michel, I3S
Alain Mille, LIRIS
Pascal Molli, Université de Nantes - LINA
JosianeMothe, IRIT
Clementine Nemo ATOS
Pascal Neveu, INRA
Nathalie Pernelle, LRI
Gabriella Pigozzi, LAMSADE
Thomas Polacsek, Onera
Dominique Rieu, LIG
Chantal Reynaud, Université Paris-Sud - LRI
Mathieu Roche, CIRAD
Marie-Christine Rousset, LIG

Résumé

Ce document contient les actes du trente-troisième congrès INFORSID (INFormatique des ORganisations et Systèmes d'Information de Décision) qui s'est déroulé à Biarritz du 26 au 29 mai 2015.

Le processus de sélection des articles publiés a été organisé à deux niveaux avec un Conseil du Comité de Programme (CoP) additionnel au Comité de Programme habituel (CP). Les membres du CoP ont organisé une méta-évaluation d'un pool d'articles qui leur ont été affectés. La méta-évaluation a consisté à organiser les discussions entre les membres du CP relecteurs de chaque article afin de résoudre les conflits d'évaluation et d'aboutir, dans la mesure du possible, à un consensus. Les membres du CoP ont rédigé, à la fin du cycle de discussions, une brève évaluation de synthèse pour chacun des articles de leur pool d'articles. Seuls les membres du CoP ont participé à la réunion de sélection finale.

