Introduction Forum Jeunes Chercheuses Jeunes Chercheurs Inforisd 2018

Thomas Polacsek

ONERA
Toulouse

Le Forum Jeunes Chercheuses Jeunes Chercheurs (JCJC) d'Inforsid a eu lieu cette année à Nantes, lors de son congrès annuel. Événement bisannuel, le JCJC permet à des doctorantes de première ou deuxième année de venir présenter leurs travaux. Pour la communauté des chercheuses centrées sur l'informatique des organisations et les systèmes d'information, les JCJC sont l'occasion de prendre le pouls de la recherche francophone, de voir les thèmes de recherche qui animent la communauté. Les JCJC sont un instantané qui donne à voir un panorama des questionnements scientifiques en cours et la neuvième édition du Forum n'a pas dérogé à la règle. Ainsi, ce sont six doctorantes et doctorants qui sont venus présenter leurs travaux en session plénière puis devant leur poster.

La première impression qui se dégage des travaux présentés est l'importance des thématiques relatives aux données massives. En effet, si les données massives représentent l'un des grands défis informatiques de la décennie, comme le souligne Stéphane Mallat, titulaire d'une Chaire « Sciences des données » créée cette année 2018 au Collège de France, les questions portant sur la collecte, la gestion, le traitement et l'emploi de ces données sont au cœur des préoccupations de la communauté Inforsid. D'ailleurs trois présentations des JCJC s'inscrivent dans cette thématique. Partant de données brutes et non structurées, la thèse de Rabah Tighilt Ferhat vise à extraire et à restructurer ces données dans le but d'effectuer des traitements décisionnels. Concernant le traitement, et plus précisément le traitement efficient, la thèse de Nabil El Malki se focalise sur l'optimisation des accès complexes et répétitifs appliqués à des données massives. Toujours sur la problématique des traitements massifs, mais avec une approche centré infrastructure, la thèse de Junior Dongo interroge les apports possibles des architectures réseaux centrée sur la donnée. Se focalisant non plus sur le traitement brut, mais sur la donnée et les informations afférentes la thèse de Franck Jeveme Panta vise à proposer une modélisation des mécanismes de filtrage de grandes collections de vidéo dans le cadre des systèmes de vidéosurveillance. Toujours sur le

thème de la gestion de l'information, la thèse de Tiphaine Van de Weghe cherche à apporter des solutions aux chercheuses et chercheurs en Sciences Humaines et Sociales pour collecter, organiser et utiliser leurs données. Pour finir, parmi l'ensemble des données produites, ce trouve aussi les profils utilisateurs que possèdent les différents services que nous utilisons et ces justement à partir de ces profils que Soufiane Faieq, dans le cadre de sa thèse, cherche à établir une composition de services sensible au contexte.

Comme nous le voyons au travers de ces six sujets de thèses présenté ici, la commuté s'est très largement emparée de la question de la massification des données, de leur traitement et, en parallèle, continue son travail sur la gestion de l'information. Cependant, en marge de ces travaux essentiels, nous pouvons espérer que, dans un avenir proche, les questions relatives à l'impact de ces collectes massives de données et aux capacités sans cesse accru de traitements seront elles aussi embrassées par la communauté. L'avenir nous le dira. Rendez-vous est pris en 2020 pour la dixième édition du Forum JCJC.

Table des matières

Vers un framework de composition de services sensible au contexte pour les environnements intelligents

Soufiane Faieq

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG*
CS 40700 - 38058 Grenoble cedex 9
F-38000, Grenoble, France

soufiane.faieq@univ-grenoble-alpes.fr

MOTS-CLÉS : *Composition de Services, Réutilisabilité de Services, Sensibilité au Contexte, Informatique Orientée Services, Informatique Ubiquitaire, Apprentissage Automatique*

KEYWORDS: *Service Composition, Service Reuse, Context-Awareness, Service-Oriented Computing, Ubiquitous Computing, Machine Learning*

ENCADREMENT. *Agnès Front (MCF), Moulay Driss Rahmani (PR), Rajaa Saidi, Hamid Elghazi*

1. Contexte

Dans le monde connecté d'aujourd'hui, les entreprises se trouvent dans l'obligation de fournir leurs services en ligne pour garder leur compétitivité. Dans ce contexte, l'Informatique Orientée Services (Service Oriented Computing) a fait ses preuves comme un paradigme permettant l'amélioration de l'agilité des organisations, ainsi que la facilitation de la coopération inter-organisationnelle avec l'utilisation des principes de conception comme la réutilisabilité, la virtualisation et la composition de services.

Alors que les entreprises cherchent toujours des moyens pour capitaliser sur la richesse des données disponibles grâce aux avancées de la technologie pour améliorer leurs services offerts, les attentes des utilisateurs concernant leurs expériences en interagissant avec les fournisseurs de services deviennent de plus en plus élevées. Cela est particulièrement vrai pour les tâches complexes et qui dépendent du contexte de

l'utilisateur, telles que la planification de voyages impliquant de nombreuses activités (par exemple, trouver et réserver des hébergements, transports, événements intéressants, etc.) où l'utilisateur doit passer au crible une grande quantité d'informations (offres) et faire beaucoup de décisions pour réaliser son objectif.

Nos travaux de recherche visent à explorer l'utilisation de la notion de contexte issue de l'informatique ubiquitaire dans la composition de services, pour supporter et intégrer les utilisateurs dans la création de services, qui sont capables de satisfaire leurs besoins ; particulièrement dans le contexte des environnements intelligents qui sont l'extension virtuelle du monde physique, capables d'offrir des services sur mesure à ses habitants et de s'adapter à leurs besoins.

2. État de l'art

L'intégration de l'utilisateur et de son contexte dans le processus de création de services est rarement étudié dans la littérature (Sheng *et al.*, 2014). En effet, c'est le fournisseur qui est au centre de l'Architecture Orientée Services puisque les utilisateurs ne peuvent utiliser que les services répertoriés par le fournisseur. Deux extensions de cette vision ont été présentés par (Tsai *et al.*, 2006; Chang *et al.*, 2006) nommées CCSOA(Consumer-Centric Service Oriented Architecture) et UCSOA(User-Centric Service-Oriented Architecture) qui mettent les consommateurs (développeurs) et les utilisateurs finaux respectivement au centre du processus de la composition de services.

Les travaux proposés dans la littérature pour supporter les utilisateurs et les intégrer dans la création de services opèrent généralement au niveau de la composition de services. Parmi ces travaux, des approches se basent sur les ontologies comme celle de (Xiao *et al.*, 2011), où les auteurs présentent une approche pour masquer la complexité du processus de la composition des services et supporter les utilisateurs dans l'exécution de ce processus. Dans cette approche, les ontologies sont utilisées pour encapsuler les connaissances du domaine ainsi que pour déduire des flux de contrôle pour améliorer la génération des processus de composition.

Une autre famille d'approches se base sur les techniques de l'apprentissage automatique pour déduire des modèles qui ont pour but d'améliorer un ou plusieurs aspects de la composition. Alors que certains travaux cherchent à modéliser et intégrer le comportement de l'utilisateur comme celui de (He *et al.*, 2015), d'autres cherchent à trouver les différents services qui sont souvent composés ensemble pour les présenter comme des recommandations pour l'utilisateur (Zhang *et al.*, 2017; Labbaci *et al.*, 2017).

3. Problématique

La création de services dans l'Architecture Orientée Services évoque naturellement une relation fournisseur-consommateur (utilisateur). Cependant, l'analyse de la littérature montre que l'on se focalise généralement beaucoup plus sur la vision du fournisseur que sur la vision de l'utilisateur. Les utilisateurs se trouvent donc incapables d'utiliser

les services créés pour satisfaire leurs besoins. Ce problème est encore plus prononcé dans les environnements ubiquitaires d'aujourd'hui où les équipements numériques se sont immiscés partout dans notre environnement et fournissent des services cruciaux et personnels aux utilisateurs, à l'exemple des maisons intelligentes.

Par ailleurs, pour avoir des services plus pertinents et utilisables, le rôle de l'utilisateur, ses besoins et ses préférences doivent être aussi définis et intégrés dans le processus de composition de services. Dans ce contexte nous visons à répondre aux questions suivantes :

– Comment peut-on intégrer l'utilisateur, à travers son contexte, dans le processus de composition de services au bénéfice du fournisseur et de l'utilisateur ?

– Quel serait l'impact de l'intégration des informations de contexte dans les algorithmes d'apprentissage automatique pour la recommandation de services sur la satisfaction de l'utilisateur ?

4. Actions réalisées

Nos recherches se sont focalisées sur l'apport du contexte dans le processus de création de services spécialement dans les environnements dits ubiquitaires ou pervasifs supportés par l'informatique en nuage (Cloud Computing), tout en examinant le rôle des utilisateurs dans ce processus.

Dans un premier temps, nous nous sommes concentrés sur l'importance du contexte dans le processus de composition de services. Nous avons proposé dans (Faieq *et al.*, 2017b), une architecture pour la composition sensible au contexte de services. Nous défendons l'importance du contexte pour (i) l'utilisateur, dans la minimisation des interactions avec les services ou le système de composition de services et (ii) le système de composition pour améliorer les performances de ce dernier ainsi que (iii) la satisfaction de l'utilisateur du service composite résultant. Nous détaillons la façon dont laquelle les informations contextuelles peuvent être bénéfiques dans chaque phase du processus de composition, ainsi que les techniques qui sont utilisées dans chacune. Cette architecture est appuyée par un méta-modèle du contexte pour représenter la situation des entités participantes dans la composition (utilisateurs et services).

Dans un second temps, nous avons étendu le travail précédent en étudiant les synergies existantes entre les différentes technologies que nous avons jugées principales dans le développement des services dans les environnements intelligents à savoir, Big data, Internet des objets, Informatique en nuage et la sensibilité au contexte. Ces environnements présentent un nouveau paradigme où les services sont fournis non seulement par des fournisseurs de services mais aussi par des équipements et appareils contenant des capteurs et des actionneurs (comme les smartphones, smartWatch, smartTV, etc.). Nous présentons donc dans (Faieq *et al.*, 2017a) un framework centré sur la notion du contexte pour le développement des services dans les environnements intelligents, dans lequel nous expliquons comment ces technologies peuvent collaborer pour créer des services qui sont puissants, pervasifs et pertinents au contexte de l'utilisateur.

5. Actions futures

Pour tester notre framework, nous visons à développer une application dédiée aux voyageurs (touristes ou affaires). Nous tenons à analyser les traces d'exécutions à travers l'application des algorithmes d'apprentissage automatique (i.e. règles d'association) pour dégager (i) les services qui sont souvent composés ensemble (ii) les clusters des utilisateurs avec des intérêts communs. Pour évaluer la pertinence du contexte dans la qualité des compositions résultantes, les résultats seront comparés contre les approches de l'état de l'art qui ne tiennent pas compte du contexte dans leurs modèles.

La phase de découverte de services est primordiale dans le processus de composition de services. à cet égard, nous souhaitons également évaluer la pertinence des langages de description de services actuels, fournis dans le contexte des environnements intelligents, dans composition de services. à cette fin, nous supposons que l'intégration des éléments de contexte dans ces langages permettront d'améliorer les résultats de la découverte et donc de la composition de services.

6. Bibliographie

Chang M., He J., t. Tsai W., Xiao B., Chen Y., « UCSOA : User-Centric Service-Oriented Architecture », *2006 IEEE International Conference on e-Business Engineering (ICEBE'06)*, p. 248-255, Oct, 2006.

Faieq S., Saidi R., Elghazi H., Rahmani M. D., « C2IoT : A framework for Cloud-based Context-aware Internet of Things services for smart cities », *Procedia Computer Science*, vol. 110, p. 151 - 158, 2017a.

Faieq S., Saidi R., Elghazi H., Rahmani M. D., « A Conceptual Architecture for a Cloud-Based Context-Aware Service Composition », *Advances in Ubiquitous Networking 2*, vol. 397 of *Lecture Notes in Electrical Engineering*, p. 235–246, 2017b.

He W., Ren G., Cui L., Li H., « User Behavioral Context-Aware Service Recommendation for Personalized Mashups in Pervasive Environments », *Web Technologies and Applications*, Springer International Publishing, p. 683–694, 2015.

Labbaci H., Medjahed B., Binzagr F., Aklouf Y., « A Deep Learning Approach for Web Service Interactions », *Proceedings of the International Conference on Web Intelligence*, WI '17, ACM, New York, NY, USA, p. 848–854, 2017.

Sheng Q. Z., Qiao X., Vasilakos A. V., Szabo C., Bourne S., Xu X., « Web services composition : A decade's overview », *Information Sciences*, vol. 280, p. 218 - 238, 2014.

Tsai W. T., Xiao B., Paul R. A., Chen Y., « Consumer-centric service-oriented architecture : a new approach », *The Fourth IEEE Workshop on Software Technologies for Future Embedded and Ubiquitous Systems, and the Second International Workshop on Collaborative Computing, Integration, and Assurance*, April, 2006.

Xiao H., Zou Y., Tang R., Ng J., Nigul L., « Ontology-driven service composition for end-users », *Service Oriented Computing and Applications*, vol. 5, n° 3, p. 159, Mar, 2011.

Zhang Y., Zhang M., Zheng X., Perry D. E., « Service2vec : A Vector Representation for Web Services », *2017 IEEE International Conference on Web Services (ICWS)*, p. 890-893, June, 2017.

Modélisation des (méta)données hétérogènes et filtrage des contenus de vidéosurveillance : application au Forensic

Franck Jeveme Panta

Université de Toulouse III, Paul Sabatier
Institut de Recherche en Informatique de Toulouse (IRIT), CNRS
118 Route de Narbonne
31062 TOULOUSE CEDEX 9

Franck.Panta@irit.fr

MOTS-CLÉS : *Vidéosurveillance, métadonnées, filtrage, hétérogénéité, forensic.*

KEYWORDS: *CCTV system, metadata, filtering, heterogeneity, forensic.*

ENCADREMENT. *Florence Sèdes (PR) et André Péninou (MCF)*

1. Contexte

Au cours des dernières années, le taux de criminalité a augmenté très rapidement dans le monde. Le nombre de caméras de vidéosurveillance censées assurer la sécurité des citoyens dans les grandes villes croit régulièrement et de manière significative. Ces caméras génèrent une masse importante de données vidéos à analyser afin d'élucider les différents crimes (terrorisme, homicide, enlèvement, ...) survenus. Le processus actuellement utilisé par les enquêteurs (en France), consiste à appliquer des traitements vidéo (détection de visages, détection de véhicules, détection et lecture des plaques d'immatriculation...) sur l'ensemble des séquences vidéos récupérées en amont et préalablement indexées (conversion au bon format, géolocalisation des caméras, gestion des horodatages...). Les traitements sont appliqués sur l'ensemble des vidéos ou uniquement sur un sous-ensemble géographique ou temporel. Après ces traitements et une extraction des données, les enquêteurs effectuent des recherches (véhicules, plaques, personnes, visages...), visionnent les séquences d'intérêt et lancent de nouveaux traite-

ments. Les affaires récentes (terrorisme, enlèvement, homicide) ont nécessité l'analyse de plusieurs dizaines de milliers d'heures de vidéo. Le gain de temps fourni par les outils d'analyse actuels reste insuffisant dans un contexte opérationnel. Beaucoup des vidéos analysées pour une enquête s'avèrent inexploitables (par ex., problème de luminosité) pour l'enquête. Écarter ces séquences inéligibles permettra d'optimiser le temps d'exploitation. Cette thèse a pour objectif de proposer une (méta)modélisation et les mécanismes afférents de filtrage de grandes collections de vidéo liées à la recherche de preuves à posteriori. Cette contribution basée sur les (méta)données passe par la révision de la norme ISO 22311 qui décrit les exigences fonctionnelles à satisfaire par les systèmes de vidéosurveillance.

2. État de l'art

L'analyse vidéo automatique est actuellement confrontée à trois défis majeurs : (i) problème de gestion du grand nombre de caméras largement distribuées, (ii) difficulté de stockage et d'organisation de l'énorme quantité de données générées par ces caméras, (iii) analyse, annotation et recherche demeurent très coûteuses en temps.

De nombreux travaux ont été mené dans le but de proposer des solutions aux problèmes d'analyse vidéo ; notre objectif n'est pas d'en faire un exposé exhaustif ni d'apporter une contribution à l'analyse de contenu proprement dite : le verrou adressé est celui qui consiste à enrichir la modélisation des (méta)données, actuellement pas ou peu exploitées dans ce domaine.

Une méthode efficace et efficiente pour la construction et l'extraction des métadonnées descriptives des vidéos Web est présentée dans (Algur SP, et al., 2014) qui propose un modèle de métadonnées descriptives pouvant faciliter le traitement des contenus vidéos. Ce modèle est orienté classification des vidéos et nécessite d'autres spécifications pour être utilisé dans notre contexte. Un framework utilisant des informations géographiques pour la récupération des vidéos sur le web est présenté dans (Han Z. et al., 2016). Les auteurs décrivent et utilisent des données géographiques liées à la vidéo, telles que la localisation vidéo, le champ de vue de la caméra et les trajectoires pour effectuer des recherches dans les vidéos.Cependant, la qualité d'images/vidéos n'est pas prise en compte.

Les récents travaux de notre équipe se sont focalisés sur la modélisation des métadonnées pour la gestion distribuée des documents multimédias et l'interopérabilité des systèmes de vidéo surveillance. (Manzat, 2013) propose un format de métadonnées associées à tout type de contenu multimédia et permettant à plusieurs systèmes d'échanger des données et de les utiliser sans effectuer de manipulations spéciales. Ce format de métadonnées a été validé dans le projet LINDO[1]. Une étude des métadonnées utiles pour la recherche dans les collections de vidéos est présentée dans (Codreanu, 2015). Cette étude a été appliquée dans le cadre du projet ANR METHODEO[2]. L'approche

1. http ://lindo-itea.eu/
2. http ://www.agence-nationale-recherche.fr/Projet-ANR-10-SECU-0006

utilisée dans (Codreanu et al., 2015) vise à aider les opérateurs humains de vidéosurveillance dans la recherche de séquences vidéo d'intérêt en leur proposant un ensemble de caméras susceptibles d'avoir filmé une scène recherchée. Une extension de cette approche est décrite dans (Panta et al,2016).

Globalement, les solutions proposées dans la littérature utilisent des approches d'interrogation de contenus basées sur des métadonnées (techniques, liées aux caméras, spatio- temporelles) mais ne proposent pas d'interrogation basée sur le contenu via les descripteurs de qualité images/vidéos ou d'autres caractéristiques liées aux contenus.

3. Problématique

Un rapide état de l'art montre donc la nécessité d'exploiter les informations concises des contenus vidéo à partir de tous les éléments disponibles, a priori ou a posteriori en fonction des (méta)données potentiellement élicitables. Ces informations constituent des métadonnées issues d'algorithmes d'analyse de contenu (caractéristiques de bas niveau, présence et/ou nombre de personnes dans la scène, ...) et peuvent être modélisées selon différents niveaux de sémantique et de granularité. La collaboration des différents niveaux de métadonnées est un verrou essentiel dans notre approche. Cette approche consiste à effectuer un filtrage de contenus vidéo avec le meilleur compromis qualité/précision/temps de réponse, en combinant les métadonnées techniques, les métadonnées décrivant le mouvement et le champ de vue de la caméra, les métadonnées issues d'algorithmes d'analyse de contenu, les métadonnées de contexte (réseaux sociaux, open data, gélocalisation indoor/outdoor, jour/nuit). Une telle approche s'apparente à un problème de modélisation multicouches et est caractérisée par l'hétérogénéité des données dont elle est issue.

Nous souhaitons lever les verrous relatifs à (i) la proposition de "descripteurs" de qualité, (ii) l'interopérabilité des métadonnées associées aux contenus vidéo, (iii) la recherche dans de grands volumes de données vidéo sans recourir aux contenus intégraux, ces problématiques visant à mettre en oeuvre le filtrage de contenus pertinents pour les enquêteurs.

4. Actions réalisées

Pour lever les verrous mentionnés dans la problématique, notre méthodologie consiste dans un premier temps à faire une étude des métadonnées de contenus utiles pour le filtrage négatif/positif des contenus vidéo. Cette tâche consiste à extraire les caractéristiques des contenus vidéo et les modéliser selon les différents niveaux de granularité (frame, segment, vidéo) et de sémantique.

Notre travail s'applique dans le contexte de la recherche des preuves à posteriori (forensic) et se déroule dans le cadre du projet ANR FILTER2[3]. Les caractéristiques sont

3. FILtrage négaTif des contEnus de vidéopRotection

extraites des vidéos en fonction des différents traitements effectués dans les enquêtes policières. Trois traitements sont retenus dans le cadre de FILTER2 : la détection de visages, la détection et la lecture automatique des plaques, et la détection des véhicules. Nous avons donc défini les différents niveaux de métadonnées de contenus suivants : (i) les métadonnées liées aux pixels (Low level) : éclairage, luminosité, contraste, taux de bruit, taux de mouvement, etc. ; (ii) les métadonnées liées aux objets (Medium level) : présence des personnes et véhicules, nombre de personnes, etc. ; (iii) les métadonnées liées à la description de la scène (High level) : intempéries, nuit/jour ,etc. Nous proposons actuellement un modèle de données intégrant toutes ces métadonnées de contenus et celles mentionnées dans l'état de l'art.

5. Actions futures

Dans le but de mettre en place le filtrage négatif/positif des contenus, les prochaines directives de travail sont : (i) la définition des critères de qualité et d'utilisabilité des vidéos grâce à la combinaison des différentes métadonnées, (ii) la généralisation des approches de modélisation des métadonnées, (iii) la mise en oeuvre des algorithmes de requêtage basés métadonnées pour le filtrage des contenus.

Bibliographie

Algur SP, Bhat P, Jain S. Metadata Construction Model for Web Videos : A Domain Specific Approach. International Journal of Engineering and Computer Science. 2014 Dec 28 ;3(12).

Manzat, Ana-Maria. "Contribution à la modélisation des métadonnées associées aux documents multimédias et à leur enrichissement par l'usage." PhD diss., École Doctorale Mathématiques, Informatique et Télécommunications (Toulouse) ; 142547247, 2013.

Codreanu, Dana. "Modélisation des métadonnées spatio-temporelles associées aux contenus vidéos et interrogation de ces métadonnées à partir des trajectoires hybrides : application dans le contexte de la vidéosurveillance." PhD diss., Université de Toulouse, Université Toulouse III-Paul Sabatier, 2015.

Codreanu D, Peninou A, Sedes F. Video Spatio-Temporal Filtering Based on Cameras and Target Objects Trajectories–Videosurveillance Forensic Framework. InAvailability, Reliability and Security (ARES), 2015 10th International Conference on 2015 Aug 24 (pp. 611-617). IEEE.

Panta FJ, Sèdes F. Mobile objects in indoor environment : Trajectories reconstruction. InProceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media 2016 Nov 28 (pp. 332-336). ACM.

Han Z, Cui C, Kong Y, Qin F, Fu P. Video data model and retrieval service framework using geographic information. Transactions in GIS. 2016 Oct 1 ;20(5) :701-17.

Elaboration d'un Data Warehouse à partir d'un Data Lake

Rabah TIGHILT FERHAT

Institut de Recherche en Informatique de Toulouse (IRIT), Université Toulouse 1 Capitole
2 Rue du Doyen-Gabriel-Marty
31042 Toulouse

rabah.tighilt.ferhat@gmail.com

MOTS-CLÉS : *Réservoir de données, Entrepôt de données, Données massives, Processus décisionnel, Métadonnées, Extraction-Transformation-Chargement.*

KEYWORDS: *Data Lake, Data Warehouse, Big Data, Decision support system, Metadata, Extracting-Transforming-Loading.*

ENCADREMENT. *Gilles ZURFLUH*

1. Contexte

Le développement des bases de données massives (Big Data) pose plusieurs problèmes. Nous citons par exemple : la gestion des données très variées pour fournir de la connaissance. De nouveaux systèmes sont récemment apparus comme une solution à ce problème (Hai et al., 2016). Il s'agit des systèmes appelés « Data Lake » ou « réservoir de données ». Un Data Lake (DL) est un référentiel de stockage et d'exploration de grandes quantités de données brutes peu ou pas structurées permettant d'acquérir de la connaissance (Chessell et al., 2014). Les DL intègrent des données dans leur format d'origine à partir de sources de type Big Data (Hai et al., 2016), (Walker et al., 2015). Généralement, les données d'un Data Lake sont décrites par des métadonnées et organisées d'une certaine manière pour qu'elles soient facilement accessibles à tout moment et à tout utilisateur autorisé à effectuer des activités analytiques (Terrizzano et al., 2015). Notre travail s'intègre dans ce contexte et concerne particulièrement l'élaboration d'un ED à partir d'une source de type Data Lake.

Un ED est une base de données conçue pour supporter des processus d'aide à la décision. Dans un ED, les données sont souvent organisées selon un modèle multidimensionnel (en étoile, en flocon ou en constellation). Dans ce type de modèles, seules les données structurées (nombres et chaînes de caractères) sont utilisées. Or, dans un Data Lake, les données sont généralement complexes (textes brutes, images, fichiers audio et vidéo, etc.). D'où l'intérêt d'adapter les ED pour prendre en considération les caractéristiques d'un DL notamment la variété des données.

Pour illustrer notre travail, nous utilisons le cas des dossiers médicaux partagés (DMP) inspiré de l'application nationale gérée actuellement par la Caisse Primaire d'Assurance Maladie (CPAM). Un dossier médical contient les données de santé et le parcours de soins d'une personne (assuré social). Il permet au médecin traitant et aux autres professionnels de santé (médecins, infirmiers, dentistes, protection civile, etc.), d'accéder et de partager des informations médicales concernant cette personne. Il doit permettre aussi aux médecins de visualiser l'historique médical d'une personne. En outre, le DMP permet aux analystes de la CPAM de traiter les données des patients ou de les confronter avec celles d'autres malades dans le but d'en avoir une meilleure connaissance des dépenses de santé.

2. État de l'art

Dans le contexte des ED, les auteurs de (Bala et al., 2013) ont proposé une approche de modélisation des Big Data dans un système décisionnel. Ils s'intéressent particulièrement à repenser et à adapter les processus ETL (Extracting, Transforming, Loading) à l'aide du modèle MapReduce pour la capture des données massives dans un ED. Par ailleurs, pour construire des ED massifs, les auteurs (Chevalier et al., 2015) ont défini des règles pour traduire un modèle multidimensionnel en étoile, en deux modèles physiques NoSQL, un modèle orienté colonnes et un modèle orienté documents. Les liens entre faits et dimensions ont été traduits sous la forme d'imbrications. (Dehdouh et al., 2015) traitent de l'implantation des entrepôts de données de grande taille (Big Data). Le processus d'implantation repose sur une architecture à 3 niveaux : conceptuel, logique et physique. Ils proposent trois approches permettant de construire des ED volumineux en utilisant le modèle NoSQL orienté colonnes. Les trois approches utilisent les spécificités des systèmes Big Data en privilégiant certains choix de stockage des faits et des dimensions. (Li et al., 2010) ont étudié les mécanismes d'implantation d'une base de données relationnelle dans le système HBase ; le but est d'élaborer un entrepôt de données NoSQL orienté colonnes. La méthode proposée est basée sur des règles de correspondance entre un schéma relationnel et un schéma HBase. Les relations entre les tables (clés étrangères) sont traduites par l'ajout des familles de colonnes contenant des références.

A notre connaissance, peu de travaux ont proposé des approches pour le stockage et l'exploration des Data Lake. Les auteurs de (Walker et al., 2015) ont proposé un système de stockage et de gestion des Data Lake. Le remplissage et l'interrogation des Data Lake reposent sur une gestion des métadonnées pour les données structurées

(relationnelles) et semi- structurées (XML, CSV, JSON, etc.). Pour les données non structurées (textes, images, fichiers audio et vidéo, etc.), les auteurs ont défini une « Mesure de gravitation » pour attribuer une densité à ces données. Dans (Hai et al., 2016), il a été proposé un système appelé « Constance », qui permet de stocker et de traiter des données dans un Data Lake. Constance se focalise particulièrement sur la gestion des métadonnées explicites et implicites. L'interrogation de données se base sur un langage de requêtes par mots-clés. De plus, Constance fournit une interface qui permet aux utilisateurs de contrôler la qualité des données en choisissant des métriques de qualité des données définies au préalable.

Il apparait à travers cet état de l'art que peu de travaux ont étudié la construction d'un ED à partir d'une source de données massives. Dans les travaux les plus proches de notre problématique notamment les travaux de (Bala et al., 2013), (Chevalier et al., 2015), (Dehdouh et al., 2015) et (Li et al., 2010) ; généralement, le schéma de données Big Data est défini préalablement (i.e. avant de commencer la saisie). Dans un Data Lake, le schéma n'est connu que partiellement. Ceci est dû au fait que les traitements décisionnels ne sont pas connus au moment du stockage.

3. Problématique

Notre problématique consiste à proposer un outil capable d'extraire les données d'un Data Lake puis de les restructurer dans le but d'effectuer des traitements décisionnels. En effet, les Data Lake ont deux problèmes majeurs, structurel et sémantique : (1) Problème structurel : le schéma de données dans un DL n'est connu que partiellement. L'absence du schéma est dû au fait que les traitements décisionnels ne sont pas définis préalablement, (2) Problème sémantique : lors de la création des DL, la sémantique des données est peu connue ; cela est en raison de l'absence d'une gestion efficace des métadonnées. Par conséquent, l'interrogation des DL s'avère être difficile. Ainsi, nous visons à proposer une approche pour la construction d'un ED à partir d'un Data Lake en résolvant à la fois le problème structurel et le problème sémantique des données. Nous proposons une approche pour la construction d'un ED à partir d'un Data Lake.

Notre problématique générale se résume à travers les trois questions suivantes :

– Comment obtenir le schéma des données brutes ?

– Comment relier les éléments du schéma sur un plan sémantique ?

– Comment restructurer les données brutes dans un ED ?

4. Actions réalisées

Comme actions réalisées, nous avons (1) commencé l'implantation de l'étude de cas présentée dans la section contexte, sur un cluster de 3 machines. Chaque machine est de type Intel Core i5, 8 Go de RAM et 2 To de disque. L'une de ces machines est configurée pour agir comme maitre et les autres comme esclaves, (2) fait un état de

l'art sur les systèmes de gestion des données massives, en particulier les systèmes Data Lake.

5. Actions futures

Nous allons proposer des solutions aux sous problèmes considérés dans la section problématique. D'une part, nous visons à pouvoir extraire le schéma de données en nous basant sur les métadonnées explicites et implicites. Pour ceci, le travail (Hai et al., 2016) peut nous être utile. D'autre part, nous allons proposer un outil permettant de restructurer les données en fonction des besoins des décideurs.

Bibliographie

Chessell, Mandy, Scheepers, Ferd, Nguyen, Nhan, et al.Governing and managing big data for analytics and decision makers. IBM Redguides for Business Leaders, 2014.

Terrizzano, Ignacio G., Schwarz, Peter M., Roth, Mary, et al. Data Wrangling : The Challenging Yourney from the Wild to the Lake. In : CIDR. 2015.

Hai, R., Geisler, S., Quix, C. (2016, June). Constance : An intelligent data lake system. In Proceedings of the 2016 International Conference on Management of Data (pp. 2097-2100). ACM.

Walker, Coral et Alrehamy, Hassan. Personal data lake with data gravity pull. In : Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on. IEEE, 2015. p. 160-167.

Bala, Mahfoud et Alimazighi, Zaia. Modélisation de processus ETL dans un modèle MapReduce. In : Conférence Maghrébine sur les Avancées des Systèmes Décisionnels (ASD'13). 2013. p. 1-12.

Chevalier, M., M. E. Malki, A. Kopliku, O. Teste, et R. Tournier (2015). Entrepôts de données multidimensionnelles nosql. EDA.

Dehdouh, Khaled, et al. "Using the column oriented NoSQL model for implementing big data warehouses." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015.

Li, Chongxin. Transforming relational database into HBase : A case study. In : Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on. IEEE, 2010. p. 683-687.

Accélération par pré-agrégations des accès complexes et répétitifs aux Big Data

Nabil El malki

Institut de Recherche en Informatique de Toulouse, Univ. Toulouse III Paul Sabatier
118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9

Nabil.El-Malki@irit.fr

MOTS-CLÉS : *apprentissage automatique, analyse des données, agrégation, k-moyennes*

KEYWORDS: *machine learning, big data analytics, k-means, aggregation*

ENCADREMENT. *Olivier Teste et Franck Ravat*

1. Contexte

L'humanité produit des quantités de données numérisées dans des proportions et avec un rythme sans commune mesure avec le passé. Ces masses de données, désignées communément comme Big Data, sont entreposées dans des clusters de stockage où les données sont plus ou moins structurées. Ces masses de données sont ensuite exploitées par des analystes (« data scientists ») qui utilisent des chaînes complexes de traitements, afin d'extraire les phénomènes contenus dans les masses de données. Ces traitements consistent à explorer les données, les classifier suivant des approches supervisées, semi-supervisées ou encore non supervisées. Par exemple, les données radars dans l'aviation civile sont stockées sous forme binaire par traces radars reconstituées. Un accès répétitif consiste à extraire toutes les trajectoires dans une fenêtre spatio-temporelle de l'espace aérien 3D. Un tel traitement réclame de nombreux accès aux données brutes pour constituer une réponse. L'aspect répétitif est induit notamment lorsque plusieurs requêtes demandent des calculs élémentaires communs répétés sur les données brutes. Par exemple, calculer le nombre de trajectoires par semaine dans une fenêtre de l'espace aérien revient à agréger 7 calculs de trajectoires quotidiennes.

2. État de l'art

Dans cette section nous nous intéressons uniquement aux travaux relatifs aux accès complexes et répétitifs appliqués à des données massives.

Dans le domaine de l'analyse statistique, les auteurs des travaux de (Wasay *et al.*, 2017) proposent un système baptisé Datacanopy qui repose sur un cache intelligent destiné à l'analyse statistique exploratoire. Datacanopy pré-calcule et pré-agrège des calculs statistiques pour éviter les accès répétitifs aux données de base. Pour ce faire, il décompose les calculs en opérations élémentaires et les données de base en blocs unitaires (chunks). Ces chunks sont stockés dans un arbre binaire agrégeant de manière récursive les calculs des feuilles jusqu'à la racine. La construction d'un tel arbre dépend des requêtes utilisateur. Cette contrainte conduit l'approche Datacanopy à supporter uniquement les requêtes respectant cet ordonnancement.

Dans le domaine de l'apprentissage automatique, les méthodes, telles que k-means, ont recours à des accès répétitifs aux données de base. Le but de k-means est de diviser l'ensemble des individus Xi en un certain nombre de classes homogènes défini préalablement par un utilisateur. Cette méthode repose sur un agorithme itératif consistant à intégrer ou déplacer des points dans des classes. L'utilisation de la méthode k-means nécessite un temps d'exécution proportionnel au produit du nombre de classes et du nombre de points par itération. Ce temps d'exécution total est coûteux en terme de calcul, en particulier pour les grands ensembles de données. Par conséquent, l'algorithme de clustering k-means ne peut pas satisfaire le besoin en temps de réponse rapide pour certaines applications. K-means (Lloyd, 1982), effectue des répétitions des calculs de distance et de moyenne sur les mêmes blocs de données. Plusieurs extensions de la version standard du k-means ont été proposées pour accélérer les temps d'exécution :

– Accélération par la parallélisation de l'algorithme via le MapReduce (Li *et al.*, 2015) ou le MPI (Zhang *et al.*, 2013) qui sont des modèles de programmation conçus pour traiter des volumes importants de données de manière parallélisée et distribuée.

– Accélération par réduction du nombre de calculs à effectuer. Les algorithmes d'Elkan (Elkan, 2003) et de Hamerly (Hamerly, 2010) se basent sur la propriété de l'inégalité triangulaire pour éviter de calculer à chaque itération la distance entre un point donné et tous les centres de gravité des classes.

– Accélération par organisation ou structuration des données. Dans les travaux (Hung *et al.*, 2005), les auteurs proposent un algorithme accélérant k-means par découpage du jeu de données en blocs unitaires égaux. Ceux-ci contiennent au moins un individu. Ensuite k-means est déroulé sur les centres de gravité de ces blocs et non sur les points contenus.

Ces extensions accélèrent en temps d'exécution k-means standard mais n'utilisent pas l'approche de pré-agrégation des résultats intermédiaires utilisés dans d'autres contextes et qui pourrait offrir des perspectives d'amélioration intéressantes.

3. Problématique

Les chaînes de traitements induisent des accès parfois complexes et répétitifs aux données, alourdissant sensiblement ces traitements (Wasay *et al.*, 2017). La répétition est provoquée par les méthodes qui sont généralement peu optimisées. Ainsi ils ne conservent pas les calculs communs entre deux requêtes. Par exemple, calculer une variance et une covariance sur une même colonne numérique nécessite de calculer à chaque fois la somme des valeurs contenues dans la colonne. La complexité d'accès est due à la multiplicité des architectures de stockage du big data (BD Nosql, polystore...)(Chevalier *et al.*, 2015). L'objectif de nos travaux est de proposer une approche pour optimiser les accès répétitifs, en considérant différentes architectures, basée sur des pré-calculs agrégeant les calculs intermédiaires répétés.

4. Actions réalisées

Dans un premier temps, nous nous intéressons aux algorithmes d'apprentissage de données (machine learning), plus particulièrement, à l'algorithme de k-means, l'un des algorithmes de clustering couramment utilisé. Nous avons mené deux actions visant à répondre au problème de la répétitivité des accès dans k-means :

– Proposition d'une structure arborescente stockant des pré-agrégats (des moyennes) pour être utilisé par k-means. Lors de l'exécution de ce dernier, il n'effectue pas d'opérations de moyennes mais il récupère les résultats des opérations depuis une structure de données arborescente. Cette solution permet à k-means de ne pas parcourir les données de base pour calculer les moyennes pour réduire le temps de son exécution. La structure est une composition de plusieurs sous-structures dans lesquelles chacune ne stocke que les moyennes des ensembles ayant le même nombre d'entités de base. Les sous-structures dites M2 (regroupement de deux entités de base) et M3 (regroupement de trois entités de base) sont calculées à partir des données de base. Par contre la sous-structure M4 est calculée à partir de M2, celle de M5 à partir de M2 et M3, celle de M6 à partir de M3 et ainsi de suite. Il existe plusieurs chemins de constructions des sous-structures, à titre d'exemple M6 peut être construite de deux M3 ou bien de trois M2.

– L'autre action est la réduction de l'espace de calcul des moyennes, c'est-à-dire que l'on ne calcule a priori que les moyennes susceptibles d'être requises par k-means.

5. Actions futures

L'objectif du projet de thèse est d'aller au-delà de la simple utilisation des outils existants permettant d'explorer et d'analyser les masses de données. La thèse abordera plusieurs aspects scientifiques non résolus dans la littérature :

– Modélisation des données. Pour permettre la manipulation des données, il conviendra dans un premier temps de définir le modèle de représentation des données.

L'enjeu de cette phase est notamment de prendre en compte la grande variabilité des données dans le contexte des Big Data (approche « schemaless »). La modélisation des données a pour enjeu de permettre la définition rigoureuse des composants de base manipulés par les requêtes des utilisateurs, et de définir les mécanismes garantissant des pré-agrégations cohérentes de ces composants.

– Noyau algébrique d'opérateurs élémentaires. Ces accès seront décomposés sur la base d'un noyau algébrique d'opérations élémentaires additives permettant de modéliser la chaîne de traitements complexes par composition de ces opérations élémentaires. Les calculs élémentaires répétés pourront alors être pré-calculés par le système de gestion du Big Data accédé sous la forme de pré-agrégats.

– Apprentissage automatique des pré-agrégats. Des mécanismes d'apprentissage automatique pourront être introduits dans le système. Le but de cet apprentissage sera de permettre une maintenance prédictive des pré-agrégats en fonction de l'évolution des traitements effectués par les utilisateurs. Ces mécanismes confèreront au système des capacités d'adaptation automatique en fonction de l'évolution des traitements mais également de la masse de données.

Nos expérimentations pourront s'appuyer sur la plateforme OSIRIM de l'IRIT, offrant une baie de stockage massif (36 disques de 3To) et d'un cluster de calcul de 640 cœurs, étendus par cartes GPU.

6. Bibliographie

Chevalier M., El Malki M., Kopliku A., Teste O., Tournier R., "Implementation of multidimensional databases with document-oriented NoSQL", *Int. Conf. on Big Data Analytics and Knowledge Discovery, Dawak'15*, p. 379–390, 2015.

Elkan C., "Using the Triangle Inequality to Accelerate k-Means", *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*p. 147–153, 2003.

Hamerly G., "Making k -means even faster", *2010 SIAM international conference on data mining (SDM 2010)*p. 130–140, 2010.

Hung M.-C., Wu J., Chang J.-H., "An Efficient k-Means Clustering Algorithm Using Simple Partitioning", *Journal of Information Science and Engineering 21*, vol. 1177, p. 1157–1177, 2005.

Li Z., Song X., "K-means Clustering Optimization Algorithm Based on MapReduce", , , p. 198–203, 2015.

Lloyd S. P., "Least Squares Quantization in PCM", *IEEE Transactions on Information Theory*, vol. 28, n° 2, p. 129–137, 1982.

Wasay A., Wei X., Dayan N., Idreos S., "Data Canopy", *Proceedings of the 2017 ACM International Conference on Management of Data - SIGMOD '17*p. 557–572, 2017.

Zhang J., Wu G., Hu X., Li S., Hao S., "A Parallel Clustering Algorithm with MPI – MKmeans", *Journal of Computers*, vol. 8, n° 1, p. 10–17, 2013.

Approche Big Data sur un réseau centré sur la donnée

Exemple du réseau NDN (Named Data Networking)

Junior Dongo

Université Paris-Est - Laboratoire d'Algorithmique, Complexité et Logique
61 Avenue du Général de Gaule
94010 Créteil

junior.dongo@u-pec.fr

MOTS-CLÉS : *Big data, Réseau centré données, Réseau par nommage de données, système de fichiers distribués, stratégie de réplication.*

KEYWORDS: *Big data, Content Centric Networking, Named Data Networking, Distributed File System, replication strategy*

ENCADREMENT. : *Fabrice Mourlin et Charif Mahmoudi*

1. Contexte

Dans la communauté Big Data, le squelette MapReduce a été considéré comme l'une des principales approches permettant de répondre à une demande permanente et croissante des ressources informatiques imposées par des données massives. Son importance s'explique par l'évolutivité du paradigme MapReduce qui permet une exécution massivement parallèle et distribuée sur un grand nombre de nœuds de calcul. L'émergence de nouvelles architectures réseaux telles que Content-Centric Networking (CCN) offrent des perspectives nouvelles pour déléguer le support du Big Data depuis la couche applicative vers la couche réseau. L'une des architectures réseaux CCN dominantes est Named Data Networking (NDN) (Zhang *et al.*, 2010) qui est une architecture centrée sur la donnée. Nous passons donc des communications d'hôte à hôte, à un modèle de communication basé sur la donnée.

NDN est financé par la Fondation Nationale Américaine pour la science (NSF) dans le cadre du projet Future Internet Architecture (FIA). Il dispose d'une spécification solide en plus de plusieurs implémentations et d'un déploiement sous forme d'un réseau universitaire multinational.

2. État de l'art

Le traitement Big Data peut être considéré en deux grandes phases : le stockage des données sur un Système de Fichiers Distribués (DFS) et le lancement de calculs distribués sur les données. Différentes approches de calcul distribué ont été proposées dans le cadre des traitements Big Data, le plus utilisé étant le MapReduce (Zhao *et al.*, 2009). Généralement les données sont stockées sur le système de fichiers distribués Hadoop (HDFS (Shvachko *et al.*, 2010)) et les calculs sur les données effectués en utilisant MapReduce. Une approche Big Data centrée sur la donnée a été proposée par les travaux réalisés à l'Université de L'Arizona (Gibbens *et al.*, 2017). Leur approche consiste en un portage de Hadoop sur NDN. En effet NDN étant basé sur la donnée et non sur l'IP, une profonde modification de Hadoop a été nécessaire afin de permettre une exécution sur NDN. La limitation de Hadoop concernant le point unique de défaillance reste d'actualité avec cette approche.

Une approche de système de fichiers distribués a été proposée pour les réseaux NDN en vue de faciliter les traitements Big Data(Chen *et al.*, 2015). Cette approche n'utilise pas de réplication de données, qui est d'une grande importance dans le calcul Big Data, et nécessite une profonde refonte de l'architecture NDN.

3. Problématique

Dans NDN, nous avons deux types de composants : des Consumers et Producers. Un Producer, produit la donnée. Il expose le préfix des noms de données qu'il est capable de servir. Un Consumer envoie un intérêt pour une donnée en la nommant. Chaque élément de données est identifié par un nom unique. Les 2 types de paquet sur NDN sont : Interest et Data, comme le montre la structure de la Figure 1). Aucune information concernant la source ou la destination d'une donnée n'est contenu dans un intérêt ou une data. NDN permet la mise en cache de la donnée sur le réseau. Chaque paquet étant signé lors de sa création, l'intégrité des données peut être facilement vérifiée par n'importe quel nœud.

L'utilisation du mode actuel de communication d'hôte à hôte, laisse apparaître des problèmes. Par exemple dans le cadre d'une réplication de données, si nous avons 3 nœuds qui doivent répliquer une donnée depuis 1 nœud source, cela produira 3 paquets de la source vers les 3 destinations. Il n'y a donc pas d'optimisation de paquet, vu qu'en l'espèce, il s'agit de la même donnée.

NDN a été proposé comme l'architecture Internet du Futur. Son adoption nécessite donc la proposition de solution Big Data basée sur cette architecture. Aussi, la majorité

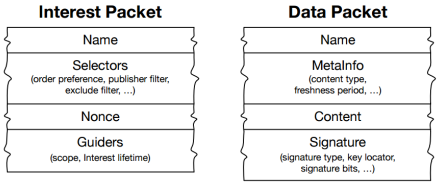

Figure 1. *Types de packets NDN*

des systèmes de fichiers actuels utilisent un composant central pour la gestion des données. Cela est le cas par exemple de Hadoop, où le NameNode stocke les métadonnées ainsi que l'arborescence de tous les fichiers du système de fichiers. Ainsi, en cas de perte du NameNode, il y a un risque de perte de toutes les données. Il représente donc un point unique de défaillance.

L'objectif est d'apporter une solution pleinement distribuée qui évite l'utilisation d'un composant central et qui permet une agrégation des paquets réseaux afin de réduire la consommation réseau.

4. Actions réalisées

Notre première action a consisté en la définition d'un système de fichiers distribués basé sur NDN. Ce DFS a les caractéristiques suivantes :

– Pleinement distribué : ne contient pas un unique point de défaillance ;

– Résilient : mécanisme de recouvrement intégré pour tous les composants ;

– Sécurisé : chaque donnée est signée et encryptée ;

– Adaptable : supporte les données de petite et de grande taille.

Notre système est composé de 3 composants :

– Client : utilisé par l'utilisateur pour initier une demande de réplication. La donnée est découpée en segment et rendue disponible sur le réseau. Un intérêt pour une demande de réplication est émis à destination des nœuds de stockage.

– Storage : responsable du stockage des réplicas. Lors du traitement d'une demande de réplication, si un autre réplica est nécessaire, le storage prend en charge la tâche d'effectuer la demande de réplication.

– Heartbeat : Un mécanisme utilisé pour vérifier la disponibilité des réplicas sur le réseau. L'algorithme proposé dans notre approche, est un algorithme dans lequel un nœud vérifie un autre nœud. Cela permet le maintien d'un pointeur circulaire sur les nœuds répliquant la donnée. Cela aide à avoir cet aspect complètement distribué. En effet, les réplicas n'ont pas à se signaler périodiquement à un composant central.

Une implémentation de l'approche en utilisant le simulateur réseau ndnSim (Afanasyev *et al.*, 2012) basé sur ns-3 (Lacage *et al.*, 2006), nous a permis lors de simulation d'obtenir les résultats suivants :

– une augmentation du nombre d'utilisateurs permet de réduire le temps nécessaire pour récupérer une donnée. Cela s'explique par l'utilisation de cache réseau au niveau de NDN.

– l'impact du facteur de réplication sur la demande des données depuis la source est linéaire. Cela signifie que les intérêts sont agrégés par NDN. Cela montre l'efficience du système, car un nombre minimal de paquet est uniquement transmis vers la source.

– à partir d'un réplica disponible sur le réseau, le système est capable de se reconstruire à un état stable si des nœuds de stockage sont disponibles.

5. Actions futures

L'étude d'une approche de calcul distribué de type MapReduce basée sur notre approche de DFS sur NDN est en cours. L'idée est de distribuer le calcul, mais aussi d'éviter d'exécuter un même calcul plusieurs fois si celui-ci a déjà été exécuté par un autre nœud (mise à disposition des résultats des calculs intermédiaires). Nous ferons une analyse des résultats issus de l'implémentation cette approche dans le simulateur ndnSim.

Une implémentation et expérimentation de l'approche sur des machines physiques est prévue afin de confirmer les résultats obtenus, mais aussi une comparaison de ceux-ci avec ceux de Hadoop.

6. Bibliographie

Afanasyev A., Moiseenko I., Zhang L. et al., « ndnSIM : NDN simulator for NS-3 », *University of California, Los Angeles, Tech. Rep*, 2012.

Chen S., Cao J., Zhu L., « NDSS : A Named Data Storage System », *2015 International Conference on Cloud and Autonomic Computing*, p. 196-199, Sept, 2015.

Gibbens M., Gniady C., Ye L., Zhang B., « Hadoop on Named Data Networking : Experience and Results », *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, n° 1, p. 2 :1–2 :21, juin, 2017.

Lacage M., Henderson T. R., « Yet another network simulator », *Proceeding from the 2006 workshop on ns-2 : the IP network simulator*, ACM, p. 12, 2006.

Shvachko K., Kuang H., Radia S., Chansler R., « The hadoop distributed file system », *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, IEEE, p. 1–10, 2010.

Zhang L., Estrin D., Burke J., Jacobson V., Thornton J. D., Smetters D. K., Zhang B., Tsudik G., Massey D., Papadopoulos C. et al., « Named data networking (ndn) project », *Relatório Técnico NDN-0001, Xerox Palo Alto Research Center-PARC*, 2010.

Zhao J., Pjesivac-Grbovic J., « MapReduce : The programming model and practice », 2009. Tutorial.

Méthodologie et environnement pour le traitement de données appliquées aux Sciences Humaines et Sociales

Tiphaine VAN DE WEGHE

Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour
Avenue de l'Université
64000 PAU

t.van-de-weghe@univ-pau.fr

MOTS-CLÉS : *données, planification, modélisation, collecte, traitement, représentation*

KEYWORDS: *data,planning,modelization, collecting, processing, representing*

ENCADREMENT. *Philippe Roose (MCF) et Marie-Noëlle Bessagnet (MCF)*

1. Contexte

Les recherches génèrent de nouvelles informations. Si on prend le nombre moyen de projets par Université, multiplié par le nombre d'Universités en France, on est face à un grand volume de données. Imaginons donc ces éléments à l'échelle mondiale. En Sciences Humaines et Sociales (SHS), ces renseignements peuvent se multiplier très vite. Les recherches se font sur des supports retrouvés en archives, ou encore sur des études déjà effectuées, des événements étudiés etc. Les SHS regroupent des disciplines qui analysent les humains et les sociétés, qui ont existé et existent. Ces sciences possèdent une matière première diverse et complexe (images, vidéos, sons, textes non numérisés,etc.). Durant cette thèse, nous allons traiter plusieurs points, pendant lesquelles les données subirons des transformations. On dénombre parmi ces étapes : (i) la collecte de données, (ii) la modélisation, (iii) le stockage, (iv) le traitement/analyse de l'information (v) la valorisation. **L'objectif de cette thèse est de couvrir l'ensemble de ces phases et d'aider au mieux les chercheurs en SHS à valoriser leur données de la recherche.** Dans cet article, nous allons présenter l'intérêt d'une collaboration

entre les SHS et les Sciences Exactes, avec, pour fil conducteur l'intervention de l'informaticien pour la gestion des données, tout en faisant un bref état de l'art. Par la suite, nous pourrons présenter les problématiques autour des cinq points principaux (i à v) de cette thèse énumérés ci-dessus. Puis, nous déploierons les actions réalisées et futures, et nous conclurons.

2. État de l'art

L'OCDE[1] définit les données de la recherche comme " des enregistrements factuels (chiffres, textes, images et sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires à la validation des résultats de recherche ". L'Université Humboldt de Berlin (Schöpfel *et al.*, 2017) distingue deux types de données : les sources (brutes) et les résultats (traitées). Elles suivent l'acheminement du cycle de vie de la donnée qu'offre le schéma de UK Data Service. (Reymonet *et al.*, 2018) montrent qu'il y a plusieurs intervenants (archivistes, documentalistes, informaticiens, chercheurs, etc.) dans le plan de gestion des données. L'implication de l'informaticien et sa mise en relation avec le chercheur en SHS est incontournable. En effet, le chercheur en SHS possède des ressources ou supports informatisés ou non. L'informaticien l'aide à les structurer dans des bases de données, ou encore, aider à réaliser de l'extraction d'information semi-automatique, comme l'ont stipulé (Kergosien *et al.*, 2016). Le type de données que le chercheur en SHS possède est : des textes, des images, des enregistrements sonores, des vidéos etc. Mais, les études sur les données de la recherche des Universités de Lille 3 et Rennes 2 (Prost *et al.*, 2015) et (Serres *et al.*, 2017), montrent que les principales ressources des chercheurs en SHS sont des textes. Bien entendu, pour que l'informaticien intervienne, il est nécessaire que le matériel des chercheurs en SHS soient numérisées. Malheureusement, les documents textuels ne le sont pas systématiquement.

3. Problématique

Quels sont les besoins des chercheurs en SHS ? Qu'ont-ils comme support (ressources) ? Au regard de notre expérience, deux profils sont déterminés : Dans un premier cas, le chercheur a des connaissances en système d'information, tel que l'archéologue qui use des systèmes d'information géographique (qui est le plus utilisé et qui a la plus grande communauté). Il se construit sa propre base de données, et parfois il étudie lui-même en cartographiant ses données. Celles-ci sont souvent analysées statistiquement par le gestionnaire de données. D'un autre côté, il y a des chercheurs qui nécessitent un accompagnement complet, c'est-à-dire de l'étude des besoins à la valorisation des données. Ils ont des problématiques, des supports, et un savoir, mais ne savent pas comment valoriser et utiliser ces informations, à l'aide de l'outil

1. Organisation de Coopération et de Développement Économiques

informatique, ni d'ailleurs ce qu'un tel outil peut lui apporter. Notre rôle sera alors de comprendre et de modéliser ses éléments, en définissant des méthodes et outils qui les automatisent. Après avoir analysé leurs demandes, il faut penser à l'organisation, au tri de leurs données. Il s'agit de les structurer, de les enregistrer. L'utilisation d'une base de données est nécessaire. De plus, à la fin des projets de recherche, certaines valeurs doivent être codées et exportées vers des institutions pour expositions et/ou pérennisation. A partir de cette collection, qu'est-il possible de faire avec ces données ?

4. Actions réalisées

En tant qu'ingénieur d'étude au laboratoire ITEM EA 3002[2], j'ai pu observer et analyser les pratiques des chercheurs en SHS. Ce laboratoire de recherche est composé de 25 enseignants chercheurs (anthropologie, archéologie, histoire, histoire de l'art, et étude hispaniques) et d'une cinquantaine de doctorants. Les principaux axes de recherches sont :

– territoires, mobilités

– identités, patrimoines

– méthodologie de la recherche : "archives et corpus"

Cela m'a permis de recenser les pratiques des anthropologues, des archéologues et des historiens, en terme de collecte de données et surtout les formats de données utilisés par ces chercheurs. Actuellement, le principal comportement des chercheurs est la consultation de bases de données bibliographiques spécialisées, sites Internet, archives, etc. Nous avons recensé, dans le laboratoire, différents types de support qui sont : des textes, des images, des vidéos, des enregistrements sonores, des valeurs quantitatives, ou encore des données qualitatives. Les documents textuels du laboratoire sont généralement des lettres de correspondance, des articles, etc. Nous modélisons de façon à être interopérable dans le futur. C'est pour cela que nous avons également étudié la norme Dublin Core avant de déterminer des variables plus appropriées aux recherches. Le Dublin Core, comme le définit la BnF (Bibliothèque Nationale de France) est une norme internationale de variables (titre, créateur, langue, contributeur, etc.) qui est utilisée par de nombreuses institutions, ce qui facilite leur exportation. Dans certains cas, il est nécessaire de gérer des données sources autrement (collecte et stockage), puis, de respecter la norme Dublin Core, avec des données résultats (sélection des informations pertinentes pour le chercheur). Pour des raisons de confidentialité, parfois un stockage suffit pour des éléments sources, seuls les résultats sont importants (éléments sensibles). Cependant, pour une insertion des valeurs, les chercheurs doivent pouvoir les stocker afin de pérenniser leur travail.

---

2. Identités, Territoires, Expressions, Mobilités Équipe d'Accueil

5. Actions futures

Après la collecte des données suivent les phases de traitement et d'analyse qui permettront de représenter les données, de manière compréhensive par tous. Nous extrairons les informations répondant aux problématiques des chercheurs. Ces analyses entraînent un traitement avec des méthodes statistiques et/ou informatiques. Quelle sera alors leur complétude ? En effet, l'informatique permet d'extraire les informations des corpus selon 3 dimensions : thématique, spatiale et temporelle. Nous appliquerons notamment des méthodes liées à la fouille de texte, ainsi que des démarches statistiques et de la cartographie. Notre question de recherche sera alors : comment combiner judicieusement ces types d'analyses ? Les supports sont souvent présentés sous la forme d'anciens écrits. Le chercheur en SHS va tenter de récupérer une majorité de ces documents. Pour des raisons politiques ou des droits d'utilisation, il rencontre des difficultés à se les procurer. Ses recherches sont donc basées sur un minimum de textes de taille plus ou moins grande. Alors, comment adapter les méthodes de fouille de texte sur de tels corpus ? Les documents en langues étrangères et anciennes apparaissent comme un autre défi. Quels sont les méthodes et outils informatiques pour aider les chercheurs en SHS dans l'analyse de tels textes ? Le chercheur en SHS demande également des outils et méthodes pour une meilleure visualisation des résultats. Parfois sur des corpus conséquents, la recherche d'information sera développée. Est-ce un type de valorisation ? En quelque sorte, cela met en avant les données. Il existe, tout de même, d'autres techniques qui seront définies. En général, des publications, des applications, etc. sont la suite logique de la valorisation des données. Mais aussi, l'exportation des données sources vers d'autres établissements où la fusion se fait avec d'autres données de recherche. Chaque institution choisit son mode de structuration de données, avant de procéder à leurs insertions. Il est important de se renseigner sur leurs moyens de travail (langage utilisé) afin de s'adapter. Aujourd'hui, un projet de recherche doit faire face à cette exportation, qui se fera en XML (langage de balisage extensible).

Cette thèse a démarré en janvier 2018. Elle reprend les points du cycle de vie des données, en développant des automatismes qui correspondent au mieux aux attentes des chercheurs en SHS, après avoir fait une analyse des besoins. L'enquête tend vers des projets de recherches en SHS comme Acronavarre[3] (actes royaux de Navarre), le patrimoine d'encre pyrénéen et TCVPYR[4] (Thermalisme, Culture, Villégiature dans les Pyrénées). Des modèles de données ont été créés pour la collecte. Ces programmes de recherches fournissent la matière appropriée à cette expérimentation. Ces derniers demandent un travail complet sur les données, c'est-à-dire, de la planification à leur valorisation.

3. projet ANR : https ://acronavarre.hypotheses.org/
4. projet européen FEDER : http ://tcvpyr.iutbayonne.univ-pau.fr/

6. Bibliographie

Kergosien E., Bessagnet M.-N., Sallaberry C., Le Parc-Lacayrelle A., Royer A., « Analyse géographique de séries de publications : application aux conférences EGC », *EGC'2016 (Extraction et Gestion des Connaissances)*, p. 371–382, 2016.

Prost H., Schöpfel J., Les données de la recherche en SHS. Une enquête à l'Université de Lille 3., Technical report, Lille 3, 2015.

Reymonet N., Moysan M., Cartier A., Délémontez R., « Réaliser un plan de gestion de données « FAIR » : guide de rédaction », 2018.

Schöpfel J., Kergosien E., Prost H., « « Pour commencer, pourriez-vous définir'données de la recherche' ? » Une tentative de réponse », *Atelier VADOR : Valorisation et Analyse des Données de la Recherche ; INFORSID 2017*, 2017.

Serres A., Malingre M.-L., Mignon M., Pierre C., Collet D., Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2, Technical report, Université Rennes 2, 2017.