INFORSID 2016

34e édition - Grenoble

INFORSID

Actes du 8e Forum Jeunes Chercheurs du congrès INFORSID

31 mai 2016 à Grenoble

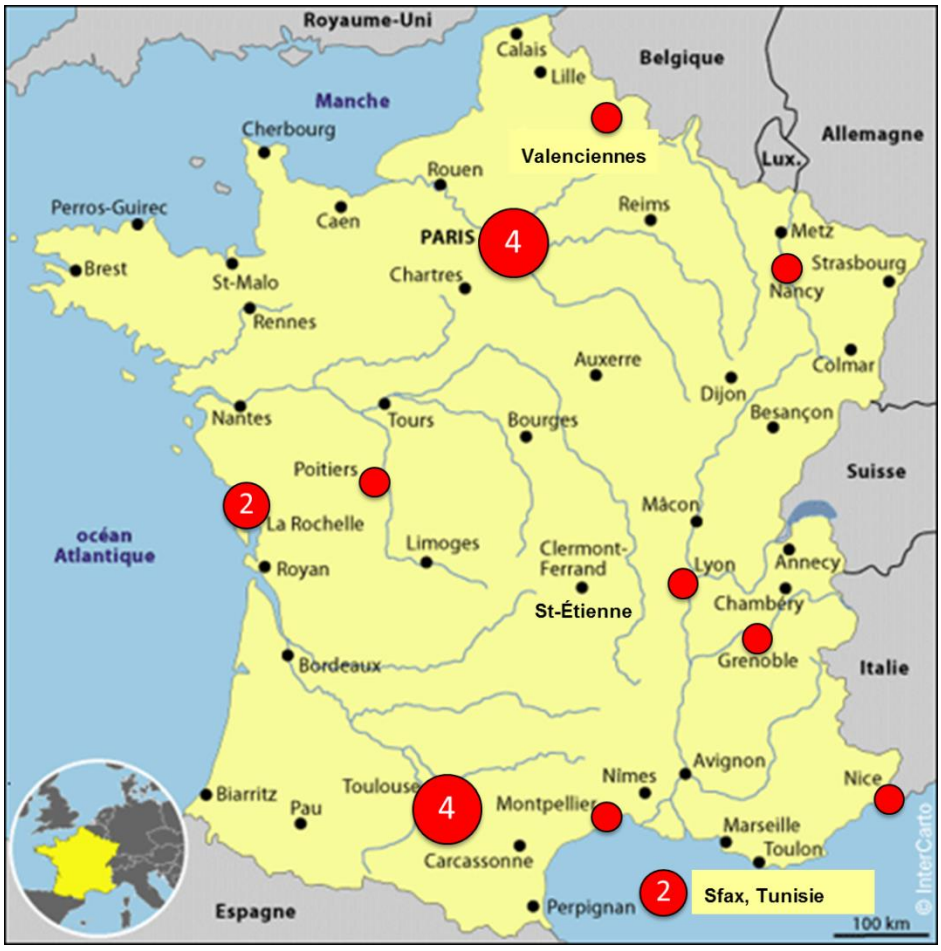
Préface

La huitième édition du Forum Jeunes Chercheurs est organisée à Grenoble le 31 mai 2016, dans le cadre du 34e congrès INFORSID. L'objectif de cette manifestation scientifique est double :

- permettre aux jeunes chercheurs en 1re ou 2e année de doctorat de présenter leur problématique de recherche et d'établir des contacts avec des équipes travaillant sur les domaines similaires ou connexes,
- offrir un aperçu des axes de recherche actuels et ainsi élargir le champ des connaissances des jeunes chercheurs.

Cette année, 32 soumissions ont été proposées par des doctorants de première ou deuxième année.

Les actes du Forum Jeunes Chercheurs recueillent les 19 articles qui ont été sélectionnés. Ils ont été rédigés par des doctorants issus de divers laboratoires de recherche en France et en Tunisie. Ces articles sont présentés lors de la session plénière du congrès consacrée au forum. Durant toute la durée du congrès, l'exposition des posters préparés par les doctorants offre également des opportunités de rencontres et de discussions avec les chercheurs de la communauté des systèmes d'information.



Répartition spatiale des articles présentés au Forum Jeunes Chercheurs 2016

Le nombre et la qualité des articles ainsi que la variété des thématiques abordées par les doctorants participant au Forum Jeunes Chercheurs sont autant d'éléments qui attestent de la vigueur des recherches dans le domaine des systèmes d'information.

De nombreuses personnes ont contribué au bon déroulement de l'organisation de cette huitième édition du forum. Je tiens ici à les remercier chaleureusement.

- Tout d'abord, merci au bureau de l'Association INFORSID de sa confiance pour l'organisation de ce forum : Régine Laleau sa présidente, Franck Ravat son vice-président, Philippe Roose son trésorier, Agnès Front sa secrétaire et Elöd Egyed-Zsigmond son chargé de communication.

- Merci à la présidente du Comité de Programme d'INFORSID 2016, Mireille Blay-Fornarino pour le travail mené en concertation.

- Un grand merci à la présidente du Comité d'Organisation d'INFORSID 2016, Dominique Rieu, et à toute l'équipe organisatrice grenobloise pour le support apporté sur les différents plans. Un merci particulier aux personnes de cette équipe avec qui j'ai été en contact pour le forum : Agnès Front, Cyril Labbé, Dominique Rieu, Claudia Roncancio, Marlène Villanova-Oliver (site Web, hébergement des doctorants, logistique des posters, …).

- Je tiens également à adresser mes remerciements à Guillaume Cabanac, organisateur de la précédente édition du Forum Jeunes Chercheurs à Lyon, qui, en fournissant les archives de l'édition précédente, m'a facilité différentes tâches, en plus de m'aider à assurer également la promotion du forum.

- Enfin, je remercie vivement l'ensemble des doctorants pour leur contribution, sans qui ce forum n'aurait pas lieu.

Cécile FAVRE
Université de Lyon (Lumière Lyon 2)
Laboratoire Entrepôt, Représentation et Ingénierie des Connaissances (ERIC)
Responsable du Forum Jeunes Chercheurs 2016

Table des matières

Articles de doctorants inscrits en 1re année de thèse :

Articles de doctorants inscrits en 2e année de thèse :

Du recueil d'expertises à la production de logiciels fiables

Application aux systèmes de neurostimulation médicale

Clément Duffau

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis
2000, route des Lucioles
06903 Sophia Antipolis Cedex

duffau@i3s.unice.fr

MOTS-CLÉS : génie logiciel, ligne de produits logiciels, connaissance, expert, medical

KEYWORDS: software engineering, software product line, knowlegde, expert, medical

ENCADREMENT : Mireille Blay-Fornarino (PR)

1. Contexte

La problématique portée par la société AXONIC est la production de systèmes de neurostimulation médicale (SNSM) dédiés à différentes pathologies (e.g., obésité, douleur fantôme). Un SNSM se présente comme un neurostimulateur (partie hardware), ses électrodes de stimulation et un logiciel de pilotage. Ce logiciel est un système d'information utilisé par différents corps de métiers (chirurgien, clinicien, patient). Les systèmes ainsi développés doivent répondre à des normes strictes qui reposent entre autres sur un cycle de développement qui inclut de nombreuses phases de tests et en particulier, une fois la Vérification & Validation passée, sur des campagnes d'expérimentations mettant en jeu différents sujets (de l'animal à l'homme avec différentes caractéristiques). En fonction des phases de tests, des propriétés différentes sont recherchées (*e.g.* seuils de douleur, efficacité, identification des effets indésirables). On parle d'expérimentation, lorsque l'on mène sur un même sujet un ensemble de stimulations. Chaque stimulation donne lieu à des résultats physiques "bruts" (*e.g.* ressenti du patient, imagerie médicale). Les résultats obtenus dans le cadre d'une expérimentation sont ensuite interprétés par des experts métiers (*e.g.* cliniciens, chirurgiens) qui en déduisent alors la validation ou non des propriétés recherchées ; l'ensemble stimulation, résultat, et interprétation devient alors une *connaissance*.

Sur la base de ces connaissances, il s'agit alors de contraindre les SNSM pour, par exemple, interdire les stimulations qui conduisent à de la douleur. Aujourd'hui, cette étape repose principalement sur le développeur qui doit configurer le logiciel pour traduire ces différentes "connaissances" en contraintes et recommandations. Or chaque SNSM présente ses propres caractéristiques (*e.g.,* type de

stimulateur, implantation ou non, forme de l'onde électrique), ce qui conduit à une grande variabilité des SNSM. Les systèmes eux-mêmes doivent présenter des plages de paramétrages différentes en fonction des sujets, des pathologies, des propriétés recherchées.

Dans cette thèse nous nous intéressons à faciliter la production de SNSM adaptés aux sujets et aux pathologies en intégrant la prise en compte des expérimentations dans le cycle de production de ces SNSM. Bien qu'appliquée dans le contexte de cette thèse au domaine médical porté par la société AXONIC, cette problématique est généralisable à la mise en œuvre de grands systèmes, dès que la connaissance sur ces systèmes dépend d'expérimentations.

2. État de l'art

Les lignes de produits logiciels (LPL) sont des solutions bien adaptées pour produire des logiciels fiables, automatiquement (Pohl *et al.*, 2005). Dans notre cas d'étude, la ligne doit évoluer pour intégrer non seulement de nouvelles fonctionnalités (*e.g.*, nouveau stimulateur, nouvelle pathologie) mais également et surtout, s'enrichir des connaissances acquises. En cela, nous nous approchons d'une problématique d'écosystèmes (Bosch, 2009). Toute la difficulté est alors de faire évoluer les assets associés (au sens de la LPL, artefacts nécessaires à la production de codes)(Seidl *et al.*, 2012).

L'évolution du logiciel a largement été étudiée (Mens *et al.*, 2005). Plus spécifiquement, nous nous intéressons aux évolutions au cours du cycle de vie, à quel niveau ont-elles un impact et comment les mettre en œuvre (Buckley *et al.*, 2005). Nous axons notre travail sur le challenge de l'enrichissement d'une LPL à partir d'expérimentations qui est un point non abordé dans la littérature à notre connaissance.

En amont il s'agit aussi d'être capable de capitaliser les connaissances recueillies lors des expérimentations. Ainsi Polacsek (Polacsek, 2016) propose d'organiser les connaissances sous la forme de graphes d'argumentation. Toute la difficulté pour nous est alors de n'autoriser que des assertions et de détecter les incohérences. Les stratégies doivent alors être formalisées par exemple pour se ramener à de la logique (Clarke *et al.*, 2010). Ainsi contrairement à des approches de data mining (Donoghue *et al.*, 2015), nous ne cherchons ni à nous substituer aux cliniciens ni à optimiser les paramètres de sélection des stimulations, mais uniquement à construire automatiquement un système qui respecte les contraintes et situations apprises des expérimentations.

3. Problématique

La figure 1 présente notre point de vue actuel sur la problématique abordée dans cette thèse : construire automatiquement un Système d'Information (SI) qui respecte les contraintes apprises à partir d'expérimentations. On distingue 3 catégories d'acteurs : (1) les experts apportent leur expertise technique en amont du projet pour définir les contraintes globales du système en s'appuyant sur l'état de l'art ou des documents techniques ; (2) les "Product Designers" sont en charge de configurer la ligne pour générer le SI dédié ; (3) les expérimentateurs sont les utilisateurs finaux. Nous identifions les défis suivants.

D1, D6 : Alimenter la base de connaissances Aujourd'hui les expérimentations sont mémorisées à travers des fiches techniques, des documents ou tableurs synthétisant les résultats d'une expérimentation. Ces données sont hétérogènes, non liées entre elles et constituent un ensemble documentaire volumineux. Les informations sont donc diffuses et peu maîtrisables par la totalité des experts. Le défi est de maintenir l'expression de ces expertises séparée et incrémentale, tout en garantissant la cohérence des connaissances qui en résultent. Or une expérimentation peut être

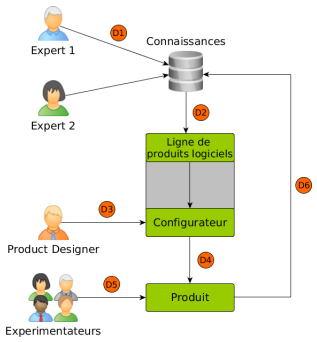


Figure 1. *Vision globale de la problématique*

contredite ultérieurement. Il faut donc être capable de gérer ses incohérences et évolutions. La problématique du passage à l'échelle se pose aussi, notamment sur la remontée de connaissances issues d'expérimentations.

D2 : Relation entre la connaissance et la LPL La maîtrise de la « ligne de produits logiciels » repose aujourd'hui sur l'ingénieur logiciel, à l'intersection de plusieurs corps de métiers. Pour libérer le développeur et en même temps garantir la prise en compte des connaissances acquises, le défi principal est d'être capable d'exploiter automatiquement les connaissances au niveau des différents types d'assets qui constituent la ligne de produits et de garantir différentes propriétés sur le système résultant comme par exemple que l'ensemble des configurations de stimulation proposées pour un traitement chez l'Homme ont été "testées" auparavant, avec toutes les ambiguïtés à lever derrière ce terme.

D3 : Utiliser la LPL pour configurer des SIs dédiés Contrairement aux approches classiques, les produits issus de la ligne, suivant leur degré de maturité, peuvent être utilisés pour des besoins différents : R&D, telle qu'une recherche sur une nouvelle pathologie (*e.g.* état de l'art, affiner un traitement) ou production quand l'étape de R&D est jugée suffisamment concluante. Le SI dédié se trouve fortement impacté par cette distinction. Ainsi, suivant les besoins, le Product Designer qui sera amené à le configurer sera différent. Ceci impacte les choix dans la configuration du produit notamment sur les rôles utilisateurs. Il semble qu'il y ait un niveau intermédiaire de configuration (Hubaux *et al.*, 2009).

D4, D5 : Générer un produit utilisable issu de la configuration Le produit généré sera composé de deux modules qui doivent interagir : le logiciel pour réaliser les expérimentations et celui pour effectuer le recueil de résultats d'expérimentations. Le premier pilote les expérimentations tandis que le second permet de capitaliser sur les résultats de ces expérimentations. La problématique est donc de générer ces modules en tenant compte du modèle métier et de l'usage qui en sera fait.

4. Actions réalisées

***Domain Specific Languages* (DSL) dédiés** Un DSL, pour le clinicien, a été défini permettant de collecter des contraintes (D1). Nous avons également traité la transformation de ces contraintes dans un modèle exploitable par la LPL abordant ainsi le défi D2.

LPL La caractérisation du domaine a été réalisée pour les produits existants au sein de la société via une représentation sous forme de Feature Models (Kang *et al.*, 1990). Nous avons aussi travaillé

autour du module de recueil d'analyse expérimentations pour en définir une maquette de la GUI répondant au besoin des expérimentateurs, qui sont les cliniciens et plus généralement les professions médicales dans le cadre de l'entreprise (D3, D4).

Représentation de la connaissance Le travail sur cette GUI nous a permis de réaliser un modèle abstrait des informations à afficher et à recueillir (D5).

5. Actions futures

***Domain Specific Languages* (DSL) dédiés** Nous travaillons actuellement à définir un DSL, métamodèle et *Graphical User Interface* (GUI), pour un autre expert dans le but de se confronter à l'hétérogénéité des données (D1) et à la collecte des expérimentations (D5, D6). La gestion des incohérences qui peuvent apparaître dans la base de connaissance (D1, D6) fait partie des actions futures.

LPL La correspondance entre les contraintes issues de la base de connaissances et les assets reste un défi pour lequel les pistes envisagées portent sur des techniques de méta-modélisation (D2)(Favre *et al.*, 2006). Le développement du premier prototype du GUI d'expérimentation doit nous permettre d'établir les bases pour la production automatique idoine des GUI (D4, D5).

Représentation de la connaissance Il nous reste à valider ce modèle sur des expérimentations réelles puis de travailler à leur transformation en données exploitables. Une piste nous vient de la théorie de l'argumentation (Polacsek, 2016)(D6).

6. Bibliographie

Bosch J., « From Software Product Lines to Software Ecosystems », , vol. 1, p. 111–119, 2009.

Buckley J., Mens T., Zenger M., Rashid A., Kniesel G., « Towards a taxonomy of software change », *Journal of Software Maintenance and Evolution*, vol. 17, n° 5, p. 309–332, 2005.

Clarke D., Proença J., « Towards a Theory of Views for Feature Models », *Proceedings of FMSPLE'10*, 2010.

Donoghue J. O., Roantree M., Boxtel M. V., « A Configurable Deep Network for high-dimensional clinical trial data », *Neural Networks (IJCNN), 2015 International Joint Conference on*, p. 1-8, July, 2015.

Favre J.-M., Establier J., Blay-Fornarino M. (éd.), *L'ingénierie dirigée par les modèles : au-delà du MDA*, Hermes-Lavoisier, Cachan, France, 2006.

Hubaux A., Classen A., Heymans P., « Formal Modelling of Feature Configuration Workflows », *SPLC'09*, IEEE, p. 221–230, 2009.

Kang K. C., Cohen S. G., Hess J. A., Novak W. A., Spencer Peterson A., Feature-oriented domain analysis (FODA) feasibility study, Technical Report n° November, The Software Engineering Institute, 1990.

Mens T., Demeyer S., Wermelinger M., Hirschfeld R., Ducasse S., Jazayeri M., « Challenges in software evolution », *International Workshop on Principles of Software Evolution (IWPSE)*, vol. 2005, p. 13–22, 2005.

Pohl K., Böckle G., Linden F. J. v. d., *Software Product Line Engineering : Foundations, Principles and Techniques*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

Polacsek T., « Arbre d'argumentation : un nouvel acteur dans vos diagrammes », *INFormatique ORganisation et Systèmes d'Information et de Décision (INFORSID)*, 2016.

Seidl C., Heidenreich F., Assmann U., « Co-evolution of Models and Feature Mapping in Software Product Lines », *Proceedings of the 16th International Software Product Line Conference (SPLC)*, vol. 1, p. 76–85, 2012.

A Visualization Platform for Epidemiology Surveillance

Samiha Fadloun

LIRMM (Université de Montpellier, CNRS)
Campus St Priest, 860 rue St Priest,34095 Montpellier Cedex 5, France
samiha.fadloun@lirmm.fr

MOTS-CLES : *Analyse Visuelle, Fouille de Texte, Fouille de données*

KEYWORDS: *Visual Analytics, Text mining, Data mining*

SUPERVISORS: *Arnaud Sallaberry (MCF), Pascal Poncelet (PR), Mathieu Roche (DR), Julien Rabatel (POST DOC)*

1. Context

Today epidemiologists have to face a huge amount of information to investigate animal disease outbreaks. Usually, this information is available on the Web, through news articles or official reports. One of the major issues, they have to face is how to find the relevant information. In this paper, we propose a new visualization platform for epidemiology surveillance, it combines text mining and information visualization. In principle, we do not plan to apply text mining and then visualization rather we aim at providing an interactive system. With our visual analytics platform, interactivity between users, visualization and text mining are the key for helping epidemiologist researchers.

2. State-of-the-art

Visual Analytics is “the science of analytical reasoning facilitated by interactive visual interfaces” (Cook *et al.*, 2005), it combines multiple techniques (Figure 1), and makes the processing of data and information transparent for an analytic discourse,

in addition to constructive evaluation, correction and rapid improvement of processes and models (Keim *et al.*, 2008). In order to facilitate the knowledge and data mining processing, there are many visualization tools, using either a single or a combination of techniques. For example Cho *et al.* (2016) propose the ViaRoma platform dedicated to the history of Roma by combining different visualization techniques such as map, tree and sunburst views. A review of visualization tools focused on the landscape of infectious diseases for public health is proposed in Carroll *et al.* (2014). They mainly evaluate visualization relying on geographic information systems such as Healthmap (Brownstein *et al.*, 2008), molecular epidemiology or social network analysis. They also investigate the user needs, preferences, efforts and barriers to adopt such a kind of abstractions.

Now, from a text mining point of view, epidemiologists have to consider a huge amount of documents. In order to help them, they usually apply some text mining. For instance, Arsevska *et al.* (2016) propose to apply both text and web mining to extract, and validate disease outbreaks for animals. Furthermore, they are able to automatically extract relationships between host species and symptoms.

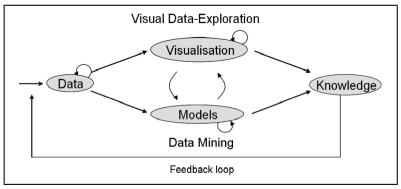


Figure 1. *The Visual Analytics Process*

3. Problem Statement

As stated in the introduction, epidemiologists need to monitor and analyze a huge amount of data. It must be clearly and easily managed, in order to facilitate its interpretation and help making the right decision.
Most of the visualization tools used by epidemiologists are maps, for showing the distributions of some diseases over the world and, these tools frequently do not offer a real interaction. Our concern is different since we address the way of finding relevant documents according to the interest of the user. Let us consider the following illustrative example: the distribution of clinical signs of the Foot-and-mouth disease such as production losses, vesicular stomatitis and vesicular disease, on animals such as cattle, pig and deer. Experts have a strong knowledge on the disease for these animals and they would like to know if Foot-and-mouth disease can affect camelids. Usually, to do so, they manually query one search engine to find some recent news or reports related to Foot-and-mouth disease and camelids.
Our platform is initialized by the needs of text mining research (Arsevska *et al.*, 2016), how can experts analyze the data results?, how can experts use their knowledge? etc.

We assume that our platform aims to create a visual analytic tool for helping the user to express what he/she is focusing on, and automatically query different search engines whatever the complexity of the query. Furthermore with text mining approaches, we propose to rank the results in order to provide the end user with the most representative and useful documents. Finally some tools are proposed to validate/invalidate the results and then store them to perform his/her own research.

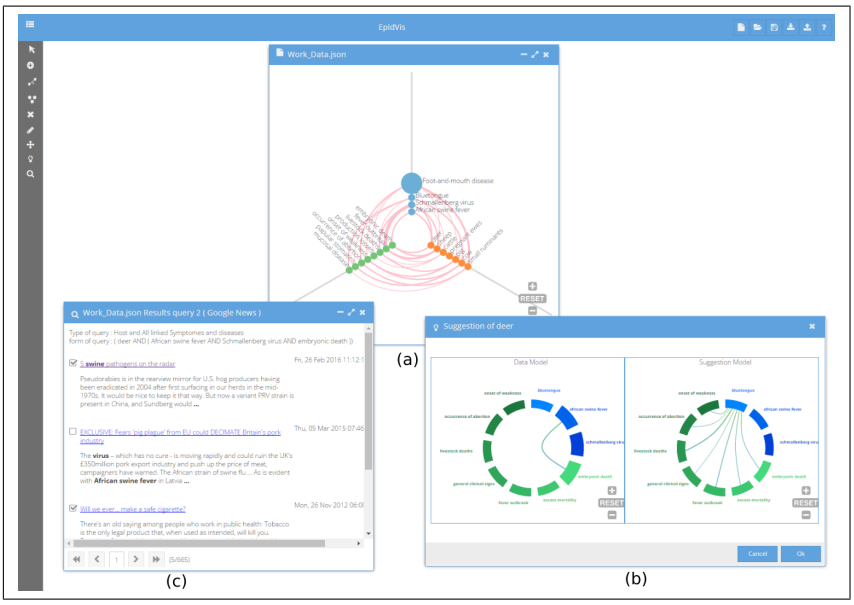


Figure 2. *Screenshot of EpidVis*

4. Our current approach

Few researches address the visualization of animal epidemiology data. We propose a visual analytics tool combining visualization, text mining, and expert knowledge (Figure 2). It uses data obtained by text mining approaches, i.e. names of diseases, hosts and symptoms, and association values between them (Arsevska *et al.*, 2016). Our method is based on three components:

1) **Main interface** (Figure 2.(a)). In this interface, the end user is provided with an abstraction for visualizing terms and relationships: hosts, symptoms and diseases, and the links between them. It is based on a Hive plot (Krzywinski *et al.*, 2011) technique: the data are represented on three axes. Each axis plots a set of nodes from a category, and the relationships between them are represented by edges. Experts can dynamically create new nodes and links, the platform used CSV and JSON with require format.

2) **Data Suggestion** (Figure 2.(b)). This radial visualization is proposed to suggest new relationships to the epidemiologist.

3) **Queries and Information Retrieval** (Figure 2.(c)). This interface shows the results of one query which has been automatically created by the platform, with access to the server. Similar results can be obtained with different search engines. Different ranking techniques based on text mining measures are proposed in order to highlight relevant documents to experts. A quantitative and qualitative evaluation of the tool is currently performed. This is based on new information (i.e. returned web pages) suggested by our system.

5. Future Work

We plan to improve the visualization tool to take into account spatial and temporal information. Furthermore to help for instance the understanding of the spread of an epidemy, additional information will be considered (e.g. symptoms, number of cases). They will be extracted directly by analyzing the content of the returned documents. Other data mining techniques, such as clustering, will be also investigated for the visualization. Finally we will try to generalize our approach on different epidemiology domains like human epidemiology.

Acknowledgments. We would like to express our sincere gratitude to Elena Arsevska (Cirad, CMAEE) for the data, fruitful discussions and evaluation of results. This work has been partially funded by Labex NUMEV (ANR-10-LABX-20), and by the Ministry of Higher Education and Scientific Research of Algeria.

6. References

Arsevska E., Roche M., Hendrikx P., Chavernac D., Falala S., Lancelot R., Dufour B., "Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web", *Computers and Electronics in Agriculture*, vol. 123, p. 104–115, 2016.

Brownstein J. S., Freifeld C. C., Reis B. Y., Mandl K. D., "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project", *PLoS Med*, vol. 5, n° 7, p. e151, 2008.

Carroll L. N., Au A. P., Detwiler L. T., Fu T.-c., Painter I. S., Abernethy N. F., "Visualization and analytics tools for infectious disease epidemiology: a systematic review", *Journal of biomedical informatics*, vol. 51, p. 287–298, 2014.

Cho I., Dou W., Wang D. X., Sauda E., Ribarsky W., "VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History", *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, n° 1, p. 210–219, 2016.

Cook K. A., Thomas J. J., Illuminating the path: The research and development agenda for visual analytics, Technical report, Pacific Northwest National Laboratory, WA (US), 2005.

Keim D., Andrienko G., Fekete J.-D., Görg C., Kohlhammer J., Melançon G., *Visual analytics: Definition, process, and challenges*, Springer, 2008.

Krzywinski M., Birol I., Jones S. J., Marra M. A., "Hive plots rational approach to visualizing networks", *Briefings in bioinformatics*p. 069, 2011.

Vers une plateforme de monitoring de flottes d'objets connectés

Asmaa Achtaich

Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac, 75013 Paris France
Asmaa.achtaich@malix.univ-paris1.fr
SIWEB Team, Ecole Mohammedia d'Ingénieurs, Univ. Mohammed V-Rabat
Avenue Ibnsina ,765 Agdal Rabat Maroc
asmaaachtaich@research.emi.ac.ma

MOTS-CLÉS : *IdO, Objets Connectes, variabilité, Flotte De Mobile, EMM, MDM, Lignes De Produits Logiciels, Constraints Programming, Monitoring, Auto-Adaptation.*

KEYWORDS: *IoT, Connected Objects, Variability, Mobile Fleet, EMM, MDM, Software Product Lines, Constraint Programming, Monitoring, Self-Adaptation.*

ENCADREMENT : *Camille Salinesi (PR), Ounsa Roudies (PR), Nissrine Souissi (HDR), Raul Mazo (MCF)*

1. Contexte

Les appareils mobiles ont transformé nos modes de communication, réflexion et travail. 72% des employés en 2015 ont utilisé leurs appareils mobiles à des fins professionnelles (Citrix 2015). L'intégration des dispositifs dans le lieu de travail est une grande opportunité pour les entreprises mais peut aussi s'avérer d'une grande complexité. Les appareils mobiles non gérés posent des risques élevés pour les entreprises (Leavitt 2013). Par conséquent, une nouvelle catégorie de solutions logicielles appelée Enterprise Mobility Management (EMM) a été introduite. Ces solutions permettent d'intégrer pleinement les technologies mobiles dans un environnement sécurisé en permettant l'inscription à l'infrastructure, l'application de configurations, le stockage sécurisé des données, l'accès réglementé aux applications, le monitoring des ressources et enfin, le reporting sur l'activité des employés. Cependant, les objets connectés de manière plus générique sont de plus en plus déployés dans le contexte professionnel (avec les chaînes de production industrielle) autant que personnel (avec la domotique) ou gouvernemental (avec les villes intelligentes). La conception des objets connectés dans l'infrastructure présente d'énormes opportunités économiques, mais aussi des défis majeurs et de nouveaux verrous scientifiques. Tout comme les EMMs, des solutions de gestion de dispositifs de l'Internet des Objets (IdO) s'avèrent primordiales. Notamment pour la

détection et l'enrôlement de dispositifs, le support de communication hétérogène, la sécurité des données, les échanges, le stockage d'informations ou encore pour maitriser la taille, les caractéristiques et le contexte très variable de ces objets.

Nos travaux de recherche visent à concevoir une plateforme pour la surveillance et le monitoring de flottes d'objets connectés d'entreprise. Notre contribution prend en compte la nature variable de ces composants et se base sur les bonnes pratiques de construction de solutions EMM.

2. État de l'art

Dès les premières discussions autour de l'IdO dans le MIT Auto-ID Center, il était clair que ces enjeux sur l'entreprises sont de taille. Toutefois, pour que ce paradigme devienne réalité, le développement d'une infrastructure solide devait être entrepris. Plusieurs problématiques furent donc abordées par les chercheurs.

La détection des objets en mouvement fut une des premières problématiques. (EPCglobal 2013) et (Koshizuka & Sakamura 2010) – entre autres – ont proposé respectivement l'ONS et le uCode Resolution Protocol, deux processus d'adressage qui se basent sur le Protocol DNS pour identifier les objets de manière unique. Ensuite, les échanges entre des dispositifs hétérogènes qui présentent des capacités très différentes du point de vue computationnel et communication est un autre défi. Ainsi, plusieurs efforts de standardisation ont été mis en place, Notamment (EPCglobal 2014) pour les technologies RFID, la norme IEEE 1451 qui définit un format standard des données d'échange et de protocoles de communication, ou encore (Zigbee Alliance 2006) qui maintient et publie les normes ZigBee. La sécurité et la confidentialité des échanges sont aussi des difficultés majeures pour la mise en œuvre de l'IdO. Certaines solutions dans ce sens sont revus dans (Suo et al. 2012). Ensuite, un des chalenges actuels de l'IdO est la scalabilité : au cours des quatre prochaines années, 25 milliards d'objets seront connectés (Rivera & Meulen 2014). De ce de fait, les middlewares supportant cette évolution sont devenus une exigence clé pour soutenir la croissance actuelle de l'IdO, chose abordée dans (Gomes et al. 2014). Finalement, pour surmonter le grand nombre de dispositifs, la surveillance et le monitoring s'avèrent une nécessité. Le traitement des données recueillies sur les appareils et sur leur environnement permet l'ajustement intelligent, instantané et dynamique des configurations. Plusieurs propositions de Framework, d'architectures ou de modèles sur le monitoring des objets connectés ont été appliqués aux secteurs de la santé (Hansen 2014), de l'énergie (Wei & Li 2011), ou de la distribution (Li et al. 2013).

3. Problématique

Bien que les travaux de monitoring mentionnés répondent à des problématiques pertinentes, elles présentent certaines limites. En effet, d'un côté, l'internet des objets est composé de dispositifs hétérogènes qui évoluent très rapidement. La diversité et l'évolutivité doivent donc être intégrées dans la plateforme de gestion de la flotte. Ensuite, même si les objets connectés partagent plusieurs caractéristiques

communes, leurs applications et leurs contextes restent extrêmement variables : d'abords les caractéristiques matérielles et logicielles d'objets connectés sont très différentes (état de la batterie, espace de stockage libre, capacité de traitement, type et taille d'écran, mobilité, système d'exploitation, technologie de communication, …). Ensuite, l'usage du dispositif est distinct d'un élément de la flotte à un autre. Ceci dépend du rôle du dispositif, de sa configuration ou encore de la nature de l'utilisateur. Et finalement, l'environnement qui entoure les objets connectés est propre à un groupe particulier de dispositifs (localisation, température, son, mouvement, poids …).

Ceci étant, les Framework et les architectures de monitoring proposés au profit de l'IdO ne prennent pas en compte la diversité des objets connectés, la variabilité de leurs contextes et ainsi leurs configurations, et leur évolutivité. Notre objectif consiste donc à capturer, modéliser et raisonner sur la variabilité dans une flotte d'objets connectés, ce en se basant sur les principes de l'ingénierie des lignes de produits. Plus précisément, par le biais de la Programmation Par Contraintes (PPC) de par son efficacité à spécifier les Lignes de Produits Logiciels (LPL) et à analyser la variabilité (Sawyer et al. 2012) (Salinesi et al. 2011).

4. Actions réalisées

Notre objectif est de proposer une plateforme de monitoring d'objets connectés, en se basant d'une part sur les principes et architectures des solutions EMM, et d'une autre part sur l'ingénierie des lignes de produits logiciels pour la gestion de la variabilité. Les EMMs étant eux-mêmes des applications de l'IdO sont testés et approuvés par l'entreprise. Ce qui représente pour nous une base de bonnes pratiques sur lesquelles notre plateforme peut s'appuyer. Nous avons donc procédé par un état de l'art de ces solutions afin d'en discerner la contribution ainsi que les principales limites. Les axes d'amélioration portent principalement sur le traitement de l'information sur le contexte du dispositif pour d'éventuelles reconfigurations, le monitoring automatique, intelligent et dynamique, la communication avec des supports hétérogènes ainsi que la gestion des ressources limitées. Nous nous intéressons particulièrement au monitoring, cependant, nous étendons le spectre de dispositifs de manière à ce qu'il englobe les objets connectés dans un sens large. Finalement, pour la création de solutions qui prennent en compte et raisonnent sur la variabilité, nous proposons une méthodologie qui repose sur le paradigme des Lignes de Produits Logiciels (LPL). Celui-ci offre les outils nécessaires pour la création de famille de produits qui partagent des caractéristiques en commun, tout en satisfaisant les exigences particulières d'un contexte donné.

5. Actions futures

L'étape en cours consiste à déterminer l'architecture de notre plateforme. Celle-ci reposera sur les standards des EMMs, de l'IdO et des LPLs. La prochaine étape est reliée à la surveillance des objets connectés. Il s'agit de définir une ontologie pour modéliser les éléments du domaine ainsi que les règles de gestion. Ensuite,

pour assurer la configuration conditionnée des objets connectés, la démarche envisageable est d'adopter la PPC pour la définition des features, et pour raisonner sur la variabilité. Finalement, dans le cadre de leur validation, nos contributions feront l'objet d'études de cas réelles, probablement dans les secteurs médical (Smart Hospital) et domotique (Smart House).

6. Bibliographie

Citrix, 2015. 7 Enterprise Mobility Statistics You Should Know. Available at: https://www.citrix.com/articles-and-insights/workforce-mobility/jun-2015/7-enterprise-mobility-statistics-you-should-know.html.

EPCglobal, 2013. GS1 Object Name Service (ONS). , (2), pp.1–34.

EPCglobal, 2014. *The GS1 EPCglobal Architecture Framework*,

Gomes, M., da Rosa Righi, R. & da Costa, C.A., 2014. Internet of things scalability: Analyzing the bottlenecks and proposing alternatives. In *2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE, pp. 269–276.

Hansen, F.O. vergaard, 2014. Ambient assisted living healthcare frameworks, platforms, standards, and quality attributes. *Sensors (Basel, Switzerland)*, 14(3), pp.4312–4341.

Koshizuka, N. & Sakamura, K., 2010. Ubiquitous ID: Standards for Ubiquitous Computing and the Internet of Things. *IEEE Pervasive Computing*, 9(4), pp.98–101.

Leavitt, N., 2013. Today's Mobile Security Requires a New Approach. , pp.16–19.

Li, S. et al., 2013. Distribution equipment monitoring system based on the internet of things. In *2013 IEEE Global High Tech Congress on Electronics*. IEEE, pp. 41–45.

Rivera, J. & Meulen, R. van der, 2014. Gartner Says Two-Thirds of Enterprises Will Adopt a Mobile Device Management Solution for Corporate Liable Users Through 2017.

Salinesi, C. et al., 2011. Constraints: The core of product line engineering. In *2011 FIFTH INTERNATIONAL CONFERENCE ON RESEARCH CHALLENGES IN INFORMATION SCIENCE*. IEEE, pp. 1–10.

Sawyer, P. et al., 2012. Using Constraint Programming to Manage Configurations in Self-Adaptive Systems. *Computer*, 45(10), pp.56–63.

Suo, H. et al., 2012. Security in the internet of things: A review. *Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012*, 3, pp.648–651.

Wei, C. & Li, Y., 2011. Design of energy consumption monitoring and energy-saving management system of intelligent building based on the Internet of things. *2011 International Conference on Electronics, Communications and Control, ICECC 2011 - Proceedings*, pp.3650–3652.

Zigbee Alliance, 2006. Zigbee specification.

CP-Based Framework for Software Product Lines Engineering

Angela Villota-Gomez

Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac
75013 Paris

angela-patricia.villota-gomez@malix.univ-paris1.fr

MOTS-CLÉS : *Lignes de produits, Ingénierie des lignes de produits, Programmation par contraintes, Solveurs de contraintes.*

KEYWORDS: *Product lines, Product line engineering, Constraint programming, Constraint solvers.*

ENCADREMENT. *Camille Salinesi (PR) et Raúl Mazo (MCF)*

1. Context

The Product Line Engineering (PLE) and Software Product Line Engineering communities employ the term constraint referring to relations between elements in a system, non-functional requirements, feature relations, etc. It is almost natural that some authors formalize these concepts as boolean satisfiability problems or constraint satisfaction problems. Moreover, in the past years, many different works have formalized variability concepts and product line preferences in the shape of logic formulas or constraint satisfaction problems to support PLE activities like analysis, testing, verification, etc. with the purpose of taking advantage of Constraint Programming (CP) approaches. However, there is no proposal integrating CP in all the development stages of a PLE in a consistent manner. In this research project, we aim to design and develop a generic framework specially designed for PLE. With this framework, we want to consolidate the constraints, domains, and solvers to propose a more suitable approach for supporting PLE.

2. State of the art

We combine two paradigms in this project: CP and PLE. CP is a programming paradigm to solve combinatorial problems modeled as a Constraint Satisfaction Prob-

lem (CSP). CSPs are defined in terms of variables, domains and constraints. A CSP may have one or more solutions, and they are produced employing one of the many consistency algorithms or by searching them using a constraint satisfaction tool called solver (Rossi *et al.*, 2006). On the other hand, a product line or system family is a collection of similar products sharing common characteristics and satisfying the requirements of a particular mission or market segment. Products in a product line are assembled from a common set of core assets in a prescribed way. There are previous proposals exploiting the CP paradigm in PLE by assisting activities such as: analysis, configuration, derivation, synthesis, testing, simulation, and verification of product lines in different contexts (Benavides *et al.*, 2005; Czarnecki *et al.*, 2005; Mazo *et al.*, 2015). Constraints are also used to represent product line variability. Most studies propose rules for transforming variability descriptions into constraint satisfaction problems. Nevertheless, works as (Salinesi *et al.*, 2011; Mazo *et al.*, 2011) consider constraint (logic) programming as an expressive method to describe variability. Moreover, the work of (Salinesi *et al.*, 2011) proposes a constraints language to describe product lines from a high-level meta-model general description. Additionally, Mazo et al. propose the use of abstract constraints to represent product lines with a high-level and unique notation that encompass different constraint languages (e.g., over Booleans, Integers, Reals, trees, lists, etc.) (Mazo *et al.*, 2011).

3. Problem

We have evidence of previous works that CP is a paradigm and a viable model for supporting a PLE process. However, there is no proposal integrating CP in all the development stages of a PL in a consistent manner. This may happen because the CP approaches are modeled or adapted for a particular objective or application. Instead, we want to design and develop a generic framework gathering constraints, domains, and constraint solvers specially designed for PL. Therefore, our framework will emerge as a suitable approach for supporting PLE consisting on: (i) a constraint language for representing PL models in a more accurate way, consolidating the constraints included in previous works, plus other not included yet. Moreover, the proposed constraint language will continue the proposal of (Mazo *et al.*, 2011) (ii) a solving mechanism that supports the proposed constraint language.

4. Actions taken

We apply the design science paradigm in our research project. This paradigm is one of the paradigms proposed to conduct research in information systems disciplines (Peffers *et al.*, 2007). According to above methodology, this project can be divided into four stages: solution design and development, demonstration, evaluation, and communication. Table 1 shows the objectives related to each stage.

As a first step, we conducted a systematic mapping study with 6 research questions over publications containing evidence of the application of the constraint programming paradigm in the developing of product lines. With this study, we aimed to provide an overview of the research on the intersection of PLE and CP subject. As results of our mapping study, we proposed a *Classification Framework* for CP-based approaches in

Stage	Objectives
Solution design and development	Objective 1: Elicitate the types of constraints considered relevant in the product lines engineering. The elicitation must be carried out by studying the literature and also by surveying stakeholder in the academic and industrial context Objective 2: Propose and develop a Generic Constraint System for modeling variability as a constraint satisfaction problem. Objective 3: Propose and develop a solving mechanism supporting the proposed constraint system.
Demonstration and evaluation	Objective 4: Validate this proposal with several case studies and benchmarks.
Communication	Objective 5: Publish and disseminate this project in conferences and index journals.

Table 1. Project stages and objectives

the context of PLE, and built a comprehensive collection of constraints for PLE. The proposed classification framework is a four-dimensional framework composed by four views: expressivity, translation, application, and support. Consequently, the proposed classification framework allows the publications characterization from four different points of view. The second contribution of our mapping study emerges as a consequence of the classification of constraints used in the PLE. To perform the classification and answer the question: *what is expressed as a constraint, and how?* we gather the constraints employed in the different publications regarding their semantics and their implementation. Therefore, regarding the semantics, we obtained two collections of constraints: constraints to document variability, and constraints to express product line preferences. Additionally, constraints in PLE are implemented using arithmetic constraints, boolean constraints, global constraints, and reified constraints.

One of the findings in our mapping study is the usage of different approaches for solving constraints problems. Furthermore, the utilization of specialized solvers for each approach. Therefore, we plan to include the different paradigms and solvers taking advantage of their strengths. In consequence, the first proposal for a CP-based general framework specially designed for PLE considers the integration of different paradigms for solving constraint problems such as Logic-Formula Satisfiability Problems (LSP), and CSP. The CP-based framework for PLE is a four-level framework, as shown in Figure 1. The first level, *constraints meta-model* contains a generic constraint language gathering the meta-model for constraints relevant in product lines engineering. The second level, *constraints instantiation* transforms a generic constraint into a constraint for a particular paradigm (i.e.: logic formulas, constraints) with the help of a compiler, and federator components. The third level, *solver paradigm meta-model* contains a component for translating a constraint (represented in a paradigm) into constraints to be used in a particular solver paradigm (e.g: SAT or BDD solvers for determine logic formula satisfiability). Finally, the fourth level, *solver implementation* transforms the constraints obtained in the third level into a program regarding the compelling a solver's implementation syntax. In consequence, the framework will include the strengths of different paradigms and solvers.

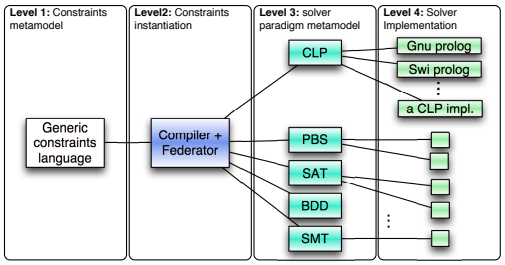


Figure 1. CP-Based Framework for Software Product Lines Engineering

5. Future Work

In this paper, we presented the proposal and first results of the research project to design and develop a CP-based general framework specially designed for PLE. In the first stages of our study, we conducted a systematic mapping study to examine the state of the art, and provide an overview of the application of CP in the PLE. Since we will further the proposal of the CP-based general framework specially designed for PLE described in section 4, the direction of our current and future work is threefold: the first step is to continue developing the collection of constraints for PLE. We built a collection of constraints for PLE as a result of our mapping study. Therefore, the next step is validate these constraints with the PLE community via a case study. The next step consist in the integration of the constraints with a solving mechanism. Currently, we consider the framework must include a mechanism to solve each problem with the most suitable paradigm and solver regarding the literature review. Finally, we plan to validate the integration constraints-solving mechanism with a case study from the industry, preferable.

6. References

Benavides D., Trinidad P., Ruiz-Cortés A., "Automated reasoning on feature models", *Advanced Information Systems Engineering*, Springer, pp. 491–503, 2005.

Czarnecki K., Kim C. H. P., "Cardinality-based feature modeling and constraints: A progress report", *International Workshop on Software Factories*, pp. 16–20, 2005.

Mazo R., Muñoz Fernández J. C., Rincón L., Salinesi C., Tamura G., "Variamos: an extensible tool for engineering (dynamic) product lines", *SPLC*, ACM, pp. 374–379, 2015.

Mazo R., Salinesi C., Diaz D., "Abstract Constraints: A General Framework for Solver-Independent Reasoning on Product Line Models", *INSIGHT-Journal of International Council on Systems Engineering (INCOSE)*, vol. 14, num. 4, pp. 22, 2011.

Peffers K., Tuunanen T., Rothenberger M., Chatterjee S., "A Design Science Research Methodology for Information Systems Research", *J. Manage. Inf. Syst.*, vol. 24, pp. 45–77, 2007.

Rossi F., Beek P. v., Walsh T., *Handbook of Constraint Programming*, Elsevier Science Inc., New York, NY, USA, 2006.

Salinesi C., Mazo R., Djebbi O., Diaz D., Lora-Michiels A., "Constraints: The core of product line engineering", *Research Challenges in Information Science (RCIS), 2011 Fifth International Conference on*, pp. 1–10, 2011.

Cross-Company Collaboration Analyzed Through Process Mining

A Method for Analyzing the Working Environment and Measuring Impacts of Change

Ornela Çela

Laboratoire d'informatique de Grenoble (LIG)
LIG - Bâtiment IMAG - CS 40700
38058 GRENOBLE CEDEX 9
Ornela.cela@imag.fr

KEYWORDS: *Process Mining, Software Process Mining, Bug tracker, Collaboration*
ENCADREMENT : *Agnes Front (MCF) et Dominique Rieu (PR)*

1. Context

Cross companies collaboration is considered nowadays a competitive advantage, so business entities attempt to use it for increasing their benefits. Nevertheless the economic benefits, this collaboration can cause lots of problems in the working environment. The non-uniform workload between users and the vagueness of user's functions and roles in the work schema are two of the main problems caused during the collaboration.

The goal of our project is to provide a way of analyzing the working environment and the collaboration between companies, by processing event log through process mining techniques. Our works will combine different aspects of process mining [1] in order to analyze the environment. We will propose a method which analyzes the cross collaboration hierarchy and introducing the idea of discovering a model of communication between the companies. Moreover the method will be enriched with the possibility of measuring the impact that possible changes in the hierarchical structure might have in the collaboration performance.

2. State-of-the-art

Process mining is a mean of extracting non trivial and useful information from process execution logs also called audit trails, or transaction logs [2]. The main perspective of process mining is the control flow perspective. The aim is creating a model of the business process using the event log files that are the starting point based

on the traces left by the users in the information system supporting the activity. There are identified three main focuses of process mining as shown in figure 1:

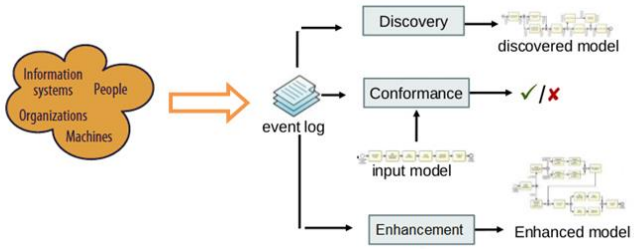


Figure 1: The main focuses of process mining.

1. Discovery aims to elaborate the event log and other data provided as input and to provide a model of the process.

2. Conformance targets the problems such as checking if the event log or reality processing is conform or not to the business model already available, or if the process model created by the analyst is respected into practice, and improving the model discovered into the first type of process mining.

3. Enhancement offers the user the possibility to play in, play out and replay the model. These actions make possible for the user to make recommendation, take decisions, analyze bottlenecks and make predictions based upon event logs.

Another perspective of process mining is the organizational and social perspective. The mains centers of attraction of this perspective are the resources, identifying the actors involved and how they are related between each other and to the tasks of the process.

3. Problematic

The main topic of our work is the usage of process mining into the environment of software development, which is defined as software process mining in [3]. The trend of applying software process mining has augmented, but to the best of our knowledge, there is no work done in this domain that takes into consideration the organizational and social perspective.

Our study is based on analyzing the collaboration between companies in the domain of software support providing. We are working on data provided from a software developing company which offers software support to its clients. This collaboration between the company and the client is coordinated through Mantis bug tracker software and there is a predefine schema of communication between the different services of each parties. Nevertheless this configuration, the parties complain over the time required to solve issues, the workload and the interaction between the companies.

Our goal is, through process mining techniques to analyze the working environment, identify problems and propose solutions for solving performance issues. At the moment this tasks are performed by analyzing the working environment manually, since the environment lacks in terms of providing in depth analysis. Our work will be to propose a method based on process mining technique as a mean to analyze this environment and measure the impact of future changes. The impact of future changes is going to be measured by combining simulation to process mining techniques as introduced in [4], in order to get a real view on the ongoing process and more realistic prediction.

4. Performed actions

We worked on a schema of collaboration between the parties: the client, the support provider and the finance consultant. We began to analyze a group of 203 closed issues derived from the Mantis Bug Tracker system. Based on the use of process mining techniques, we were able to provide ways for observing:

1. Models of workflow for the issues, which allow the detection of the tasks which are more time consuming and bottleneck in the process.
2. Reports of tasks performed over time in order to detect patterns of work gathering.
3. Models of communications between the services of the companies in order to see if communication is conform to the predefined communication schema and if there were deviations as shown in the figure 2, represent in dashed lines.

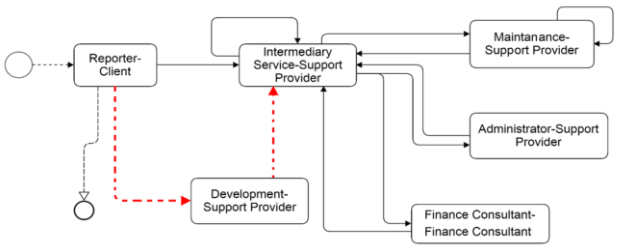


Figure 2: The model of communication between the services.

4. A model of user interactions in order to detect patterns of collaboration between resources.

As a final result, we propose the basis for a method which empowers performing a general analyze of the work environment and analyze the collaboration between companies. Finally we introduced the idea of measuring the impact of a change, giving so the opportunity to the decision maker to better analyze the impact of the proposed change. The method we proposed is composed of three parts: the general analyze, the cross company communication analyze and the measuring of the future impacts.

5. Future actions

Our future work aims to consolidate and design each one of the phases of the proposed method and their requirements:

- The general analyzing phase will provide reports generated through process mining tools which focus on the workflow of the process, group tasks over time, model of workload dispersion and also model of user collaboration pattern.
- The cross-company analyze phase will aim to analyze the collaboration between the services and put in evidence if the hierarchy of the companies is respected inside the collaborative environment.
- The measuring phase will aim to study the impact of changes thought the combination of process mining with simulation techniques. The fusion of this two disciplines allows testing of changes in the workflow of the process.

Once the method is consolidated, we are going to validate its usage. The validation will focus on evaluating the method's usage and utility. We also plan to focus in enriching our method with additional elements such as:

- prediction perspective: predicting the time required for the correction of an issue,
- estimating the usage of the method into other support providing environment different from the software support provider and also in the software development domain and other domains
- adding text mining techniques in order to analyze if there is or not sufficient input information when passing the issue to the next step,
- creating of an evolutionary model in order to analyze the people's way of working changes over time. This model provides a way of seeing how the model change over time and to provide a way of detecting possible causes of such changes.

6. References

Van der Aalst W.M.P., *Process Mining Discovery, Conformance and Enhancement of Business Processes*, ISBN 978-3-642-19344-6, 2011.

Rozinat A., Mans R.S., Song M., van der Aalst W.M.P., " Discovering simulation models ", *Journal Information Systems*, 2009.

Rubin V., Lomazova I., van der Aalst W.M.P, " Agile Development with Software Process Mining ", *International Conference on Software and System Process*, 2014.

Wynn M.T., Rozinat A., van der Aalst W.M.P., ter Hofstede A.H.M, Fidge C.J., " Process Mining and Simulation ", *Modern Business Process Automation*, 2010.

Etude des processus de vigilance pour la sécurité de systèmes décisionnels

Amal Chaker [1, 2]

[1] *Institut de Recherche en Informatique de Toulouse (IRIT), 118 Route de Narbonne, F-31062 Toulouse Cedex*
[2] *Trimane, 57 rue de Mareil, 78100 Saint-Germain-en-Laye*
amal.chaker@irit.fr

MOTS-CLÉS : Sécurité, Systèmes d'information, Systèmes décisionnels, « Big Data », Prévention des risques, Base de signatures.

KEYWORDS: Security, Information systems, Decision systems, Big Data, Risk prevention, Signature database.

ENCADREMENT (THESE CIFRE) : Gilles Zurfluh [1](PR) et Fatma Abdelhédi [1,2] (Directrice du laboratoire de recherche CBI² à Trimane).

1. Contexte

Grâce aux nouvelles technologies, les entreprises mettent en place des systèmes d'information de plus en plus sophistiqués pour répondre aux besoins d'information de différents acteurs : direction, employés opérationnels, clients, etc. Les systèmes d'information apparaissent donc comme le garant d'un fonctionnement efficace des entreprises au quotidien. Pour atteindre leurs objectifs, notamment pour garantir une continuité de service auprès des utilisateurs, ces systèmes doivent résister à divers éléments perturbateurs tels que les pannes, les atteintes à la confidentialité et les agressions volontaires ou involontaires provenant de l'entreprise et de son environnement.

Nos travaux de recherche s'inscrivent dans le cadre de la sécurité des systèmes d'information et nous nous intéressons plus particulièrement aux systèmes décisionnels. Ces systèmes facilitent le pilotage des plans d'actions par les décideurs de l'entreprise. Ils présentent des spécificités en matière de traitements et de données qui les rendent très vulnérables à différents risques qui peuvent toucher aux données sensibles des entreprises. Tout dysfonctionnement pourrait ainsi entrainer des conséquences néfastes pour la gouvernance de l'entreprise. D'autre part, les dernières évolutions, montrent que les systèmes d'information traitent de plus en plus des données massives

associées aux principes du « Big Data » : volume, variété et vélocité (Gandomi et Haider, 2015). Ceci induit une complexité des traitements et des données qui rend la protection de ces systèmes et la prévention des risques particulièrement aléatoires et généralement peu efficaces.

La société Trimane est une ESN spécialisée dans le développement d'applications décisionnelles dans un environnement Internet. Nos travaux se focalisent sur la sécurité des sites web dédiés à la prise de décision. Ces sites présentent la particularité de recevoir des requêtes via des réseaux de communication internes ou externes à l'entreprise. Ils sont donc particulièrement vulnérables aux agressions extérieures qui ne cessent d'évoluer et s'avèrent très difficiles à prévenir.

2. État de l'art

De nombreuses solutions concernent la sécurité des systèmes d'information et se répartissent en deux grandes catégories : les systèmes de détection et les systèmes de prévention d'intrusions. Elles permettent de surveiller le trafic lié au système d'information dans le but d'identifier et d'empêcher les actes malveillants. Ces systèmes disposent de deux approches différentes : l'approche d'analyse comportementale et l'approche à base de signatures (Dagorn, 2006). Nous nous intéressons aux systèmes à base de signatures. Une signature est définie par un ensemble de règles décrivant une attaque connue. Le principe de ces systèmes consiste à rechercher dans le trafic, les correspondances avec les signatures connues. Cette approche n'est efficace que pour les attaques qui sont déjà répertoriées dans la base de signatures. Par conséquent, cette base doit être alimentée à chaque nouvelle attaque. Cependant, cette mise à jour est toujours pénalisée par un temps de retard qui est dû à la création manuelle de signatures considérée comme une tâche fastidieuse pour les experts.

Plusieurs travaux de recherche ont abordé le problème de l'automatisation de la mise à jour de la base de signatures lors de nouvelles attaques. L'article de (Esposito *et al.*, 2005) a proposé un processus de reconnaissance de signatures à partir d'un trafic d'attaques extrait manuellement du trafic global. Les auteurs de l'article (Hamdi *et al.*, 2012) ont proposé une autre approche basée sur la programmation logique inductive. Cette technique consiste à générer automatiquement des signatures à partir d'une classification du trafic en deux catégories : la classe du trafic appartenant à l'attaque et la classe du trafic complémentaire (ne contenant pas d'attaque). Cependant, dans les travaux de (Hamdi *et al.*, 2012), la classification est manuelle puisqu'elle nécessite l'intervention d'un expert. Récemment, l'article de (Godefroy *et al.*, 2014) a proposé une nouvelle approche qui consiste à décrire les attaques sous forme d'arbres et à établir des règles de corrélation à partir de ces arbres. Cependant

l'intervention d'un expert reste, encore une fois, nécessaire pour la construction des arbres d'attaques.

Dans tous les travaux que nous venons de citer, l'évolution de la base de signatures repose généralement sur l'intervention d'un expert du domaine. Ceci provoque un décalage au niveau de la mise à jour de la base de signatures. Par conséquent, les solutions proposées jusqu'à présent s'avèrent insuffisantes pour assurer une sécurité efficace des systèmes d'information décisionnels.

3. Problématique

L'objectif de notre travail est d'améliorer la sécurité des systèmes d'information décisionnels sous forme de sites web. Comme souligné précédemment, plusieurs travaux ont tenté d'améliorer les performances des systèmes à base de signatures. Néanmoins, ils ont présenté certaines limites qui les rendent insuffisantes pour assurer la sécurité des systèmes d'information de façon satisfaisante. Nous souhaitons donc compléter et adapter les travaux proposés auparavant afin de protéger au mieux ces systèmes et mettre en place des protections pérennes contre les perturbations externes.

4. Actions réalisées

Je prépare mon doctorat dans le cadre d'une bourse CIFRE. L'entreprise m'a confié la mission d'administrer la sécurité de ses sites web. Ceci me permet de travailler au quotidien sur des cas concrets en relation avec la sécurité des systèmes d'information.

J'ai aussi réalisé une étude bibliographique qui s'articule autour de trois domaines :

- les systèmes décisionnels : les modèles, méthodes et outils décisionnels ;
- le Big data pour stocker et analyser en temps réel de gros volumes de données ; par exemple, les données de réseau et de communication, les activités du système d'information, le comportement des usagers.
- la cybersécurité : les méthodes et les techniques actuelles pour la sécurité des systèmes d'information, leurs atouts et limites ainsi qu'une comparaison entre les différentes méthodes.

J'étudie également la problématique de la cybersécurité appliquée aux systèmes d'information décisionnels. L'exploitation en temps réel de flux d'information (principalement des requêtes) contenant des signaux faibles d'attaques, nécessite des techniques de filtrage performantes. De plus, l'analyse de grandes masses de données permettant de prévenir des attaques futures, requiert de nouveaux outils d'aide à la décision qui pourraient reposer sur des techniques d'apprentissage.

5. Actions futures

Nous allons étendre certains travaux étudiés dans l'état de l'art, notamment ceux de (Hamdi *et al.,* 2012) et (Godefroy *et al.,* 2014). Nous envisageons de spécifier et mettre en place un processus pour corréler les menaces et automatiser la mise à jour de la base de signatures. Ensuite, il parait nécessaire d'évaluer l'efficacité de la base de signatures pour détecter les attaques futures. L'évaluation nécessite une phase d'expérimentation qui correspond à simuler un trafic sain et un trafic malveillant. Nous évaluons ainsi le comportement du système à base des signatures ajoutées face au trafic subi.

Dans un deuxième temps, nous aborderons le problème d'identification des nouvelles attaques. Nous envisageons de suggérer des solutions capables de répondre efficacement aux signaux faibles générés par la préparation des attaques. En effet, un verrou à résoudre est que des signaux détectés ne soient pas significatifs. Enfin, les méthodes proposées seront expérimentées et évaluées par des tests portant sur des applications réelles.

6. Bibliographie

Dagorn N., "Intrusion Detection for Web Applications (Short Version)", *International Conference on Security and Cryptography*, 2006, p. 32-39.

Esposito M., Mazzariello C., Oliviero F., Romano S. P., Sansone, C., "Evaluating Pattern Recognition Techniques in Intrusion Detection Systems", *PRIS*, 2005, p. 144-153.

Gandomi A., Haider M., "Beyond the hype: Big data concepts, methods, and analytics". *International Journal of Information Management*, 2015, vol. 35, n° 2, p. 137-144.

Godefroy E., Totel E., Majorczyk F., Hurfin M., "Génération automatique de règles de corrélation pour la détection d'attaques complexes", *9ème conférence sur la Sécurité des Architectures Réseaux et des Systèmes d'Information (SAR-SSI)*, 2014, p. 10.

Hamdi O., Maissa M., Krief F., "Génération automatique de signatures par apprentissage pour les systèmes de détection d'intrusions", *7ème Conférence sur la Sécurité des Architectures Réseaux et Systèmes d'Information*, Cabourg, 22-25 mai 2012.

Gestion adaptative des contenus numériques

Proposition d'un framework générique par apprentissage et scénarisation dynamique.

Damien Mondou

Laboratoire L3I
Faculté des Sciences & Technologies Avenue Michel Crépeau
17042 La Rochelle

damien.mondou@univ-lr.fr

MOTS-CLÉS : *Apprentissage, analyse de comportement, adaptation dynamique*

KEYWORDS: *Learning, behavior analysis, dynamic adaptation*

ENCADREMENT. *Arnaud Revel(PR) et Armelle Prigent(MCF)*

1. Contexte

L'analyse des comportements des visiteurs d'un site web ou des touristes au travers de l'utilisation d'une application mobile dédiée pose une problématique commune : l'extraction rapide d'un parcours type afin de remonter jusqu'à une caractérisation de profils d'utilisateurs et de leurs comportements afin de réaliser une adaptation dynamique du contenu au regard de ces comportements et du contexte spécifique de la navigation. Différents travaux ont permis de s'intéresser aux moyens d'acquérir et d'analyser dynamiquement les données de navigation dans les contenus.Deux méthodes permettent de répondre à cette problématique : l'une à base d'**apprentissage** (de type algorithme par renforcement) permettant d'extraire des informations représentatives sur la base de logs de navigations passées et l'autre à base d'**observation dynamique du comportement** de l'utilisateur. Nos travaux ont donc pour objectif de concevoir un cadre générique d'analyse des comportements des utilisateurs et d'adaptation dynamique des contenus utilisant une approche hybride mêlant apprentissage et observation temps réel. Le domaine d'application des travaux pour validation du modèle concerne

principalement les expériences interactives dans les musées mais d'autres sujets sont étudiés (gestion de crise, narration interactive, pilotage de système multi-agents(SMA)).

2. État de l'art

Qu'il s'agisse de la narration dans l'univers des jeux vidéos, de la navigation web ou encore des outils mobiles pour les expériences pervasives (applications ludo-pédagogiques dans les lieux de patrimoine par exemple), l'adaptation dynamique est un atout majeur des systèmes pour garantir une expérience de qualité. Nous pouvons classer les différentes approches en trois grandes classes : **les approches scénarisées**, **les approches basées sur des agents** et **les approches hybrides**.

Les approches scénarisées (Vega *et al.*, 2003), (Delmas *et al.*, 2007) reposent sur un système expert contrôlant l'expérience de l'utilisateur. Elles génèrent la narration se rapprochant le plus des attentes de ce dernier au regard de l'observation qui en est faite. Cette génération dynamique nécessite un système d'observation et un système décisionnel. La modélisation d'une telle architecture est complexe car l'ensemble des chemins possibles pour l'utilisateur doit être défini par le concepteur. Les approches basées sur des agents permettent quant à elles une plus grande modularité dans la conception de l'expérience et des règles d'adaptation. Ces agents autonomes possèdent leurs propres comportements permettant d'interagir avec les autres agents et vers l'environnement dans lequel ils évoluent (Mateas *et al.*, 2003). Néanmoins, cette approche permet plus difficilement de maîtriser la qualité et la cohérence du modèle généré (chose que les approches scénaristiques garantissent par leur structure même). Pour palier aux manques des deux approches précédentes, les approches hybrides tentent de trouver le point d'équilibre entre liberté de l'utilisateur (et donc un niveau élevé d'adaptation) et respect de la trame du concepteur (c'est à dire de conserver une forme de scénario dans la diffusion). (Brian Magerko, 2003) propose un système de narration interactive, où l'auteur spécifie un cadre générique sans limiter les actions de l'utilisateur. Le contrôle de la qualité du système est réalisé par un gestionnaire qui analyse dynamiquement les comportements des utilisateurs, afin de détecter des comportements non prévus. Dans ce cas, le système modifie le comportement des personnages et l'univers dans lequel se déroule la scène pour encourager l'utilisateur à rester dans le cadre proposé par l'auteur.

Ces trois approches présentent des améliorations possibles. La difficulté majeure réside dans la représentation de l'ensemble des interactions possibles entre l'utilisateur et le système. Il convient de disposer d'un modèle garantissant un fort niveau de modularité, de réutilisabilité et suffisamment extensible pour répondre aux modifications potentielles. De plus, la vue que peut avoir le concepteur est susceptible de ne pas représenter l'ensemble des comportements possibles de l'utilisateur : l'observation pour l'adaptation ne doit pas se limiter à la détection du contexte mais doit prendre en compte les comportements propres de l'utilisateur (n'étant peut-être pas inclus dans le modèle). Nos travaux s'orientent donc vers une méthode basée sur un modèle haut niveau d'adaptation des contenus (approche hybride tel que cité précédemment) permet-

tant d'intégrer de nouveaux comportements et de nouveaux profils par un apprentissage réalisé sur l'ensemble des observations.

3. Problématique

La question est donc la suivante, comment combiner au sein d'un même système la supervision, le pilotage de l'activité (approche **top-down**) et l'intégration dans le pilotage de règles déterminées sur la base de l'observation du comportement réel des utilisateurs (approche **bottom-up**) ?

L'approche **bottom-up** permet, à travers une phase d'apprentissage, d'extraire des informations pertinentes en analysant un système interactif pendant son exécution. Les techniques d'apprentissage par renforcement (Kaelbling *et al.*, 1996) offrent un processus itératif et exploitable tout au long de son exécution. Ces techniques ont pour but de trouver une séquence idéale pour atteindre un but dans un environnement donné. L'approche **top-down** vise quant à elle, à produire une représentation d'un système à partir d'un profil (utilisateur ou de contexte). (Wang *et al.*, 2010) proposent un nouveau langage, ADAM, pour spécifier les structures adaptatives. Celles-ci sont définies au travers de trois dimensions, spatiale, temporelle et adaptative (gestion des exigences d'adaptivités de langue, qualité et d'interaction), et seront utilisées pour générer un contenu en adéquation avec le profil de l'utilisateur. Également basé sur le profil utilisateur, (De Virgilio, 2012) propose un langage de modélisation, AML, pour la conception adaptative d'applications web au niveau conceptuel. Sur la base de primitives élémentaires, la réponse et la structuration de l'information apportée à l'utilisateur est adaptée au contexte d'exécution de l'application.

Ainsi, contrairement aux approches classiques visant à proposer une adaptation sur la base d'un modèle et d'un ensemble de règles de re-scénarisation de l'activité, nous souhaitons ici proposer une approche hybride basée sur un pilotage top-down combiné à une analyse bottom-up. L'objectif étant de faire converger ces deux approches et de proposer une boucle d'itération du modèle entre apprentissage et représentation formelle afin d'améliorer la connaissance et de l'intégrer dynamiquement au modèle par retour de pertinence.

Les différents domaines d'application présentent chacun des spécificités. Ainsi, si dans le cadre de la narration l'interaction entre entités est importante, l'application du modèle aux expériences interactives dans les lieux de culture nécessite d'augmenter le modèle pour la prise en compte de la dimension spatiale.

4. Actions réalisées

Nous reprenons les travaux de (Rempulski, 2013) (effectués au L3I) qui proposent une représentation à base d'automates à entrées sorties de systèmes complexes. Le modèle manipule de nombreuses entités, en proposant au concepteur une modélisation simplifiée à base de paradigmes tels que la modularité, le multiple-héritage et la réutili-

sation de composants. Sur cette base, nous développons actuellement une interface entre l'outil de modélisation (#Telling) issu de ces travaux et un SMA (GAMA (Taillandier *et al.*, 2012))Tout comme notre système, les SMA manipulent des agents autonomes interagissant entres eux sur la base d'un ensemble de comportements. Cette intégration proposera une approche adaptative des SMA, basée sur un outil de modélisation. De plus, des améliorations du modèle pour augmenter son expressivité ont été réalisées afin d'entamer les travaux de modélisation d'un serious-game dans les musées (mini-jeux proposés par des robots Nao pilotés par #Telling).

5. Actions futures

Nous souhaitons intégrer prioritairement l'approche bottom-up dans notre outil afin d'apprendre à notre modèle de nouveaux comportements, que l'on fera valider par le concepteur dans un premier temps. Nous souhaitons, à terme, intégrer les contraintes temporelles que ce soit au niveau de la détection du contexte ou de la prise de décision (intégration de la sémantique du temps des modèles UPPAAL (Larsen *et al.*, 1997) dans le modèle de (Rempulski, 2013)).

Une autre problématique est l'analyse des comportements au travers de modalités de navigation variées. L'objectif est de proposer un framework unifié d'observation et d'analyse des comportements sur un domaine de navigation dont le périmètre peut être modifié en fonction des besoins. Une suite des travaux à venir est d'extraire des traces d'exécutions à partir du modèle de pilotage. Pour conclure, l'intégration de l'approche top-down nous permettra d'aboutir à un système complet permettant également d'adapter le contenu à partir du profil utilisateur ou du contexte.

6. Bibliographie

Brian Magerko J. L., « Building an interactive drama architecture », 2003.

De Virgilio R., « AML : A modeling language for designing adaptive web applications », 2012.

Delmas G., Augeraud M., « Plot Monitoring for Interactive Narrative Games », p. 17–20, 2007.

Kaelbling L. P., Littman M. L., Moore A. W., « Reinforcement Learning : A Survey », 1996.

Larsen K. G., Pettersson P., Yi W., « UPPAAL in a Nutshell », 1997.

Mateas M., Stern A., Tech G., « Façade : An Experiment in Building a Fully-Realized Interactive Drama », 2003.

Rempulski N., « Synthèse dynamique de superviseur pour l ' exécution adaptative d ' applications interactives », 2013.

Taillandier P., Vo D.-a., Amouroux E., « GAMA : a simulation platform that integrates geographical information data , agent-based modeling and multi-scale control », 2012.

Vega S. N., L., « A Petri Net Model for the Analysis of The Ordering of Actions in Computer Games », *GAME ON 2003*, 2003.

Wang C., Wang D. Z., Lin J. L., « ADAM : An adaptive multimedia content description mechanism and its application in web-based learning », 2010.

Authentification de documents par la vérification des informations textuelles

Chloé Artaud

L3I (Laboratoire Informatique, Image, Interaction) et DGA
Avenue Michel Crépeau
17042 La Rochelle cédex 1
chloe.artaud@univ-lr.fr

MOTS-CLÉS : recherche d'information, extraction d'information, ingénierie des connaissances, faux documents

KEYWORDS: Information Retrieval, Information Extraction, Knowledge Engineering, Fraudulent Documents

ENCADREMENT : Antoine Doucet (PR) et Jean-Marc Ogier (PR)

1. Contexte

En 2015, 68% des entreprises françaises interrogées dans le cadre d'une étude de PricewaterhouseCoopers[1] déclarent avoir été victimes d'une fraude au cours des 24 derniers mois (changements de RIB, fraudes comptables, fraudes aux achats...) Dans un autre registre, les administrations publiques sont également victimes de fraudes à travers de fausses déclarations. La Sécurité sociale a ainsi détecté en 2013 un préjudice de 350,5 millions d'euros lié à la fraude aux prestations sociales[2]. Les institutions ne sont pas les seules à être victimes de fraudes : les particuliers eux-mêmes peuvent être victimes d'arnaques. Lors d'un achat de voiture, certains fraudeurs peuvent par exemple falsifier la carte grise en changeant la date de la voiture, pour l'achat d'une maison, les diagnostiques énergétiques…

Les documents frauduleux jouent souvent un rôle important dans les fraudes, quel que soit le domaine ou le type de document, et doivent donc être particulièrement contrôlés dans le cadre de la lutte contre les fraudes. Si des solutions existent pour sécuriser et contrôler les documents d'identité, les autres types de documents, qu'ils soient électroniques ou papier, souvent en format A4 et simplement imprimés, sont facilement modifiables. Le sujet de notre travail s'inscrit

[1] Global Economic Crime Survey 2016 "La fraude explose en France. La cybercriminalité au cœur de toutes les préoccupations", PwC

[2] http://www.securite-sociale.fr/La-fraude-sociale

donc dans ce contexte : il s'agit de détecter les documents faux ou falsifiés, notamment en vérifiant les informations qu'ils contiennent.

2. État de l'art

La détection des fraudes sur les documents est très peu étudiée, malgré des besoins importants. Nos recherches traversent donc plusieurs domaines, de l'analyse de document à l'ingénierie des connaissances en passant par l'analyse d'image, la recherche d'information et les *digital forensics* (que l'on pourrait traduire par « informatique légale », par analogie à « médecine légale »).

La plupart des travaux sur la détection des faux documents concernent des indices graphiques, comme la différence d'inclinaison, de taille, d'alignement ou de bruit d'un caractère par rapport aux autres (Bertrand *et al.*, 2013), la différence de police ou d'espacements des caractères au sein d'un mot (Bertrand *et al.*, 2015), le décalage d'une ligne par rapport aux marges (van Beusekom *et al.*, 2010) ou encore l'inclinaison différente des lignes les unes par rapport aux autres (van Beusekom *et al.*, 2009). Dans ces travaux, l'hypothèse de départ est que les fraudeurs doivent souvent modifier les éléments du document dans la précipitation et que la modification n'est donc pas toujours parfaite, ce qui permet de la détecter.

D'autres travaux en analyse d'image permettent de détecter ou de suspecter des modifications, comme une étude sur les imprimantes et scanners utilisés dans la fabrication du document (Elkasrawi *et al.*, 2014) ou la vérification des tampons (Micenková *et al.*, 2015). (Poisel *et al.*, 2011) dresse un état de l'art des recherches de fraudes sur les données multimédias, dont celles sur les images. Plus largement, de nombreux travaux existent sur les modifications des images de scènes naturelles.

Parallèlement à l'analyse d'image, nous nous intéressons aux informations contenues dans le document. Si les journalistes s'intéressent de plus en plus au *fact checking* (vérification des faits), il est difficile de trouver des travaux scientifiques sur la vérification d'information. Nous pouvons cependant citer (Goasdoué *et al.*, 2013) et leur outil FactMinder, à la croisée de l'extraction d'information, de la recherche d'information sur le Web et de l'ingénierie des connaissances. Nous pourrons également nous inspirer des travaux sur la détection de plagiat.

3. Problématique

Si la recherche sur les faux documents se concentre essentiellement sur des analyses graphiques, nous pensons qu'il est possible de lier leur détection à celle de faux contenus ou de contenus incohérents. En effet, le contenu d'un document est composé d'informations diverses, textuelles et graphiques, ayant toutes un sens. Ces informations sont liées sémantiquement les unes aux autres et peuvent être vérifiées, seules et collectivement. Par exemple, l'adresse d'une entreprise est liée à son nom, et souvent à un numéro de téléphone, un numéro de SIRET…

Aujourd'hui, lorsqu'une personne veut vérifier une information, son premier réflexe est de faire une recherche sur Internet, à l'aide de moteurs de recherche. Généralement, lorsque les mots de la requête sont bien choisis, la page de résultats (SERP, pour *Search Engine Results Page*) donne une idée de la véracité de l'information. Sinon, il faut explorer les liens ou reformuler la requête. Nous envisageons d'automatiser complètement ce processus afin de le simplifier et de l'intégrer à notre système de détection des faux documents. Nous pensons qu'il serait intéressant de créer un système capable d'extraire automatiquement les informations du document et de les comparer avec l'information accessible en ligne.

4. Actions réalisées

Nous avons élaboré le plan d'un processus qui permettrait de vérifier les informations simples. Il s'agit tout d'abord d'extraire les informations du document et de les annoter afin d'en peupler une ontologie. Nous pouvons pour cela créer des règles d'extraction pour chaque information récurrente des documents. Cette extraction peut se faire avec des outils comme Unitex.

Une fois les informations extraites et identifiées, il s'agit d'en faire des requêtes pour un moteur de recherche. Les requêtes peuvent être composées d'une seule information, qui correspondra alors à une instance de l'ontologie (comme « 12345678901234 », instance de la classe « numéro SIRET »), ou de plusieurs mots combinant instances et classes (comme « numéro SIRET Exentrep » qui associe la classe « numéro SIRET » et l'instance « Exentrep » de la classe « nom société »). Nous avons choisi d'utiliser une ontologie pour pouvoir justement associer des informations grâce aux propriétés.

Ces requêtes permettent de récupérer une SERP dont on peut extraire les informations, en utilisant les mêmes règles d'extraction que précédemment. Il faut ensuite vérifier que les informations extraites sont identiques en les normalisant. Si elles sont identiques, elles deviennent une information unique que nous pouvons comparer avec celle du document à vérifier. S'il y a plusieurs résultats différents, il faudra déterminer lequel est le plus fiable. On peut vérifier les différents résultats en les requêtant à leur tour et en analysant les résultats de la recherche. Nous pouvons ensuite comparer les informations du document et celles de la SERP en utilisant des méthodes de calcul de similarité et affecter un indice de fiabilité à chaque information.

5. Actions futures

Le processus proposé se fera en deux phases : nous chercherons d'abord à évaluer chaque étape en les réalisant manuellement, puis nous automatiseront intégralement le processus, si la première phase se révèle concluante. Il s'agira donc à terme d'automatiser le processus tout en laissant à l'utilisateur la possibilité de le superviser et d'intervenir pour corriger d'éventuelles erreurs. Il faudra également

évaluer les performances de ce système en comparant les résultats avec ceux obtenus manuellement.

Nous pensons également explorer les liens proposés dans la SERP : nous pouvons pour des informations plus complexes avoir besoin de fouiller le web plus en profondeur. Ainsi, nous pouvons imaginer un système qui tenterait d'extraire les informations sur chaque page explorée et continuer à explorer les liens tant qu'il n'a pas trouvé suffisamment d'informations pertinentes.

Dans un dernier temps, nous chercherons à associer ces indices sémantiques aux indices graphiques décrits dans l'état de l'art pour réaliser un système de détection des fraudes complet où l'utilisateur n'aura qu'à entrer le document à vérifier pour que le système lui donne un taux de fiabilité pour ce document.

6. Bibliographie

Bertrand, R., Gomez-Kramer, P., Terrades, O. R., Franco, P., Ogier, J. M., " A system based on intrinsic features for fraudulent document detection ", *International Conference on Document Analysis and Recognition (ICDAR)*, 2013, p. 106-110.

Bertrand, R., Terrades, O. R., Gomez-Kramer, P., Franco, P., Ogier, J.-M., " A Conditional Random Field model for font forgery detection ", *13th International Conference on Document Analysis and Recognition (ICDAR),* 2015, p. 576-580.

Elkasrawi, S., Shafait, F., " Printer Identification Using Supervised Learning for Document Forgery Detection ", *11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 146-150.

Goasdoué, F., Karanasos, K., Katsis, Y., Leblay, J., Manolescu, I., Zampetakis, S., " Fact checking and analyzing the web ", *International Conference on Management of Data (SIGMOD '13)*, 2013, p. 997-1000.

Micenková, B., van Beusekom, J., Shafait, F., " Stamp verification for automated document authentication ", *Computational Forensics*, 2015, 8915, p. 117-129.

Poisel, R., Tjoa, S., " Forensics investigations of multimedia data: A review of the state-of-the-art ", *6th International Conference on IT Security Incident Management and IT Forensics (IMF 2011)*, 2011, p. 48-61.

van Beusekom, J., Shafait, F., Breuel, T., " Automatic Line Orientation Measurement for Questioned Document Examination ", *Computational Forensics*, 2009, 5718, p. 165-173.

van Beusekom, J., Shafait, F., Breuel, T. M., " Document inspection using text-line alignment ", *9th IAPR International Workshop on Document Analysis Systems*, 2010, p. 263-270.

Toward free spam social networks: detecting and tracking spammers in Twitter

Mahdi Washha

Institut de Recherche en Informatique de Toulouse
Avenue de l'étudiant
31400 Toulouse

Mahdi.Washha@irit.fr

MOTS-CLÉS : *Réseaux Sociaux, Spam, Utilisateurs Légitimes, Machine Learning, Temps.*

KEYWORDS: *Social Networks, Spam, Legitimate Users, Machine Learning, Time.*

ENCADREMENT. *Florence Sedes (PR)*

1. Context

Social networks are now so well established with playing an important role at communication level between people. However, the propagation of noisy information, so-called "spam", in social networks is growing up daily at unusual rates where unethical goals stand behind publishing spam information. The spreading of spam affects negatively in different aspects, summarized in : (i) polluting real-time search ; (ii) interfering on statistics computed by mining tools ; (iii) consuming significant resources from both humans and systems ; (iv) decreasing the performance of search engines that use explicitly social signals ; and (v) violating the privacy of user's which occurs because of viruses and phishing methods. Thus, this research aims at finding out solutions suitable for the applications that take social networks as a source of information. Also, we target to overcome the existing limitations in the current state of art solutions.

2. State of art

2.1. *Spam definition and spammers' goals*

The first appearance of spam problem has been in electronic messaging systems. Researchers have defined the spam as unsolicited and undesired messages sent to the systems users by unethical individuals, so called "spammers" (Qaroush *et al.*, 2012). However, spammers have extended their targets to include social networks area because of the huge volume of users, estimated at hundreds of millions.

Spammers have a set of goals, e.g. spreading advertisements to generate sales, disseminating pornography, viruses and phishing. They achieve their spamming behavior

through exploiting the services that social networks provide such as hashtags and shorten URLs. Also, spammers leverage the availability of free APIs provided by social networks to automate the publishing of spam.

2.2. *Implicit solutions provided by social networks*

The social networks administrators have attempted to break up the spam phenomenon. Their attempts are centered around using users' reports, and defining general rules. Exploiting the users' reports can be viewed as a kind of manual collaboration. However, such an attempt needs a manual effort from both administrators (review sent reports) and users (send report about accounts). The rules concept can contribute in detecting spammers by suspending permanently upon violating the rules defined. However, the current social networks like Twitter use general rules such that spammers can easily bypass them.

2.3. *Automated spam detection solutions*

The limitations in the solutions of social networks have motivated researchers to propose more powerful methods. We classify their methods into two approaches based on the degree of automation in detecting spam content or spammers : (i) machine learning approach ; (ii) social honeypot approach. In this paper, we just mention the related work of machine learning approach as a fully automated one, because of the inefficiency of social honeypot based solutions.

The existing solutions use the concept of features combined with learning algorithms to automate the detection process. However, as main differences between the solutions, machine learning techniques have been employed at three levels of detection : (i) post-level ; (ii) user-level ; and (iii) campaign-level.

Post-level. At this level, Martinez-Romo and Araujo (Martinez-Romo *et al.*, 2013) applied probabilistic language models to determine the topic of the considered post. Then, they label the post as spam when it is too diverging from the potential topic. Benevenuto (Benevenuto *et al.*, 2010) identified spam tweet through extracting a set of features like number of words from each tweet individually. They applied then the Support Vector Machine (SVM) learning algorithm on manually annotated data-set to have a binary classifier.

User-level. The work investigated in (Benevenuto *et al.*, 2010) proposed features associated to the account, including number of followers, number of friends, similarity between tweets posted or re-tweeted by the user, ratio of URLs in tweets. Much more complex features related to the graph theory have been extracted at the user level as well. For example, (Yang *et al.*, 2011) examined the relation between users by using graph metrics as features, including local clustering, node betweenness, and bi-directional relation ratio. Song, Lee, and Kim (Song *et al.*, 2011) studied some relational features such as the distance and the connectivity between sender and receiver(s) of a given message. In the same level, user profiling methods and community clustering algorithms (Mezghani *et al.*, 2015) can contribute better than using abstract graph metrics. However, the time complexity issue in such methods prevents to exploit them in fighting spammers.

Campaign-level. Chu et al. (Chu *et al.*, 2012) clustered users' accounts based on the URLs retrieved from their posted tweets. Then, a set of features from the clustered accounts were extracted to be incorporated in identifying spam campaign. Indeed, this level is not applicable to detect individual spammers as well as spammers can lunch uncorrelated small campaigns to avoid detection.

3. Problem statement

State of art solutions are still not robust enough to adapt the changes of spammers' tactics in propagating spam information. More precisely, post-level methods are an inefficient solution to identify spammers because one post does not provide "informative" information to distinguish the spammers' behavior from the legitimate users' behavior. Therefore, relying on such way in detecting spam content is not a practical solution.

The user-level methods have critical limitations and major drawbacks derived from using features that are easy to manipulate. Thus, spammers can avoid the detection when using such features. As an example, the number of followers (i.e. the accounts that follow a user) is mainly used in detecting spammers as a feature. Based on the state of art conclusions, the small number of followers has high probability for being spammer. However, such a number can be easily increased by creating a huge number of recent accounts with letting each account to follow each other, and thus keep away from being classified as spammer. Also, although of graph features give high accuracy, but they remain incompatible with real time filtering. This incompatibility is because of the time complexity issue of graph metrics and the necessity to fetch all users of social network to get the metrics values.

At the campaign level, working at this level is effective to detect big campaigns only, not for individual spammers. Furthermore, spammers are intelligent enough to design spam campaigns such that the correlations between accounts are not detectable in easy ways.

With the negative impacts of spam on social networks and the limitations in the current state of art methods, our research problem is summarized in answering the following research questions :

– To what extent can the creation date of the account and the posting date of the post contribute in detecting spammers, as unmodifiable attributes by social networks users ?

– Is it possible to have few and robust features suitable for real-time spammers detection ?

– Instead of searching in the all users of social networks, is it possible to define a heuristic function such that it can predict the names of spam accounts, as a way to locate and track spammers ?

4. Work in progress

In addressing the problem of identifying spammers, we propose a design of new generic features suitable for real-time filtering. Our features are distributed between statistical features, behavioral features. The statistical ones incorporate explicitly both the time of posting tweet and the creation date of user's account. We choose the time and date attributes because they are unmodifiable attributes overtime. Indeed, this explains our motivation in leveraging them. The behavioral features can catch any potential posting behavior

similarity over time between different instances (e.g. hashtags) available in the user's posts (e.g. correlation between different hashtags (#h1 and #h2) posting time). Our features are inspired by the following hypotheses : (i) spammers tend to create recent accounts ; (ii) spam accounts follow each other to boost up some critical attributes such as number of followers ; (iii) spammers have systematic and defined posting pattern.

We validate the generic features proposed using Twitter social network as an example. In doing so, we crawled a data-set consisting of 7,100 users from Twitter social network. Then, we annotated manually the crawled users by assigning a label (spammer or non-spammer) for each user. This forms a data set available as ground truth for other assessments. Using the annotated data-set, we employed machine learning algorithms to build a binary classifier using the crawled data-set and features designed. According to the experiments, our new features are able to classify correctly the majority of spammers with a detection rate higher than 93% when applying Random Forest as a classification algorithm. The results obtained outperform by 6% the detection rate of 70 state of art features proposed in (Benevenuto *et al.*, 2010). This work is going to be published as one contribution in spam detection area.

5. Future works and perspectives

The next steps will be dedicated to finding a simple and fast method to locate and track spammers in a social network. In doing so, we plan to study the correlation between spammers' names and the current occurring events in social networks. This may help in predicting the name of spam accounts, which forms an entry point to locate spammers. We motivate this correlation study based on a true observation found in our crawled data-set. The observation found states that spammers leverage often trending events to publish spam information using accounts having names inspired by the content and description of events.

6. References

Benevenuto F., Magno G., Rodrigues T., Almeida V., « Detecting spammers on twitter », *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*, 2010.

Chu Z., Widjaja I., Wang H., « Detecting social spam campaigns on twitter », *Applied Cryptography and Network Security*, Springer, p. 455–472, 2012.

Martinez-Romo J., Araujo L., « Detecting malicious tweets in trending topics using a statistical analysis of language », *Expert Systems with Applications*, vol. 40, n° 8, p. 2992–3000, 2013.

Mezghani M., On-at S., Peninou A., Canut M.-F., Zayani C. A., Amous I., Sedes F., *New Trends in Databases and Information Systems : ADBIS 2015, France, September 8-11, 2015. Proceedings*, Springer International Publishing, chapitre A Case Study on the Influence of the User Profile Enrichment on Buzz Propagation in Social Media : Experiments on Delicious, p. 567–577, 2015.

Qaroush A., Khater I. M., Washaha M., « Identifying spam e-mail based-on statistical header features and sender behavior », *Proceedings of the CUBE International Information Technology Conference*, ACM, p. 771–778, 2012.

Song J., Lee S., Kim J., « Spam filtering in twitter using sender-receiver relationship », *Recent Advances in Intrusion Detection*, Springer, p. 301–317, 2011.

Yang C., Harkreader R. C., Gu G., « Die Free or Live Hard ? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers », *Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection*, RAID'11, Springer-Verlag, Berlin, Heidelberg, p. 318–337, 2011.

Intégration du contexte spatio-temporel et social pour l'analyse de sentiments sur Twitter

Ophélie Fraisier

IRIT, 118 route de Narbonne, 31062 Toulouse, France
CEA Tech Midi-Pyrénées, 135 avenue de Rangueil, F-31400 France

Ophelie.Fraisier@irit.fr

MOTS-CLÉS : *Analyse de sentiments, Analyse de médias sociaux, Synthèse d'information*

KEYWORDS: *Sentiment analysis, Social media analysis, Information synthesis*

ENCADREMENT. *IRIT : Mohand Boughanem, Guillaume Cabanac et Yoann Pitarch ; CEA : Romaric Besançon*

1. Contexte

L'analyse de sentiments est un domaine attirant beaucoup d'attention depuis quelques années. Ceci est notamment dû au fait que l'explosion des médias sociaux a permis à n'importe quelle personne disposant d'un accès au net de s'exprimer publiquement. Le domaine bénéficiant d'un transfert très rapide vers le monde industriel, de nombreuses entreprises sont apparues sur le marché (Viralheat, Brand24, Linkfluence, ...).

Les sentiments sont des « états privés », car ils ne sont pas observables et vérifiables de manière objective. De nombreuses méthodes ont néanmoins été développées pour détecter et catégoriser ces états privés à partir de textes, en utilisant des techniques de traitement automatique des langues, des lexiques spécialisés et des méthodes d'apprentissage automatique (Liu, 2012). Les données analysées sont souvent issues de microblogs, dont la plateforme la plus connue et la plus utilisée est Twitter. Les messages qui y sont publiés s'appellent des tweets et sont limités à 140 caractères.

Les sentiments d'une personne peuvent évoluer dans le temps et être influencés par de nombreux facteurs. Ma thèse porte donc sur la problématique de détection de

sentiments à partir de tweets, tout en considérant des éléments contextuels tels que la localisation de l'auteur-e, l'horodatage des messages ou les relations sociales qu'il entretient avec les autres membres du réseau.

2. État de l'art

La recherche d'information est un domaine ayant toujours accordé une grande importance à l'évaluation des modèles développés. De nombreuses campagnes d'évaluation permettent de comparer son système de classification de sentiments à celui d'autres équipes de recherche sur des jeux de données communs. Parmi les campagnes d'évaluation internationales, les plus importantes sont SemEval et TREC (Ounis et al., 2006). Il existe également des campagnes francophones, telles que DEFT (Hamon et al., 2015).

Les approches de classification rencontrées dans ces campagnes peuvent se diviser en trois grandes familles : les approches lexicales basées sur des lexiques tels que SentiWordNet (Feldman, 2013), celles à base d'apprentissage utilisant des corpus annotés (Xia et al., 2011) et les approches distributionnelles définissant le sens d'un mot par rapport à son contexte (İrsoy & Cardie, 2014).

Certains travaux ont tentés d'enrichir leur analyse en mettant en relation les sentiments et le contexte dans lequel ceux-ci ont été émis. Allisio et al. (2013) ont notamment visualisé la répartition géographique des sentiments de leur corpus en les représentant sur une carte de l'Italie. Celli et Rossi (2012) ; Celli et Zaga (2013) ont pour leur part étudié l'influence de la personnalité de l'auteur-e sur ses sentiments et ses interactions avec les autres utilisateur-trice-s de Twitter. Pardo et Rosso (2013) ont eux mis en relation le sexe, les émotions et le style d'écriture de l'auteur-e.

Afin d'améliorer la prise en compte du contexte social d'un-e auteur-e, les liens d'abonnements entre utilisateur-trice-s – explicitement disponibles sur Twitter – peuvent être utilisés pour détecter des similarités. Shao et al. (2015) proposent notamment un algorithme se basant sur les distances entre nœuds du réseau pour étudier l'appartenance de l'auteur-e à une communauté.

3. Problématique

Les méthodes de classification présentées dans la section précédente se sont montrées efficaces mais ne prennent en compte que le contenu du tweet pour détecter le sentiment. Or, le sentiment éprouvé vis-à-vis de certains termes peut être intimement lié au point de vue de l'auteur-e (par exemple pour des termes tels que « capitalisme » ou « IVG »). L'hypothèse d'homophilie nous permet de considérer que les points de vue de personnes appartenant aux mêmes communautés sont similaires (Abbasi et al., 2014). Une analyse sociale – prenant en compte les relations de l'auteur-e, son âge, son sexe, le contexte spatio-temporel du message, etc. – pourrait donc apporter un éclairage nouveau sur certains messages. La problématique de ma thèse consiste

donc à assembler « analyse de sentiments » et « analyse sociale » afin de proposer un résumé multi-factoriel s'appuyant sur ces différentes composantes.

4. Actions réalisées

Afin de me familiariser avec les méthodes d'évaluation du domaine, j'ai réalisé un état de l'art des campagnes d'évaluation. Ceci a également permis de repérer quelles tâches pourraient être intéressantes, qu'il s'agisse de tâches actuelles auxquelles il serait possible de participer dans le futur ou de tâches passées pour lesquelles les jeux de données et les résultats sont disponibles (Fraisier, 2016). L'étape suivante a été de définir le front de recherche actuel afin de savoir où situer mon travail dans le domaine (voir section 2).

J'ai effectué une première expérimentation pour découvrir les algorithmes classiques, avec une approche à base d'apprentissage supervisé sur une tâche de classification automatique de tweets en opinions positives/négatives/neutres. Les données d'évaluation proviennent de la tâche 2 de la campagne SemEval 2013 et sont composées de 9655 tweets d'entraînement et de 3813 tweets de test. Les résultats sont présentés dans le tableau 1. On constate que le classifieur le plus efficace est le SVM à noyau sigmoïde avec un score F1 de 0,60 (le score du meilleur participant étant de 0,69).

Classifieur	Score F1	Classifieur	Score F1
SVM à noyau sigmoïde	**0,60**	Forêt aléatoire	0,55
SVM à noyau linéaire	0,59	Arbre de décision	0,55
Bayésien naïf multinomial	0,57		

Tableau 1. *Résultats de mes expérimentations sur la classification des tweets de la tâche 2 de la campagne SemEval 2013*

5. Actions futures

Par la suite, je vais utiliser mes premières expérimentations pour mettre en œuvre un modèle plus performant, en intégrant plus de caractéristiques. Les premiers éléments rajoutés seront des éléments propres au tweet (présence de ponctuation, d'émoticône, négation, etc.), avant d'introduire le contexte – et notamment la notion d'homophilie – dans le modèle. Après cela, je me servirai des résultats obtenus pour déterminer quelles caractéristiques conserver et comment les assembler formellement afin d'avoir un modèle pertinent. Je déterminerai également comment évaluer les différentes composantes de ce nouveau modèle. Afin de comparer notre modèle à d'autres, une participation à une campagne d'évaluation est fortement envisagée. Il pourrait s'agir de la tâche « Sentiment Analysis in Twitter » de SemEval ou la nouvelle tâche TREC combinant « Microblog » et « Temporal Summarization ».

Remerciements : Ce travail a été réalisé grâce à l'obtention d'un financement Contrat Laboratoire – Entreprise numéro 14050975 soutenu par la Région Languedoc – Roussillon Midi-Pyrénées.

Références

Abbasi, M. A., Zafarani, R., Tang, J., & Liu, H. (2014). Am I more similar to my followers or followees ? : analyzing homophily effect in directed social networks. In *ACM-HT* (pp. 200–205). ACM. doi: 10.1145/2631775.2631828

Allisio, L., Mussa, V., Bosco, C., Patti, V., & Ruffo, G. (2013). Felicittà : Visualizing and Estimating Happiness in Italian Cities from Geotagged Tweets. In *ESSEM@AI*IA* (Vol. 1096, pp. 95–106). CEUR-WS.org.

Celli, F., & Rossi, L. (2012). The Role of Emotional Stability in Twitter Conversations. In *Proc. Workshop on Semantic Analysis in Social Media* (pp. 10–17). ACL.

Celli, F., & Zaga, C. (2013). Be Conscientious, Express your Sentiment ! In *ESSEM@AI*IA* (Vol. 1096, pp. 140–147). CEUR-WS.org.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82. doi: 10.1145/2436256.2436274

Fraisier, O. (2016, février). Information Retrieval – Evaluation Campaigns. (IRIT/RR–2016–04–FR). Consulté sur `https://www.irit.fr/publis/IRIS/2016_R_F.pdf`

Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., & Grouin, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT). In *Actes de deft.* Caen, France.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on HLT*, *5*(1), 1–167. doi: 10.2200/S00416ED1V01Y201204HLT016

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., & Soboroff, I. (2006). Overview of the TREC-2006 Blog track. In *Text Retrieval Conference.* Consulté sur `http://trec.nist.gov/pubs/trec15/papers/BLOG06.OVERVIEW.pdf`

Pardo, F. M. R., & Rosso, P. (2013). On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. In *ESSEM@AI*IA* (pp. 34–46).

Shao, J., Han, Z., Yang, Q., & Zhou, T. (2015). Community Detection Based on Distance Dynamics. In *ACM-SIGKDD* (pp. 1075–1084). doi: 10.1145/2783258.2783301

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, *181*(6), 1138–1152. doi: 10.1016/j.ins.2010.11.023

İrsoy, O., & Cardie, C. (2014). Opinion Mining with Deep Recurrent Neural Networks. In *EMNLP* (pp. 720–728).

Intégration aux modèles de capacité et de maturité de processus logiciel de méthodes, techniques et normes de l'IHM

Taisa Guidini Gonçalves

Université de Valenciennes et du Hainaut Cambrésis, LAMIH – UMR CNRS 8201
Le Mont Houy
F-59313 Valenciennes CEDEX 9- France
Taisa.guidinigoncalves@etu.univ-valenciennes.fr

MOTS-CLÉS : *Modèle de capacité et de maturité de processus de logiciel, Ingénierie de l'Interaction Homme-Machine, Amélioration du processus logiciel.*

KEYWORDS: *Software process capability and maturity model, Human-Computer Interaction Engineering, Software process improvement.*

ENCADREMENT : *Christophe Kolski (PR) et Kathia Marçal de Oliveira (HDR)*

1. Contexte

Cette étude concerne deux grands domaines : (1) le génie logiciel (GL), en particulier les modèles de capacité et de maturité de processus logiciel (CMPL), et (2) l'Ingénierie de l'Interaction Homme-Machine (IHM). Les modèles de CMPL fournissent un ensemble de lignes directrices intégrées et exhaustives pour le développement des produits et services, soutenant les organisations dans la définition et l'amélioration de leurs processus.

Le modèle CMPL le plus connu est CMMI-DEV (*Capability Maturity Model Integration for Development*) (CMMI Product Team, 2010). Ce modèle est composé d'un ensemble de domaines de processus qui, à leur tour, ont un ensemble d'objectifs spécifiques jugés importants en vue d'une amélioration significative dans ce domaine. Chaque objectif spécifique est composé de pratiques spécifiques qui définissent « quoi » faire, mais pas « comment » le faire (e.g., pour le domaine de processus **Développement d'exigences**, un objectif est « Développer les exigences client », et une des pratiques pour cet objectif est « Expliciter les besoins »).

En parallèle, l'Ingénierie de l'Interaction Homme-Machine (IHM) a connu de grands progrès ces dernières années concernant la définition de méthodes, techniques et normes pour appuyer la conception, la construction et l'évaluation des systèmes interactifs. Nous pouvons citer, par exemple, des méthodes d'analyse des tâches (par ex. MAD, DIANE+ ou CTT ; cf. Limbourg et Vanderdonckt, 2003) ou

encore des techniques pour l'évaluation d'IHM (par ex. les tests d'utilisabilité ; cf. Dumas et Fox, 2009). De plus, des normes ont été également définies, telle la norme ISO/TR 18529 (2000).

Bien que le génie logiciel et IHM aient largement évolué ces dernières années, les deux domaines ont un impact différent dans la pratique. D'une part, des études montrent l'utilisation des modèles de GL dans la pratique (http://cmmiinstitute.com/resources/process-maturity-profile-july-2015) ; d'autre part, à notre connaissance, très peu de travaux signalent l'utilisation efficace et à large envergure de l'ingénierie de l'IHM dans l'industrie (cf. Bevan, 2009). Dans ce contexte, on considère que les modèles de CMPL contribuent largement à l'utilisation du GL dans l'industrie. Les professionnels utilisent les normes, modèles et méthodes de génie logiciel pour répondre aux objectifs et pratiques définis dans le CMPL. Cependant, on n'a pas le même constat avec les approches d'IHM.

2. État de l'art

À notre connaissance, il n'y a pas de travail qui intègre les pratiques d'IHM à des modèles de CMPL. Cependant, nous pouvons trouver plusieurs travaux relatifs à l'intégration du génie logiciel et de l'IHM, et également des modèles de maturité spécifiques pour l'utilisabilité.

Certaines propositions concernent la manière d'intégrer des techniques d'IHM dans le processus de développement de logiciel : Ferre *et al*. (2005) définissent un *framework* qui intègre les pratiques d'utilisabilité dans le processus logiciel. Fischer *et al.* (2011) proposent l'intégration de l'ingénierie de l'utilisabilité et de l'IHM à travers l'analyse des normes (ISO 12207 et ISO 9241-210). Le résultat de l'analyse présente des listes d'activités, des artefacts et des corrélations entre IHM et GL, qui ont été validées par des experts en IHM.

Plusieurs modèles de capacité et de maturité sont définis spécifiquement pour l'utilisabilité. Dans ce contexte, un *survey* (Jokela *et al.*, 2006) a présenté treize modèles de capacité et/ou de maturité d'utilisabilité. Ces modèles ont utilisé la structure du modèle CMMI-DEV (et sa version précédente) et de la norme ISO 13407. En général, ces modèles couvrent différents domaines organisationnels tels que : la performance des processus d'utilisabilité et la gestion des processus d'utilisabilité dans des projets de développement. Raza *et al.* (2012) ont proposé un modèle spécifique de maturité de l'utilisabilité pour le domaine des projets *open-source*, définissant onze facteurs pour l'utilisabilité. L'évaluation est effectuée pour tous les facteurs et l'estimation de l'organisation considère cinq niveaux de maturité.

3. Problématique

L'Ingénierie de l'IHM est intrinsèquement liée au GL car ce domaine s'occupe de l'ingénierie logicielle au sens large, et donc s'applique notamment aux systèmes

interactifs. Donc, une question se pose : Pourquoi les approches d'IHM ne sont-elles pas utilisées dans la pratique en industrie comme le sont les approches de GL ?

Notre hypothèse est que les professionnels qui suivent les modèles CMPL, n'utilisent pas (ou très peu) les pratiques d'ingénierie d'IHM dans le développement de système interactif autant que les pratiques d'ingénierie logicielle parce qu'ils ne les connaissent pas (ou qu'insuffisamment) et/ou parce qu'ils ne savent pas comment les utiliser.

Dans ce contexte, un objectif de cette thèse est de définir des lignes directrices, portant sur la conception, la construction et l'évaluation des systèmes interactifs, qui puissent supporter les professionnels qui utilisent le modèle de CPML. Nous soutenons en effet que l'intégration de ces connaissances dans des modèles de CMPL peut contribuer à une meilleure utilisation dans l'industrie.

4. Actions réalisées

Pour répondre à l'objectif de cette thèse nous proposons les étapes illustrées sur la Figure 1. Ces étapes incluent (voir Figure 1) : une recherche bibliographique - **étape (i)** *l'étude des modèles* de CMPL ; des propositions pour répondre à l'objectif **- étape (ii)** *l'identification des approches d'IHM* intégrées aux modèles CMPL et **étape (vi)** *définition de lignes directrices* ; des validations avec des experts **- étape (iii)** *validation des approches* et **étape (vii)** validation des lignes directrices ; un ensemble d'enquêtes réalisées dans le monde académique et industriel **- étape (iv)** *enquête sur l'utilisation de l'ingénierie de l'IHM en milieu académique* (réalisée en deux itérations) et **étape (v)** *enquête sur l'utilisation de l'ingénierie de l'IHM en milieu industriel*. Nous avons déjà réalisé les **étapes (i), (ii), (iii), (iv)** (itération 1). Après un premier ensemble d'interviews avec 20 experts en IHM pour valider les approches (**étape iv**), nous avons réalisé une première enquête en milieu académique (itération 2) pour une partie de notre proposition. Un résultat partiel des **étapes (i), (ii) et (iii)** a été récemment accepté comme article long dans la conférence RCIS 2016.

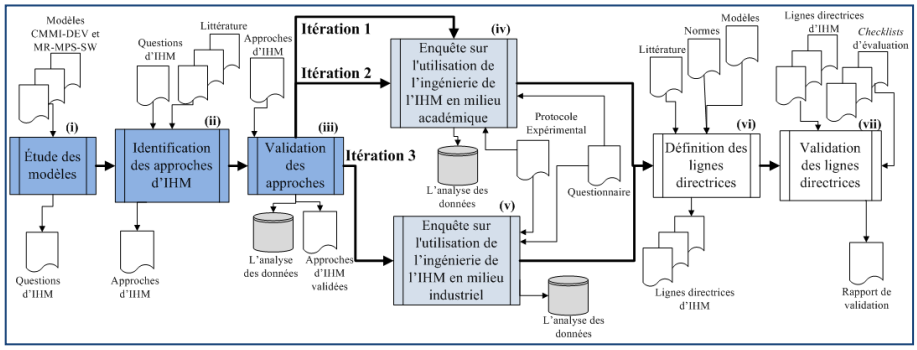


Figure 1. Méthodologie de Recherche

5. Actions futures

Nous travaillons actuellement sur la préparation de l'enquête qui sera effectuée en milieu industriel. Les actions futures concernent les **étapes (v), (vi) et (vii)**. En plus nous continuons l'enquête en milieu académique (**étape (iv)** - itération 2). La définition des lignes directrices sera basée sur des normes existantes pour le domaine de l'IHM et sur la littérature liée à l'ingénierie de l'IHM. Pour la validation des lignes directrices nous prévoyons une révision par les pairs (experts en GL et IHM).

6. Remerciements

Les auteurs tiennent à remercier CAPES - Programme Science sans frontières pour le soutien financier à ce travail.

7. Bibliographie

Bevan N., "International Standards for Usability Should Be More Widely Used", *Journal of Usability Studies*, 4 (3), 2009, p. 106-113.

CMMI Product Team. "CMMI® for Development (CMMI-DEV) V1.3", (CMU/SEI-2010th-TR-033 ed.). Pittsburgh, PA, USA: Carnegie Mellon University, 2010.

Dumas J. S., Fox J. E., "Usability testing: Current practice and future directions", *Human-Computer Interaction: Development Process*, A. Sears and J.A. Jacko (Eds.), CRC Press, 2009, p. 231-250.

Ferre X., Juristo N., Moreno A. M., "Which, when and how usability techniques and activities should be integrated," *Human-Centered Software Engineering - Integrating Usability in the Software Development Lifecycle, Human-Computer Interaction Series*, A. Seffah, J. Gulliksen, M. C. Desmarais (Eds.), Vol. 8. Springer Netherlands, 2005, p. 173-200.

Fischer H., Nebe K., Klompmaker F., "A Holistic Model for Integrating Usability Engineering and Software Engineering Enriched with Marketing Activities", *Proc. of International Conference on Human-Computer Interaction*, 2011, p. 28-37.

ISO/TR 18529, "Ergonomics of human-system interaction - Human-centred lifecycle process descriptions", 2000.

Jokela T., Siponen M., Hirasawa N., Earthy J., "A survey of usability capability maturity models: implications for practice and research", *Behaviour & Information Technology*, 25 (3), 2006, p. 263-282.

Limbourg Q., Vanderdonckt J., "Comparing Task Models for User Interface Design", *The Handbook of Task Analysis for Human-Computer Interaction*, D. Diaper and N. A. Stanton (Eds.), CRC Press, 2003, p. 135-153.

Raza A., Capretz L. F., Ahmed F., "An open source usability maturity model (OS-UMM)", *Computers in Human Behavior*, 28, 2012, p. 1109-1121.

Élasticité de l'exécution des processus métiers

Application à BonitaBPM

Guillaume Rosinosky

Bonitasoft / INRIA Nancy
32 rue Gustave Eiffel
38100 GRENOBLE

guillaume.rosinosky@bonitasoft.com

MOTS-CLÉS : *élasticité, BPM, cloud*

KEYWORDS: *elasticity, BPM, cloud*

ENCADREMENT. *François Charoy (LORIA), Samir Youcef (LORIA)*

1. Contexte

Pendant des années, les logiciels ont été installés sur des machines locales et gérées par des équipes IT ou par des utilisateurs finaux. L'arrivée du Cloud change ce paradigme en proposant des serveurs, plateformes ou services distants, payables à l'utilisation, par réservation ou à la demande. Ceci transfère la gestion de l'infrastructure à des fournisseurs, en permettant ainsi aux utilisateurs de payer seulement pour les services dont ils ont besoin.

Toutefois, pour pouvoir profiter totalement de ces avancées, des méthodes permettant le déploiement des logiciels de manière élastique sur les différents serveurs sont nécessaires. L'élasticité décrite ici correspond à l'adaptabilité d'une solution à pouvoir s'étendre avec le moins d'à-coups possibles lors de montées en charge (scalabilité), mais possédant également la capacité de se réduire quand la charge est moindre. Ce dernier point, moins important dans une architecture traditionnelle, est nécessaire pour limiter les coûts d'utilisation des ressources sous-traitées dans le Cloud.

Le BPM (Business Process Management) est le moyen de formaliser et de gérer les processus métier d'une entreprise. Un processus business est une séquence d'activités

ayant une valeur pour une entreprise. Ainsi, les outils BPM, en permettant la représentation sous forme de graphes des processus internes des entreprises, en permettant aux utilisateurs de les exécuter, et de régler eux-mêmes leurs spécificités, leur permet ainsi entre autres d'avoir une solution « à la carte ». Nous partons du principe qu'en s'appuyant sur la structure interne des schémas BPM et sur leur utilisation, nous sommes capables d'évaluer de manière efficace les besoins réels des clients du service, pour ensuite les déployer sur une infrastructure suffisante pour respecter leurs besoins en qualité de service, et ce, au meilleur coût.

Nous souhaitons proposer des approches d'élasticité pour le gestionnaire d'un BPMaaS (BPM as a Service). Nous partons du postulat que chaque instance de la solution BPM est distribuée sur une ou plusieurs instances matérielles et fait appel à une base de données séparée. La solution proposée doit également être multi-tenant, c'est à dire capable de gérer de manière isolée plusieurs clients sur une même installation.

2. État de l'art

Les capacités d'élasticité mises à disposition par les fournisseurs IaaS (Infrastructure as a Service) sont en général basées sur un système de seuils comme dans les Auto Scaling Groups d'Amazon Web Services (`http://aws.amazon.com/`) où des ressources sont ajoutées (ou retirées) en fonction de certaines métriques et des seuils correspondants. Le problème de cette approche est qu'elle fonctionne seulement de manière réactive, sur des critères liés à la configuration matérielle (RAM, CPU, etc.), et ne prend pas en compte les variations d'utilisation du logiciel.

Dans le domaine de l'élasticité pour les moteurs BPM, plusieurs approches existent telles que celles décrites dans le panorama de (Schulte *et al.*, 2014). Celles-ci ne considèrent en général que l'élasticité du moteur BPM sans tenir compte de l'aspect multi-tenant. (Sellami *et al.*, 2014) propose une approche prenant en compte ce dernier mais ignore les coûts de migrations et ne considère pas la partie base de données.

3. Problématique

Nous souhaitons proposer ici des méthodes prenant en compte simultanément : des configurations industrielles dans leur totalité (serveurs d'application et bases de données séparées), l'aspect multi-tenant, la qualité de service du client, et le coût des migrations.

Une solution simple pour assurer une bonne qualité de service pour les clients, serait de mettre à disposition les configurations les plus puissantes pour éviter les dégradations de qualité. Cette approche, souvent utilisée par les administrateurs système est séduisante pour le client, mais s'avère souvent très chère. En effet, le problème ici est que les critères coût-qualité sont souvent antagonistes. Les instances les plus puissantes sont souvent les plus chères. Un point notable est que chaque client d'un service va voir son utilisation changer avec le temps, et l'ensemble de la puissance

fournie par une configuration de taille maximale n'est pas nécessaire à chaque instant. En l'occurrence, plusieurs saisonnalités sont souvent constatées comme les horaires de bureaux, les variations jour-nuit, les week-end, etc. Une solution BPMaaS élastique doit être capable de fournir la puissance nécessaire au client au moment opportun, tout en réduisant l'empreinte financière pour le fournisseur.

4. Actions réalisées

Notre premier postulat est tout d'abord que les « migrations à chaud » des clients de l'outil sont possibles, et que celles-ci vont provoquer des ruptures de qualité de service comme l'indique (Das *et al.*, 2010) pour la migration des bases de données. Nous avons ainsi choisi comme premier axe l'optimisation du coût total des ressources *et* du nombre de migrations des différents tenants, sous contrainte de respect du débit de tâches BPM pour chacun d'entre eux. Ce débit, ainsi que la capacité et le prix des ressources doivent être préalablement établis. Nous considérons ici comme granularité temporelle l'heure, unité de facturation généralement utilisée par les fournisseurs de Cloud public. Un exemple de distribution de tenants et de ressources peut être observé en figure 1 .

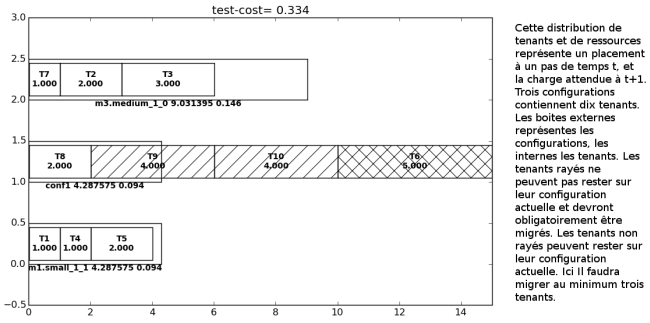


Figure 1. *Exemple de distribution de tenants sur différentes configurations.*

Nous avons proposé une première heuristique, ainsi qu'une modélisation sous la forme d'un problème d'optimisation linéaire. Il s'agit ici d'un problème bi-objectif où nous cherchons à minimiser simultanément le nombre de migrations et le coût engendré par la location de ressources. Nous avons choisi le critère de Pareto pour déterminer l'ensemble des solutions envisageables. Le principe général de l'heuristique proposée est d'observer pour chaque nombre de migrations de client les ressources, qui, en les cumulant, contiennent précisément ce nombre de tenants, et de les considérer comme « supprimées ». Il s'agit ici d'un problème de somme de sous ensemble où pour chaque nombre de migrations nous obtenons une liste de tenants « orphelins » et de ressources supprimées. Nous lançons ensuite pour chaque possibilité, une phase

de repacking où nous attribuons au mieux les tenants orphelins sur les configurations restantes de manière à minimiser l'espace disponible, suivie d'une étape de Variable Cost and Size Bin Packing basée sur les travaux de (Kang *et al.*, 2003).

Pour évaluer les performances de notre algorithme, nous avons comparé les performances et le temps de traitement de l'heuristique avec ceux d'un solveur sur le modèle décrit plus haut. La taille des configurations a été déterminée à l'aide de tests réalisés sur différentes configurations. Ces derniers ont été réalisés sur la solution BonitaBPM (`http://www.bonitasoft.com/`) (Charoy *et al.*, 2006) 7.0.3 Community avec le fournisseur Amazon Web Services. Les résultats obtenus montrent l'efficacité de notre approche. Des détails sur la méthode employée et les résultats obtenus sont retrouvables dans (Rosinosky *et al.*, 2016)

5. Actions futures

Une utilisation itérative de l'algorithme sur plusieurs pas de temps peut provoquer par exemple des migrations du même client à chaque heure, ce qui risque de poser des problèmes aux utilisateurs du service. Ainsi, nous travaillons actuellement sur un nouvel algorithme capable d'optimiser le coût et le nombre de migrations sur la base d'une journée, en prenant en compte un seuil maximum pour chaque tenant. En plus du nombre de tâches par client, d'autres critères pourront être également pris en compte, comme le nom du processus, le type de tâche (automatique ou manuelle) ou également la structure intrinsèque du schéma BPMN (BPM Notation) correspondant à chacun des processus.

6. Bibliographie

Charoy F., Guabtni A., Faura M. V., « A dynamic workflow management system for coordination of cooperative activities », *Business Process Management Workshops*, Springer, p. 205–216, 2006. 00024.

Das S., Nishimura S., Agrawal D., El Abbadi A., « Live database migration for elasticity in a multitenant database for cloud platforms », *CS, UCSB, Santa Barbara, CA, USA, Tech. Rep*, vol. 9, p. 2010, 2010.

Kang J., Park S., « Algorithms for the variable sized bin packing problem », *European Journal of Operational Research*, vol. 147, n° 2, p. 365–372, 2003. 00116.

Rosinosky G., Youcef S., Charoy F., « An Efficient Approach for Multitenant Elastic Business Processes Management in Cloud Computing environment », février, 2016, working paper or preprint.

Schulte S., Janiesch C., Venugopal S., Weber I., Hoenisch P., « Elastic Business Process Management : State of the art and open challenges for BPM in the cloud », *Future Generation Computer Systems*, 2014. 00011.

Sellami W., Kacem H. H., Kacem A. H., « Elastic Multi-tenant Business Process Based Service Pattern in Cloud Computing », IEEE, p. 154–161, décembre, 2014.

An approach towards strengthening consistency among multi-perspective business process models

Afef Awadid

Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac
75013 Paris
Afef.Awadid@malix.univ-paris1.fr

MOTS-CLÉS : *modèles de processus métier, perspectives de modélisation, cohérence inter-modèles, modélisation multi-perspective projective, modélisation multi-perspective sélective.*

KEYWORDS: *business process models, modeling perspectives, inter-model consistency, projective multi-perspective modeling, selective multi-perspective modeling.*

ENCADREMENT : *Selmin Nurcan (MCF) et Sonia Ayachi Ghannouchi (MCF).*

1. Context

In the field of process modeling, a business process (BP) defined as a set of one or more linked activities that collectively realize a business objective within the context of an organizational structure, is usually regarded as a complex system. The notion of multi-perspective modeling consisting in capturing the latter from various and complementary perspectives (such as functional, behavioral and informational perspectives) is then introduced in order to master its complexity. However, this notion implies the peculiar challenge of consistency among the produced models. The overall goal of our work is to propose an approach favoring at once the multi-perspective modeling of business processes and the consistency among the captured modeling perspectives, in order to reach a coherent understanding of each BP under study. Such understanding is vital for the reuse of the BP models as well as their improvement.

2. State-of-the-art

The need for inter-model consistency has been mostly felt since the early 1990s, when well-known enterprise modeling methodologies such as CIMOSA (Vernadat, 1992) and Zachman (Sowa et al., 1992) emerged and introduced the concept of multi-perspective modeling as a major source of inconsistency among the produced models. Nevertheless, the business process modeling area includes two further

sources of inconsistency; the variants of models depicting the same BP (using the same perspective and same formalism) and the merging of BP models. The former occurs for instance when models have been produced by different actors, or thorough the time by a same (or different) actor (s). Whereas, the latter arises for example when one seeks to obtain an overarching model from two fragments of BP models.

Based on the aforementioned sources of inconsistency, research works in this scope can be classified into three categories. A first category incorporates the approaches that deal with the consistency problems derived from multi-perspective modeling (Koubarakis et al., 2002), (Shunk et al., 2003), (Bork et al., 2014) and (Bork et al., 2015). A second category contains approaches with particular focus on inconsistency problems arising from the existence of multiple variants of the same BP (Hallerbach et al., 2010), (Pascalau et al., 2010) and (Weidlich et al., 2011), and a third one includes the approaches that shed light on the inconsistency problems resulting from the merging of BP models (Smirnov et al., 2010), (Gerth et al., 2010) and (Zemni et al., 2014).

Four major characteristics may effectively distinguish the first category from the two others: (1) the use of multiple modeling formalisms rather than a single one, (2) the subject of modeling which may refer to the enterprise as a whole instead of the BP only, (3) the scope of the approach which consists in a modeling method rather than a modeling formalism and (4) the type of multi-perspective modeling concerning only the first category and consisting in either projective multi-perspective or selective multi-perspective modeling. The former appears when one comprehensive overarching meta-model is given. All perspectives captured by all concerned modeling formalisms are defined as projections onto this central meta-model. One example of modeling standards adopting this type of multi-perspective modeling is UML which has, in its current version, the Meta Object Facility (MOF) as a common meta-model. All UML diagram types are specified by projections onto that MOF meta-model. However the latter appears when no central meta-model is given. Each perspective is captured by a distinct meta-model (Cicchetti et al., 2012).

3. Problematic

Although inter-model consistency issue has been widely tackled in the field of business process modeling, approaches dealing with this issue, focusing on a BP as the main subject of modeling and on the multi-perspective modeling as source of inconsistency are still very scarce. Furthermore, these few approaches put emphasis on the projective multi-perspective modeling, and hence on a particular modeling method or standard which limits their applicability outside of the investigated cases. To overcome these deficiencies, we intend to propose an approach to strengthening consistency among multi-perspective BP models.

This approach will be likely to respond to the following questions: How to enhance consistency among the multi-perspective BP models under study? How to guarantee the synchronization between these BP models when one of the business

processes undergoes a change in order to ensure a consistent state of the business processes ecosystem? And how to provide a generalizable solution to the issue of consistency among multi-perspective BP models (i.e. regardless of the modeling formalism being used)?

4. Performed actions

Our attempt to answer the two first questions raised in the above section consists first in granting a special attention to the Enterprise Knowledge Development (EKD) method (Loucopoulos et al., 1997). In fact, EKD (1) advocates the multi-perspective modeling of complex business processes, as a valuable way for their better understanding, by offering a set of modeling formalisms, namely actor-role, role-activity and business objects formalisms, and (2) approaches and tools supporting the building and the evaluation of EKD BP models mastering the consistency issues are still missing.

In order to enhance the syntactic and semantic quality of EKD BP models, we have proposed a set of consistency rules controlling their dependencies. Semantic consistency rules are not identified that easy. Instead, one needs to have in depth knowledge of the semantics of the various perspectives. Thereby, as their definition is far from obvious, we have modeled and investigated a plethora of business processes (e.g. book lending, freight shipping and travel offer management processes) referring to different domains. A first validation of the proposed rules have been performed. Besides, our proposal is supported by a tool in order to promote an automated consistency handling, and hence to save the time and efforts of process modelers. However, the tool in its current state supports only the syntactic consistency rules, which were implemented at meta-modeling level, in order to allow modelers to automatically obtain BP models that are syntactically consistent.

5. Future actions

The actions, which have been performed, allow us to contemplate multiple research perspectives including preoccupations to be fulfilled in the near future, which lie mainly in (1) further validating the defined semantic consistency rules by referring to other independent modelers, (2) guiding modelers to comply with the semantic consistency rules. This allows them to cope with the difficulty in understanding the poor feedback that the consistency check may produce, since it is usually expressed within the scope of the technique chosen (for instance the formal technique used), and (3) enabling a synchronized state between the produced models when one of them undergoes a change.

Moreover, we intend to investigate the possibility of applying the defined consistency rules to other business process modeling methods or standards such as BPMN, UML and ARIS, by referring on one side to the notion of selective multi-perspective modeling and on the other side to the extensibility of meta-models, as an

initial attempt to answer the last question posed in the above section 3. This idea embodies the cornerstone of our envisaged approach. Another future perspective consists in ensuring that the proposed approach will be likely to master the inter-model consistency in a scalable BPs ecosystem.

6. References

Vernadat F., CIMOSA-A European development for enterprise integration, Part 2: Enterprise Modelling, *Enterprise Integration Modeling, The MIT Press, Cambridge,* 1992.

Sowa J. F., Zachman J. A.,"Extending and formalizing the framework for information systems architecture", *IBM systems journal*, vol.*31*, n° 3, 1992, p. 590-616.

Koubarakis M., Plexousakis D., "A formal framework for business process modelling and design", *Information Systems*, vol.27*,* n° 5, 2002, p. 299-319.

Shunk D. L., Kim J. I., Nam H. Y., "The application of an integrated enterprise modeling methodology—FIDO—to supply chain integration modeling", *Computers & industrial engineering*, vol. *45*, n°1, 2003, p. 167-193.

Bork D., Karagiannis D., "Model-driven development of multi-view modeling tools the MuVieMOT approach". In *ICSOFT-PT,* IEEE*, Vienna, Austria, 2014*, p. IS-11.

Bork D., Buchman R., Karagiannis D., "Preserving Multi-view Consistency in Diagrammatic Knowledge Representation", In *Knowledge Science, Chongking, China, 2015,* p. 177-182.

Hallerbach A., Bauer T., Reichert, M., "Capturing variability in business process models: the Provop approach", *Journal of Software Maintenance and Evolution: Research and Practice*, vol. *22*, n° 6-7, 2010, p. 519-546.

Pascalau E., Awad A., Sakr S., Weske M., "On maintaining consistency of process model variants". In *Business Process Management Workshops, Springer, 2010,* p. 289-300.

Weidlich M., Mendling J., Weske M., "Efficient consistency measurement based on behavioral profiles of process models", *IEEE Trans,on Software Engineering*, vol.*37*, n° 3, 2011,p.410-429.

Smirnov S., Weidlich M., Mendling J.,"Business process model abstraction based on behavioral profiles", In *Service-Oriented Computing,* Springer, 2010, p. 1-16.

Gerth C., Luckey M., Kuster J. M., Engels G.,"Detection of semantically equivalent fragments for business process model change management", In *IEEE International Conference on Services Computing (SCC),Florida, USA, 2010*,p. 57-64.

Zemni M.A., Mammar A., Ben Hadj Alouane N., "A Behavior-Aware Systematic Approach for Merging Business Process Fragments",In *19th International Conference onEngineering of Complex Computer Systems (ICECCS), IEEE, Tianjin, China, 2014*, p. 194-197.

Cicchetti A., Ciccozzi F., Leveque T.,"A hybrid approach for multi-view modeling" ,*Electronic Communications of the EASST*, Berlin, Germany, 2012.

Loucopoulos P., Kavakli V., Prekas N, Rolland C, Grosz G., Nurcan S., "Using the EKD approach: the modelling component", ELECTRA Athena (projet ESPRIT IV), CRI, University of Paris 1 Pantheon-Sorbonne, July 1997.

Stream and Resource-Aware Elastic Stream Processing

Roland Kotto Kombi

INSA de Lyon - LIRIS
20 Avenue Albert Einstein
69100 Villeurbanne

roland.kotto-kombi@liris.cnrs.fr

MOTS-CLÉS : *flux de données, traitement distribué et parallèle, requête continue, optimisation dynamique, allocation d'opérateurs*

KEYWORDS: *data streams, distributed, parallel execution, continuous query, online optimization, operator scheduling*

ENCADREMENT. *Philippe Lamarre (PR) et Nicolas Lumineau (MCF)*

1. Context

At the era of Big Data, more and more data sources emit real-time information which require to be processed with acceptable end-to-end latency. Executing endless queries, called *continuous queries*, on data streams represents a challenge in terms of performance, scalability and robustness, in particular for complex queries (*e.g.* aggregative queries). Unlike disk-based data, data streams are potentially infinite and unpredictable sequences of items in terms of input rate and value distribution. Some *Data Stream Management Systems* (DSMS) have been developed on the last decade (Schneider *et al.*, 2009; Peng *et al.*, 2015; Xu *et al.*, 2016). They are able to execute continuous queries on a set of data streams in a distributed and parallel fashion. It allows to fit treatments to available resources.

In the context of the ANR project Socioplug[1], we suggest a platform of individual plugs for structured data stream processing. Plugs are grouped into subclusters, called communities. A community gathers plugs executing equivalent queries in order to compute each query result only once and in a parallel fashion. So, we aim at executing a variable number of continuous queries on a cluster of individual plugs with limited resources. It appears that the key feature of our approach is the ability to dynamically scale-in and scale-out and execute queries only on necessary resources. Actually, each continuous query may process greater volumes of data (unpredictability of data streams). Finally, new queries may be submitted to the system, increasing the global workload.

2. Related works

In our context, we consider that a stream is described by a schema and a potentially infinite and ordered set of timestamp. This schema is composed by attributes which describe each element. A single timestamp is associated to an element but many elements may share the same timestamp. For example, a data stream from Twitter can be composed of tuples with a schema S :<author,topic,message,retweets> and their timestamp would be the date of publication. Operators can process tuples in a chronological order thanks to their timestamps. More, many operators require to compute a result on a set of elements (*e.g.* aggregates or joins). This set of elements is defined by an interval of timestamp to consider and is denoted *computation window*.

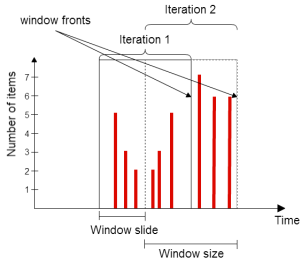


Figure 1 – Computation window

On those streams, users can define continuous queries which are represented as dataflow diagrams, called *workflows*. A workflow is a direct acyclic graph where vertices are operators and edges are data streams between operators. A topology on Twitter stream can be composed of an operator filtering tweets about a given topic and an other operator returning only tweets with more than a certain number of retweets. The main challenge is then to adapt workflow execution in order to scale-in or scale-out in order to handle to input stream variations. This adaptation, called *elastic stream processing*, aims at maintaining an acceptable end-to-end latency regardless of

1. `http://socioplug.univ-nantes.fr/index.php/SocioPlug_Project`

input stream variations. We distinguish two major steps in elastic stream processing : operator parallelization and operator scheduling. On one hand, many works (Schneider *et al.*, 2009; Xu *et al.*, 2016) looked after operator parallelization. Those solutions adapt dynamically the number of threads executing an operator in order to guarantee some properties on performance or global state of the execution support. For example, properties can be the total latency of the workflow or the average CPU load of the execution support. On the other hand, some scheduling algorithms (Peng *et al.*, 2015) have been suggested. They consider metrics which tend to optimize resource utilization (CPU, network traffic) at the scale of the cluster.

3. Issues

Given our context, we consider two complementary points of view. First, the system must deliver results to users with an acceptable end-to-end latency. More, we do not consider that users have information about the execution support and the global workload induced by other queries. So, solutions based on user expertise do not fit to this context. Then, the system can not predict how many resources are available for a specific query at any time. An acceptable tradeoff is then to use only necessary resources for a given query in order to respect latency constraint. Finding the optimal configuration for a workflow means solving two sub-problems. First, the definition , for each operator, of the appropriate number of threads in order to optimize the global latency. As presented in (Xu *et al.*, 2016), it is important to identify only relevant operators to fork. But, in opposition to existing approaches, we need to find an algorithm which avoid to test each degree of parallelism. More, users can not give indications on how many machines to allocate. Then, we have to identify the minimal subset of operator to reallocate in order to minimize reconfiguration overheads but guaranties an improvement of global performance. To achieve this goal, we need to take into account material and network constraints. In this context, the problem is then to define online mechanisms to dynamically scale in and scale out treatments without user guidance and a scheduling algorithm which minimize resource consumption.

4. Current works

First, we produced a survey (Kotto-Kombi *et al.*, 2015) which suggest a classification of representative DSMS. Then, we have deepened the two main aspects for online optimization of continuous queries. We consider that the parallelization of stateless operators can be defined with more accuracy after a learning phase. Actually, after a certain time, an average execution time can be computed. This average time allows to determine if the incoming stream rate will congest the operator. So, we can approximate how many forks are necessary to avoid a congestion. This approach is similar to what is suggested in (Xu *et al.*, 2016) but we do not consider that the parallelization is asked by users but is triggered automatically when the system detects a potential congestion. Then, we explicitly ask that the system releases an under-utilized machine after n

consecutive observations in order to automatically scale-in treatments. Our scheduling approach remains close from the online resource-aware scheduler presented in (Peng *et al.*, 2015). Nevertheless, we replace user specifications by monitoring information. In our context, users have no control or feedback on the execution support so it is not relevant to ask them hints on how to optimize their queries. We implemented those modules into Apache Storm[2] because of its growing popularity and development. In a first time, we executed some tests to evaluate performance and deployment effort. It appears that Storm delivers best performance and offers great tuning opportunities. Then we implemented a first version of an online resource-aware scheduler over Storm. Preliminary results shows that it can outperforms Storm default's scheduler and use less resources. We are currently evaluating different parallelization approaches to determine the most adapted one.

5. Future Works

In parallel of our current works, we plan to define a parallelization algorithm dedicated to join operations. Actually, join operations are particularly frequent while querying structured data. They have the particularity to potentially emit more data than they receive. Some works (Vengerov *et al.*, 2015) suggest techniques and structures to estimate join result sizes from their inputs. It brings the opportunity to dynamically set the appropriate number of instances of a join operator in a pro-active manner. We also plan to find an heuristic which generalize the online resource-aware scheduler developed in (Peng *et al.*, 2015). Actually, we want to reschedule only critical operators. That means that we aim at automatically placing operators that have a large impact on global performance on machines that will decrease significantly the end-to-end latency.

6. Bibliographie

Kotto-Kombi R., Lumineau N., Lamarre P., Caniou Y., « Parallel and Distributed Stream Processing : Systems Classification and Specific Issues », octobre, 2015, working paper or preprint.

Peng B., Hosseini M., Hong Z., Farivar R., Campbell R. H., « R-Storm : Resource-Aware Scheduling in Storm », *Proceedings of the 16th Annual Middleware Conference, Vancouver, BC, Canada, December 07 - 11, 2015*, p. 149–161, 2015.

Schneider S., Andrade H., Gedik B., Biem A., Wu K.-L., « Elastic scaling of data parallel operators in stream processing », *Parallel Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*, p. 1-12, May, 2009.

Vengerov D., Menck A. C., Zait M., Chakkappen S. P., « Join Size Estimation Subject to Filter Conditions », *Proc. VLDB Endow.*, vol. 8, n° 12, p. 1530–1541, août, 2015.

Xu, Peng G., « Stela : Enabling Stream Processing Systems to Scale-in and Scale-out On-demand », *Proc. IEEE International Conference on Cloud Engineering (IC2E), 2016*, 2016.

2. `http://storm.apache.org/`

Définition d'un modèle de système de Bases de Données Temps Réel pour les Systèmes d'Aide à la Conduite Coopératifs

Islam Elleuch

Faculté des Sciences Economiques et de Gestion de Sfax, Laboratoire MIR@CL
Route de l'Aéroport km 4, B.P. 1088, 3018 Sfax, Tunisie

elleuchislam@gmail.com

MOTS-CLÉS : *Systèmes d'aide à la conduite, VANET, Bases de données temps réel, Gestion de données.*

KEYWORDS: *Advanced driver assistance systems, VANET, Real time database, Data Management.*

ENCADREMENT : *Rafik Bouaziz (PR), avec le co-encadrement de Achraf Makni (MA)*

1. Contexte

Le nombre de véhicules circulant sur la route ne cesse d'augmenter, ce qui accroit le nombre d'accidents et les problèmes de sécurité routière. Pour pallier ces problèmes, les véhicules et l'infrastructure du réseau routier sont dotés de systèmes d'aide à la conduite (Advanced Driver Assistance Systems: ADAS) afin d'assister le conducteur dans sa tâche de conduite. Initialement, les systèmes ADAS sont autonomes. Ils se concentrent sur le comportement du véhicule à lui seul. Ainsi, le conducteur du véhicule effectue la perception de l'environnement, la décision et l'action. Il est basé uniquement sur ses propres informations et sa propre initiative, sans interagir avec d'autres véhicules. Afin d'améliorer la sécurité routière, des systèmes d'aide à la conduite coopératifs ont été proposés. Ces systèmes permettent d'échanger les données entre les véhicules ou entre les véhicules et l'infrastructure. Les véhicules en coopération peuvent partager et fusionner leur perception. Ils peuvent aussi partager leur intention de mouvement afin de coordonner leurs actions. Ainsi, un véhicule peut utiliser les données d'autres véhicules pour préciser sa localisation et pour détecter aussi la cartographie et les objets en mouvement. Nous nous intéressons dans notre thèse à la problématique de la gestion des données dans le cadre des ADAS coopératifs.

2. État de l'art

Les services fournis par les systèmes ADAS coopératifs sont basés essentiellement sur l'échange d'informations entre véhicules. Nous parlons dans ce cas d'un réseau ad hoc de véhicules (Vehicular Ad hoc Network ; VANET). Un

réseau VANET est une particularité des réseaux *Mobile Ad hoc Network* (MANET) où les nœuds mobiles sont des véhicules intelligents avec des spécificités supplémentaires (Dahiya *et al.*, 2010). Le réseau VANET est constitué de véhicules capables de s'échanger des informations et de communiquer au sein d'un groupe de véhicules à portée les uns des autres (Vehicle To Vehicle ; V2V) et avec les équipements fixes de la route à portée (Vehicle To Infrastructure ; V2I) pour fournir des informations opportunes aux conducteurs et aux autorités intéressées (Elias *et al.*, 2016). Bien que le volume de données manipulé soit de plus en plus important, les systèmes ADAS coopératifs actuels utilisent peu les Bases de Données (BD). Ils sont généralement basés sur des calculs récursifs ou probabilistes pour réaliser une action coopérative. De ce fait, la plupart des travaux de recherches se sont concentrés sur les technologies de communication et sur les approches de diffusion des données. Ainsi, plusieurs technologies de communication ont été proposées dans les communications inter-véhicules (V2V et V2I). Les auteurs de (Papadimitratos *et al*., 2009) ont comparé les technologies 802.11p WAVE (Wireless Access for Vehicular Environment), Wi-Fi, le système cellulaire et l'infrarouge, et montrent la distinction de WAVE. Egalement, plusieurs approches de diffusion de données dans le cadre de VANET ont été proposées (Ilarri *et al*., 2015). Ces approches devraient tenter d'optimiser des métriques relatives à la surcharge du réseau, à la redondance, aux taux d'utilisation, à la latence, etc. Dans (Xu *et al*., 2004), les auteurs ont proposé un mécanisme d'échange opportuniste, inspiré de l'épidémiologie, pour la dissémination d'informations entre des véhicules. Un véhicule en possession d'une information peut ainsi en contaminer d'autres rencontrés sur son trajet. Cependant, la pertinence de l'information devient faible avec le temps et il serait utile de la faire disparaître. Pour ce faire, (i) plusieurs objets détruisent l'information et (ii) peu d'objets ne la prennent dorénavant. L'approche de diffusion basée sur le contenu présentée dans (Cenerario *et al*., 2011) adapte dynamiquement la zone de diffusion en fonction des besoins en tenant compte de la pertinence des événements pour les véhicules. L'avantage majeur de ce système est sa capacité à minimiser le nombre de transmissions redondantes, et de réduire ainsi la surcharge du réseau. Cependant, la latence dans ce système est élevée, car il introduit des temps d'attente à chaque saut du processus de diffusion.

3. Problématique

La coopération inter-véhicules est intéressante et permet aux systèmes d'aide à la conduite d'évoluer rapidement. Cependant, ces systèmes sont complexes. Ils comprennent de plus en plus de capteurs et de fonctionnalités, et traitent un grand volume de données. Ces dernières doivent être mises à jour régulièrement pour refléter l'état courant de l'environnement. Malgré cela, les études ont montré que les systèmes existants ne gèrent pas le stockage des données. Par conséquent, l'échange de données devient très couteux en temps, et le risque que les transactions ratent leurs échéances augmente. Ces systèmes sont confrontés à des problèmes relatifs à la manipulation de grandes quantités de données, ce qui peut engendrer des données

perdues et des conséquences graves. Les données échangées entre les véhicules sont considérées comme des éléments à transmettre : une fois utilisées, elles sont considérées obsolètes et détruites.

Partant de ce constat, l'objectif de notre sujet de recherche est l'amélioration de la performance des ADAS coopératifs en utilisant les Bases de Données Temps Réel (BDTR). Le système à proposer doit permettre de manipuler efficacement un grand nombre de données et garantir également les contraintes temporelles relatives aux données et aux transactions. Nous cherchons à améliorer les résultats liés à l'estimation de l'état du véhicule, à la localisation coopérative et à la détection coopérative des objets en mouvement.

Les applications qui manipulent de grands volumes de données gagnent à être exploitées sous des Systèmes de Gestion de Bases de Données Temps Réel (SGBDTR). Ces applications nécessitent, en premier lieu, une modélisation des schémas des données temps réel et, en deuxième lieu, des mécanismes appropriés pour le respect des contraintes temporelles à travers la résolution des problèmes d'ordonnancement des transactions et de contrôle de concurrence. Le but final sera le maintien, l'utilisation et la mise à jour de l'état de l'environnement pour tous les véhicules.

4. Actions réalisées

L'étude de l'état de l'art, concernant (i) les systèmes d'aide à la conduite, (ii) les métriques et les approches de diffusion de données, (iii) les protocoles dans le cadre de VANET et (iv) le respect des contraintes de temps sur les données et les transactions dans les systèmes de BDTR, nous a permis d'affiner la définition de la problématique à traiter et à aborder la recherche de solutions. Nous avons commencé par proposer un modèle de système de BDTR permettant de maintenir et partager l'état courant des véhicules.

A cet effet, nous avons défini la structure des données par des triplets $d = (d_{valeur}, d_{timestamp}, d_{durée})$ où d_{valeur} représente la valeur réelle de la donnée, $d_{timestamp}$ est l'instant où cette valeur est mise à jour et $d_{durée}$ désigne la durée de validité d'une donnée. Dans les systèmes de BDTR, ces éléments sont nécessaires pour vérifier qu'une donnée n'est pas obsolète ($d_{timestamp} + d_{durée} <$ instant courant). Nous avons aussi modélisé chaque véhicule par un nœud caractérisé par sa position (X, Y), sa vitesse ($V_{x,y}$) et son sens de circulation. Cette modélisation permet de présenter l'information liée à l'aspect coopération entre véhicules. Ainsi, au niveau de chaque nœud, un état de l'environnement proche est maintenu. En plus, et comme les véhicules sont en forte mobilité (ils peuvent rapidement rejoindre ou quitter le réseau en un temps très court), des nœuds peuvent être ajoutés et d'autres supprimés de la BDTR concernée. Ceci permet de réduire la quantité des données stockées. Le réseau des nœuds aide aussi à prévoir les cas d'accidents et d'éviter les collisions même si deux véhicules ne se voient pas. Ceci est assuré à travers l'échange de données et le maintien de l'état de l'environnement proche par chaque nœud. Par

ailleurs, le maintien de l'état courant et la définition des cas d'accidents permettent de réduire la surcharge du réseau. En effet, chaque nœud peut choisir le nœud cible ainsi que l'instant approprié pour l'envoi d'information, contrairement aux systèmes existants où les données sont diffusées à tous les autres nœuds.

5. Actions futures

Nous allons explorer et proposer des voies de solution aux autres problèmes posés, comme (i) l'estimation du degré d'importance d'une donnée, (ii) la définition de transactions qui respectent au mieux les échéances Temps Réel (TR) et la validité de données, (iii) l'ordonnancement des transactions et (iv) le contrôle de concurrence. Afin de réduire la complexité liée à la conception des ADAS coopératifs reposant sur l'utilisation des systèmes de BDTR, nous avons également à étudier l'apport des patrons de conception spécifiques à ces systèmes (Marouane *et al*., 2012). Nous devons aussi évaluer les avantages conséquents de l'intégration d'un système de BDTR dans les ADAS coopératifs à l'aide du simulateur NS2. Enfin, nous comptons positionner nos contributions par rapport aux systèmes existants.

6. Bibliographie

Cenerario N., Delot T., and Ilarri S., "A content-based dissemination protocol for VANETs: Exploiting the encounter probability ",*IEEE Transactions on Intelligent Transportation Systemss*, vol. 12, n° 3, 2011, p. 771-782.

Dahiya A., Chauhan R., "A comparative study of manet and vanet environment ", *Journal of computings*, vol. 2, n° 7, 2010, p. 87-92.

Elias C., Si-Jing Z., En-Jie L., Joy E., "Advances in Vehicular Ad-hoc Networks (VANETs): Challenges and Road-map for Future Development ",*International Journal of Automation and Computinge*, vol. 13, n° 1, 2016, p. 1-18.

Ilarri S., Delot T., Trillo R., "A Data Management Perspective on Vehicular Network ", *IEEE IEEE Communications Surveys & Tutorials*, vol. 17, n° 4, 2015, p. 2420 - 2460.

Marouane H., Makni A., Bouaziz R., Duvallet C., Sadeg B., " A Real-Time Design Pattern for Advanced Driver Assistance Systems ", *Proceedings of the 17th Conference on Pattern Languages of Programs EuroPLoP'2012* , Irsee Monastery, 11-15 Juillet 2012, Germany, p. C6 :1-C6 :11.

Papadimitratos P., de La Fortelle A., Evenssen K., Brignolo R., Cosenza S., " Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation ", *IEEE Communications Magazine*, vol. 47, n° 11, 2009, p. 84-95.

Xu B., Ouksel A. M., Wolfson O., " Opportunistic resource exchange in inter-vehicle ad-hoc networks", *Fifth International Conference on Mobile Data Management MDM'04*, Berkeley, 19-22 janvier 2004, USA, IEEE Computer Society, p. 4-12.

Vers une Suite Décisionnelle dédiée aux Données de Tests

Lahcène BRAHIMI

Laboratoire d'Informatique et d'Automatique pour les Systèmes LIAS/ISAE-ENSMA
1 Avenue Clément Ader, 86360 Chasseneuil France

lahcene.brahimi@ensma.fr

MOTS-CLÉS : *Base de Données, SGBD, Entrepots de Données, Banc d'essai, Apprentissage automatique.*

KEYWORDS: *Database, DBMS, Data Warehouse, Benchmarks, Machine Learning.*

ENCADREMENT. *Ladjel Bellatreche (PR)*

1. Contexte

Actuellement, nous vivons dans l'ère de la diversité qui a touché le monde des bases de données. Il y a quelques années, il y avait une dizaine de Systèmes de Gestion de Base de Données (SGBD) que les entreprises et les organismes d'enseignement se partageaient pour développer des applications liées aux bases de données. Aujourd'hui, nous assistons à une prolifération de nouveaux SGBD développés pour répondre aux besoins d'analyse motivés par l'ère des Big Data. Le site de BD-engines[1] offre une classification de **303** SGBD selon des métriques liées à l'usage et les retours des utilisateurs. Cependant, nous avons identifié l'absence d'outils quantitatifs dédiés aux tests de performance des SGBD. La communauté des bases de données depuis des années, effectue des tests des solutions proposées. Ces dernières couvrent la performance des requêtes, le choix des plateformes de déploiement, le choix des données, etc. Nous trouvons ces tests soit dans les articles de recherche soit dans les sites Web d'organismes comme le TCP (Transaction Processing Council)[2]. Récemment,

1. `http://db-engines.com/en/`
2. `http://www.tpc.org`

un nombre important de chercheurs comme Jens Dittrick[3], a montré la difficulté auprès des chercheurs de reproduire des résultats de certains travaux[4], ce qui remet en cause leur transparence et leur fiabilité.

Dans cet article, nous expliquons tout d'abord l'importance et la transparence des tests dans le domaine des bases de données. Puis, nous proposons la conception d'un référentiel dédié à stocker l'environnement et les résultats des tests.

2. État de l'art

Les différentes solutions des bases de données doivent satisfaire des besoins non fonctionnels comme la performance. Ce besoin est souvent lié à la phase physique de cycle de vie de conception des bases de données. Durant cette phase, un nombre de structures d'optimisation est sélectionné pour réduire le temps de réponse d'une charge de requêtes. Cette sélection est souvent contrainte par l'espace de stockage, le coût de maintenance des structures sélectionnées, etc. (Khouri, 2013). L'évaluation des algorithmes répondant aux besoins non fonctionnels et aux contraintes nécessite une batterie de tests sur un ou plusieurs environnements. Ces tests nécessitent un schéma d'une base de données, des instances, une plateforme de déploiement, un SGBD, etc. Souvent, les informations sur les données sont récupérées à partir des bancs d'essais Linear Road (pour les systèmes de gestion de flux de données), les TPC, YCSB (cloud), BerlinMOD (Düntgen *et al.*, 2009) (les systèmes spatiaux temporels) and GenBASE (domaine de la biologie et de santé) (Taft *et al.*, 2014).

Par ailleurs, le TPC BenchmarkH (TPC-H)[5], largement utilisé, est un test de performance des solutions d'entreposage de données. Ce banc d'essai a été conçu pour les systèmes décisionnels. Il mesure à la fois la vitesse et la capacité des solutions de bases de données. Il représente un schéma de huit tables relationnelles et une charge de travail de *22 requêtes* en lecture seule (Poess *et al.*, 2000).

3. Problématique

Actuellement, tester une application de base de données représente un enjeu important. Cela est dû à la complexité et la multiplicité des phases de son cycle de vie. Ce dernier comporte les phases suivantes (Khouri, 2013) : la collecte des besoins, la modélisation conceptuelle, la modélisation logique, la phase de déploiement, la phase physique et la phase d'exploitation. Il est à noter que chaque phase a ses propres besoins fonctionnels et non-fonctionnels à évaluer vu le déluge des données.

Les tests dans le cadre du développement d'applications de base de données peut avoir trois principales formes : (i) *tester une base de données en tant que produit final*,

3. `https://infosys.uni-saarland.de/people/dittrich.php`
4. `https://www.youtube.com/watch?v=07Qgo6RSzmE,àpartirdela20ièmeminutesdecettevidéo`
5. http ://www.tpc.org/tpch/

(ii) *tester l'interaction entre les applications et leurs bases de données* et (iii) *tester toutes les phases du cycle de vie de la conception de base de données*.

La littérature en rapport avec les activités de tests de la troisième catégorie tente de répondre à la question suivante : comment sont-ils testés ?

Deux méthodes de test existent pour répondre à cette question : *La simulation* et *l'expérimentation matérielle*. Pour la simulation, les chercheurs utilisent principalement des modèles de coûts mathématiques et des méthodes formelles avec des paramètres d'entrée qui sont liés à plusieurs composantes des bases de données telles que le schéma, la plateforme, les requêtes, les SGBD, les structures d'optimisation (par exemple index), etc. (van der Veen *et al.*, 2012). L'expérimentation matérielle est la méthode la plus privilégiée pour les grandes entreprises telles que GAFA (Google, Apple, Facebook et Amazon) qui possèdent des ensembles de données et plateformes avancées. Comme cité dans (Haftmann *et al.*, 2007), Microsoft consacre 50% de ses coûts de développement sur les tests. Un autre exemple, le cycle de sortie des produits SAP fait 18 mois dont six mois sont utilisés uniquement pour exécuter des tests. La question que nous nous posons dans cette thèse est la suivante : *comment exploiter les données de tests et les rendre plus accessibles et surtout plus transparentes ?* L'activité de tests représentant un coût important (argent et énergie), nous proposons dans ce papier un référentiel de test qui joue le rôle d'un entrepôt de données contenant toutes les activités des tests : les ensembles de données, les métriques utilisées, la plateforme, la charge de travail, les algorithmes, etc. Ce référentiel permet ainsi de stocker tout l'environnement des tests et de simuler des résultats des tests exprimés par l'utilisateur.

4. Actions réalisées

Nous avons conçu un référentiel de test DW_TESTS dont le but est de stocker tout l'environnement des tests avec les résultats obtenus (temps de réponses des requêtes, énergie consommée, etc.). Le but est de prédire des résultats de tests en utilisant des environnements exprimés par l'utilisateur ou par des experts. La conception d'un entrepôt de données nécessite l'identification de ses dimensions. Dans notre contexte, ces dimensions comprennent : *Dim_Platform*, *Dim_Deployment*, *Dim_DBMS*, *Dim_OS*, *Dim_Dataset*, *Dim_Query*, *Dim_AccessMethods*, *Dim_Algorithms*, *Dim_Hypothesis*, *Dim_Metrics* et deux dimensions d'information : *Dim_Laboratory* and *Dim_Time*. La table de faits couvre différentes mesures utilisées par les concepteurs telles que le coût CPU, le coût IO et le coût de l'énergie. Notre entrepôt peut facilement être modélisé en utilisant un schéma en étoile, comme le montre la figure 1.

5. Actions futures

Actuellement, nous sommes en train de développer une suite décisionnelle contenant notre entrepôt de tests, sur lequel des opérations de type OLAP et reporting sont définies. Cette suite pourrait compléter l'outil *BD-engines* (cf. Contexte). Sur

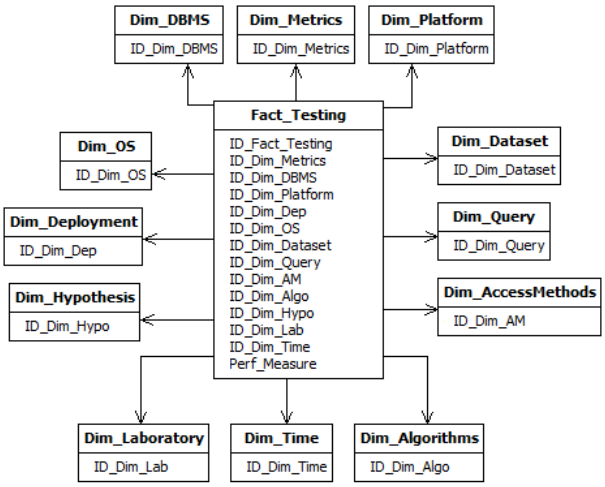


Figure 1. *Entrepôt de données DW_TESTS*

cette suite, un système de recommandation est en cours de finalisation. Il conseille les concepteurs/développeurs sur certains choix stratégiques liés aux SGBD, plateformes, les données, les requêtes, etc. Ce système utilise intensivement des techniques d'apprentissage automatique, tels que la régression sur les données de l'entrepôt.

6. Bibliographie

Düntgen C., Behr T., Güting R. H., « BerlinMOD : a benchmark for moving object databases », *The VLDB Journal*, vol. 18, n° 6, p. 1335–1368, 2009.

Haftmann F., Kossmann D., Lo E., « A framework for efficient regression tests on database applications », *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 16, n° 1, p. 145–164, 2007.

Khouri S., Cycle de vie sémantique de conception de systèmes de stockage et de manipulation de données, PhD thesis, ISAE-ENSMA Ecole Nationale Supérieure de Mécanique et d'Aérotechique-Poitiers, 2013.

Poess M., Floyd C., « New TPC benchmarks for decision support and web commerce », *ACM Sigmod Record*, vol. 29, n° 4, p. 64–71, 2000.

Taft R., Vartak M., Satish N. R., Sundaram N., Madden S., Stonebraker M., « Genbase : A complex analytics genomics benchmark », *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, ACM, p. 177–188, 2014.

van der Veen J. S., van der Waaij B., Meijer R. J., « Sensor data storage performance : Sql or nosql, physical or virtual », *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, IEEE, p. 431–438, 2012.

Ingénierie des processus ETL pour les Big Data

Hana Mallek
Université de Sfax
MIRACL (Multimedia, Information Systems and Advanced Computing Laboratory)
Le pôle technologique de Sfax, Sakiet Ezzit SFAX-Tunisie
BP 242-3021
mallekhana@gmail.com

MOTS-CLÉS : *Données Massives, Les entrepôts de données, Extraction-Transformation-Chargement (ETC), Les systèmes décisionnels, Modélisation des processus, Qualité de données.*

KEYWORDS: *Big Data, Data warehouse, Extraction-Transformation-Loading (ETL), Business Intelligence, Process Modeling, Data quality.*

ENCADREMENT : *Lotfi Bouzguenda (MCF-DR) et Faiza Ghozzi (MA).*

1. Contexte

Dans le cadre de l'analyse décisionnelle, les managers sont face à l'apparition des données massives, hétérogènes et évolutives. Ce phénomène est le résultat de la croissance exponentielle de nouvelles sources de données telles que le Web, les réseaux sociaux, les applications ubiquitaires (Smartphones, GPS,…), etc. Dans ce contexte, la quantité gigantesque de données donne naissance à la notion de "Big Data". Cette dernière est souvent définie par les 4 V : Volume, Vélocité, Variété et Véracité (Dong et Srivastava, 2013). L'apparition du Big Data avec des technologies d'intégration avancées comme Hadoop, MapReduce (Dean et Ghemawat, 2008) présente une nouvelle opportunité d'analyse pour les systèmes décisionnels. Ainsi, les Entrepôts de Données (ED), notamment les processus d'extraction, de transformation et de chargement (ETL: Extract, Transform et Load) doivent s'adapter aux nouveaux composants et architectures du Big Data afin de supporter le déluge de différentes formes et structures de données. L'intérêt de la modélisation des processus ETL est de (i) réduire les coûts de mise en œuvre d'un ED et (ii) de produire des informations pertinentes pour l'aide à la décision. En outre, la véracité présente une nouvelle caractéristique de Big Data. Elle concerne l'exactitude des données constituant le support des décisions stratégiques dans les systèmes décisionnels. Cependant, les sources de données (même du même domaine) sont de

différents niveaux de qualités avec des écarts significatifs dans la couverture, l'exactitude et l'actualité des données fournies.

2. État de l'art

Nous présentons dans cette section quelques travaux abordant la conception des processus ETL et les architectures Big Data les plus représentatifs de l'état de l'art. Plus précisément, nous distinguons principalement deux approches: une approche traitant les données métiers et l'autre qui aborde les Big Data.

Dans la première approche, les travaux de (Vassiliadis et al., 2002) ont proposé la modélisation des processus ETL en adoptant des notations spécifiques. Le travail avec les nouvelles constructions conceptuelles, s'avère être notable, mais la non standardisation reste toujours une limite. En effet, l'adoption d'un standard reste un atout essentiel au niveau de la modélisation. (Trujillo et al., 2003) a considéré que le langage UML en tant que standard présente une solution de modélisation privilégiée à travers le diagramme de classes. Aussi, (Muñoz et al., 2008) a proposé d'élaborer une modélisation conceptuelle utilisant les diagrammes d'activités qui vérifient les besoins comportementaux. Suivant le même principe, la modélisation BPMN, issue du domaine Business Process Management (BPM), est apparue avec (Akkaoui et al., 2012) qui considèrent que les processus ETL peuvent être un type particulier d'un processus métier.

Pour la deuxième approche qui considère le Web comme source de données, de nombreux travaux ont été également proposés. En effet, ces travaux ont pris en considération l'apparition du Big Data, ce qui donne naissance à de nouvelles technologies (Hadoop, MapReduce, etc.) traitant généralement le niveau physique d'intégration. D'autres travaux de l'état de l'art ont abordé l'aspect décisionnel. Plus précisément les auteurs dans (Liu et al., 2011) ont proposé une approche qui se focalise sur la performance des processus ETL et qui traite les données volumineuses et adopte le paradigme MapRduce. Aussi, (Liu et al., 2013) a défini un Framework appelé CloudETL qui supporte la schématisation en étoile d'ED. (Bala et al., 2014) a proposé un Framework P-ETL qui présente un ETL proposant de paralléliser les opérations ETL à travers le paradigme MapReduce. Ce travail traite l'aspect d'intégration mais il néglige la structuration multidimensionnelle qui est une phase primordiale pour les opérations d'analyses ultérieures.

En conclusion, ces travaux ont couvert un large éventail de problématiques reliées à la modélisation physique des processus ETL. De plus, ils se sont basés sur des modèles de programmation assez complexes. En effet, la modélisation des processus ETL doit couvrir le niveau conceptuel à l'échelle des Big Data tout en faisant recours aux opérations de base des processus ETL et la structure multidimensionnelle.

3. Problématique

L'adaptation des processus ETL, au contexte Big Data, constitue une piste de recherche prometteuse (Bala et al., 2014 ; Liu, 2014). Elle pose de nombreux problèmes qui sont liés aux 4V et qui doivent être pris en considération au niveau de la modélisation des processus ETL. En effet, le volume et la vélocité sont traités généralement par l'intégration de nouvelles technologies (MapReduce, Hive, Hadoop,…) dans les processus ETL (Bala et al., 2014 ; Liu et al., 2013) afin d'ajouter la notion de parallélisme aux différents processus.

En outre, afin d'améliorer les résultats des opérations d'analyse OLAP (On-Line Analytical Processing), il faut garder la structure multidimensionnelle d'un entrepôt de données (Fait, dimensions). Or, la majorité des travaux ont négligé cette structuration. Par conséquent, notre but est de supporter la modélisation des processus ETL adaptés au contexte Big Data tout en considérant les 4V clés mentionnés ci-dessous. Nous résumons notre problématique à travers les deux questions suivantes :

- Comment modéliser les processus ETL, tout en intégrant les nouvelles technologies de Big Data (Hadoop, MapReduce, . . .), pour supporter la construction d'un entrepôt de données ayant une structure multidimensionnelle ?

- Comment conserver la qualité de grandes sources de données hétérogènes, lors de l'enchaînement des processus ETL proposés ?

4. Actions réalisées

Pour pallier aux problèmes d'intégration, nous nous sommes focalisés sur l'adaptation des processus d'intégration (ETL) classiques dans le contexte des Big Data. Nous visons à augmenter la performance et le temps de réponse des processus ETL en redéfinissant les opérations ETL sous le paradigme MapReduce. Ainsi, nous avons proposé une modélisation conceptuelle des opérations ETL de base (union, jointure, projection, nettoyage…) utilisant ce paradigme tout en conservant l'aspect multidimensionnel de l'ED. L'approche d'intégration que nous proposons est baptisée BigDimETL (Big Dimensional ETL). Cette approche présente les processus d'intégration conventionnels et qui s'exécutent en parallèle à travers le paradigme Map-Reduce. Notre but est de construire la structure multidimensionnelle dès la phase d'extraction des processus ETL. En effet, notre processus de traitement commence par une étape d'extraction et de partitionnement vertical de la source pour garantir la correspondance entre la structure des données et les éléments de notre schéma multidimensionnel. Cette étape réalise, ainsi, le partitionnement des données en affectant chaque dimension à une partition. Ceci permet de réduire le nombre de jointures entre les différentes partitions tout en évitant de disperser et de réassembler les dimensions. Puis, une étape de structuration est mise en place, elle consiste à structurer les données de manière à les adapter au paradigme Map-Reduce. Le traitement des dimensions est réalisé par la fonction Map qui regroupe

les différentes opérations nécessaires pour chaque partition indépendamment des autres partitions, afin de préparer les données de sortie sous la forme (clé, valeur). Par conséquent, la phase suivante consiste à faire appel à différentes fonctions d'agrégation et de réduction (SUM, MAX, AVG). En effet, cette phase est responsable de construire les tables de fait qui regroupent les différentes dimensions traitées dans la phase précédente.

5. Actions futures

Après l'établissement de l'étude bibliographique, et la définition des différents composants de notre architecture BigDimETL, nous envisageons d'aborder en profondeur la modélisation conceptuelle avec une évaluation des différentes phases de notre approche par rapport les travaux existants.

Aussi, nous prévoyons de traiter la deuxième partie relative à l'identification et la résolution des problèmes reliés à l'assurance de la qualité de grandes sources de données hétérogènes (Big data), au cours de l'enchaînement des processus ETL proposé.

6. Bibliographie

Bala M., Boussaid O., Alimazighi Z., Bentayeb F."PF-ETL : vers l'intégration de données massives dans les fonctionnalités d'etl", *CongrèsINFORSID*, 2014, p. 61–76.

Dean J. et Ghemawat S., "MapReduce: simplified data processing on large clusters", Magazine Communications of the ACM Vol. 51, 1, 2008, p. 107-113

Dong X. L. and Srivastava D.. "Big data integration". *Very Large Data Bases*, Vol. 6, No. 11, 2013, p. 1188–1189.

El AkkaouiZ., MazònJ., VaismanA., ZimànyiE. "Bpmn-based conceptual modeling of ETL processes". *In Data Warehousing and Knowledge Discovery*, 2012, p. 1–14.

Liu X.et, ThomsenC., Pedersen T. B.. "Etlmr: A highly scalable dimensional etl framework based on mapreduce", *Data Warehousing and Knowledge Discovery* – vol.8, 2013 p. 96-111.

Liu X., ThomsenC., PedersenT. B. "Cloudetl: scalable dimensional ETL for hive".2014 P. 195–206.

Vassiliadis P., Simitsis A., Skiadopoulos S. "Conceptual modeling for ETL processes", *Data Warehousing and OLAP*, 2002 p. 14-21.

Muñoz L., Mazón J.N., et al., "Modelling ETL processes of data warehouses with UML activity diagrams". *Move to Meaningful Internet Systems*, vol 5333, 2008, p. 44-53.

Trujillo J., Luján-Mora S., "A UML Based Approach for Modeling ETL Processes in Data Warehouses", *ER*, vol 2813, 2003 p. 307-320.

Développement d'une application Analytics pour les ressources humaines : approche orientée problèmes dans un contexte multi-clients

Lynda Atif

PSL, Université de Paris Dauphine/ Talentsoft
LAMSADE UMR CNRS 7243,
Place du Maréchal de Lattre de Tassigny,
75016, Paris
latif@talentsoft.com

MOTS-CLÉS : *SIAD, Processus de décision RH, Approche orientée problèmes, développement collaboratif*

KEYWORDS: *DSS, HR decision making, problem-driven approach, collaborative design*

ENCADREMENT : *Camille Rosenthal Sabroux (PR) et Michel Grundstein (Chercheur-Associé)*

1. Contexte

A l'heure du numérique et de la concurrence accrue, les éditeurs informatiques développent des milliers de logiciels pour aider les organisations dans la gestion de leurs activités internes. La conception de ces systèmes informatiques passe par une phase amont du cycle de vie de développement logiciel appelée « phase d'analyse». Cette phase a pour objectif de mettre en lumière les exigences du système à développer pour satisfaire les utilisateurs finaux.

Depuis de nombreuses années, diverses études se sont penchées sur les raisons d'échec et de réussite des projets de développement de systèmes informatiques. Le rapport CHAOS du Standish Group de 1994 fait partie des plus connus. Régulièrement mis à jour depuis la première publication en 1994, les résultats de 2002, 2006, puis de 2013, montrent que la phase analyse est l'un des facteurs clés de la réussite ou d'échec d'un projet informatique. De même pour les systèmes interactifs d'aide à la décision (SIAD), plusieurs études ont pu démontrer que la cause principale d'échec de projets de leur développement était relative à cette phase. Nous focalisons notre recherche sur les SIAD orientés données appelés aussi « Application Analytics ».

2. État de l'art

Le Système Interactif d'Aide à la Décision (SIAD) est un concept complexe. Sa complexité est liée en grande partie au nombre croissant de décisions de diverses natures que peut supporter ce système, à l'évolution rapide des technologies ou encore à l'émergence de nouvelles pratiques managériales. De nombreuses définitions existent dans la littérature (Gorry & Morton, 1971) et toutes concourent à dire que le système doit aider un décideur dans la résolution de problèmes non programmés et mal structurés. De la même manière que les définitions varient en fonction des auteurs, il n'existe pas de typologie standard des systèmes interactifs d'aide à la décision. On peut s'appuyer sur différentes typologies. Selon Rosenthal-Sabroux (1996) *« un système interactif d'aide à la décision peut prendre des formes extrêmement diverses selon l'importance que prennent des caractéristiques telles que : orientation données ou modèles, décision unique ou répétitive structurée ou non structurée, environnement statique ou dynamique et enfin décisions individuelles ou collectives*».

Le champ d'études des SIAD a vu apparaitre des méthodes de développement spécifiques et adaptées à chaque contexte contrairement au domaine des systèmes d'information numérique de gestion où on retrouve des méthodes de conception à visée généralistes. La plupart de ces méthodes citées dans la littérature se focalisent sur la modélisation de données. Les méthodes sont spécifiques à des outils : entrepôts de données, cubes, requêteurs, etc., et restent très centrées sur les phases aval de conception. De plus, la majorité minimise la phase d'analyse d'exigences à l'amont du projet de développement voire l'occultent. Notons que les termes exigence et besoin sont parfois considérés comme synonymes dans la littérature, parfois employés systématiquement ensemble. Les approches d'analyse des exigences pour l'aide à la décision, en particulier pour les SIAD orientés-données, sont souvent classées selon trois grandes catégories: l'approche dirigée par les données, l'approche dirigée par les besoins et enfin l'approche dirigée par les buts. Ces approches restent encore non probantes à ce jour (STA, 2015). Selon nous, les principales raisons sont :

- la non-spécification des liens entre les problèmes de décision auxquels font face les différents utilisateurs finaux et celles du SIAD développé en aval. Soulignons aussi que si, dans la majorité des approches, les problèmes sont regardés comme préexistants au processus d'analyse des exigences, ils ne sont cependant pas considérés comme nécessaire à être extraits ou élicités,
- . On sait aussi qu'« *un problème bien posé est un problème dont le caractère crucial vient d'une estimation produite collectivement et d'une formulation estimée acceptable par toutes les parties »* (Soubie & de Terssac, 91).

3. Problématique

Notre recherche part alors du postulat que développer un SIAD en particulier pour aider les décisions peu ou non structurées, ne peut se faire sans s'intéresser aux problèmes de décision qui se posent aux différents utilisateurs finaux, à la façon de les représenter collectivement, au contenu des décisions, à leur sens et au processus de prise des décisions. On pose ainsi la problématique du domaine dans une perspective pluridisciplinaire : Pourquoi et comment pratiquer une approche par les problèmes pour développer un SIAD ?

La perspective générale est ici celle de l'ingénierie des systèmes, non pas dans le but de décrire en détail telle ou telle méthode ou tel ou tel outil technique, mais pour s'arrêter sur les questions à se poser au cours du processus d'ingénierie, et en particulier sur l'impact des choix faits par les différentes parties prenantes du système à l'amont du cycle de développement.

Nous avons noté que le terme « décision » renvoie aux actions et processus de résolution de problèmes : un problème qui se pose à un individu ou à une organisation (Lévine, 1989). Il existe selon Longueville (2003) cinq types de problèmes de décision distincts :

- Description : problèmes associés à la caractérisation réelle de l'état courant de l'organisation;
- Investigation : problèmes associés aux relations entre deux ou plusieurs éléments de données ou phénomènes ;
- Explication : problèmes associés à l'établissement d'une relation cause à effet ;
- Prédiction : problèmes associés à la projection future basée sur des données historiques ;
- Prescription : problèmes associés à la projection normative basée sur des données historiques.

À ce titre, notre approche orientée problèmes pour développer un SIAD se base sur une catégorisation des problèmes que se posent les utilisateurs finaux à l'amont du processus de développement puis la conception en miroir de la solution Analytics à l'aval. Il en ressort une nouvelle typologie de SIAD que nous proposons : Descriptif Analytics, Diagnostic Analytics, Explicatif Analytics, Predictif Analytics, Prescriptif Analytics.

4. Actions réalisées

Nos travaux de recherche se déroulent dans le cadre industriel de la société Talentsoft dans le cadre d'une bourse CIFRE. Une société spécialisée dans le développement de progiciel de gestion de ressources humaines (RH) en Saas à destination de différentes entreprises clientes. Afin d'apporter plus de valeur ajoutée, Talentsoft a décidé de concevoir un SIAD orienté données (une application Analytics) en complément du progiciel de gestion des ressources humaines afin d'aider les RH dans la résolution de leurs problèmes de décision non structurés (de

description, d'investigation, d'explication, de prédiction et de prescription) qui se posent dans chacun de leur processus de gestion : Le recrutement, l'évaluation des collaborateurs, la formation et le processus d'administration.

Nous avons sollicité cinq clients de tailles différentes pour faire partie de l'équipe projet jusqu'à la mise en production du SIAD puis nous avons appliqué notre approche de développement qu'on peut synthétiser en cinq étapes :

- La compréhension du domaine et l'explicitation collective des problèmes de décision qui se posent aux différents utilisateurs finaux du même profil mais d'entreprises différentes ;
- Formalisation et validation des problèmes de décision à traiter ;
- Traduction des problèmes de décision en spécifications fonctionnelles, et la négociation entre les parties prenantes pour parvenir à une solution acceptable ;
- La spécification des caractéristiques du système à construire et leur documentation ;
- La vérification et la validation de l'adéquation entre la solution Analytics développée et les problèmes de décision qui se posaient aux utilisateurs finaux dans ce processus RH.

5. Actions futures

Notre approche est toujours en cours de validation sur le terrain. L'identification des problèmes et l'élaboration de la solution s'effectuent en interaction, on doit alors procéder à de nombreuses transformations des représentations associées à l'artefact afin de construire une représentation de plus en plus détaillée du but à atteindre.

6. Bibliographie

Annoni E., Éléments méthodologiques pour le développement des systèmes décisionnels dans un contexte de réutilisation, Thèse de doctorat, Université de Toulouse I, 2010.

Arnott D., Dodson G., " Decision Support Systems Failure ", *Handbook on Decision Support Systems*, Berlin Heidelberg, Springer Verlag, 2008, p. 763-790.

Giorgini P., Rizzi S., Garzetti M., " GRAnD: A goal-oriented approach to requirement analysis in data warehouses ", *Decision Support Systems*, 45, 2008, p. 4-21.

Gorry G.A., Scott Morton M., " A framework for management information systems ", *Sloan Management Review*, vol. 13, n° 1, 1971, p. 50-70.

Lamsweerde A.V., " Requirements Engineering in the Year 00, A Research Perspective ", *International Conference on Software Engineering*, 2000.

Lévine P., Pomerol J.-C., *SIAD et systèmes experts*, Paris, Editions Hermès, 1989.

Longueville B. and Gardoni M., " A survey of context modelling: approaches, theories and use for engineering design researches ", *ICED 03*, Stockholm, Sweden, 2003.

Rosenthal-Sabroux C., Contribution méthodologique à la conception de systèmes d'information coopératifs, HDR, Université Paris Dauphine, 1996.

Standish Group, Chaos Report, Standish Group Report, 2015.

Recherche d'Information Contextuelle en Temps-Réel dans les Microblogs

Cas particulier de données Twitter

Thomas Palmer

Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9

Thomas.Palmer@irit.fr

MOTS-CLÉS : *Recherche d'Information - RI Contextuelle - Temps-Réel - Microblogs*

KEYWORDS: *Information Retrieval - Contextual IR - Real-Time - Microblogs*

ENCADREMENT : *Gilles Hubert (MCF) et Karen Pinel-Sauvagnat (MCF)*

1. Contexte

Nos travaux se situent dans le contexte de la Recherche d'Information (RI) en flux de données, et plus particulièrement dans les plateformes de micro-blogging telles que Twitter. Dans ce cadre, contrairement à la RI traditionnelle, la collection est constamment en évolution : de nouveaux documents arrivent en continu et en grande quantité pendant que d'autres peuvent être supprimés. Cette nature évolutive de la collection implique, pour de nombreuses tâches de RI, un traitement en temps-réel. Nos travaux se focalisent sur le filtrage temps-réel de tweet suivant un profil utilisateur, soit un (ou plusieurs) centre(s) d'intérêt(s). Une spécificité des tweets est leur limitation à 140 caractères, restreignant leur exploitation thématique. En revanche, ce type de documents apporte des éléments supplémentaires utilisables, que nous qualifions de contexte d'un tweet. Ces éléments de contexte concernent l'utilisateur qui a posté le message, mais aussi l'utilisation de composants spécifiques comme des images, des liens URL, ou encore des entités comme les hashtags.

2. État de l'art

Dans le domaine de la RI en flux de données, de nombreux travaux utilisent des données Twitter. En effet, leur accès est facile et fournit dans le même temps un certain nombre de caractéristiques contextuelles. Ces informations autour du document en lui-même (le contenu du tweet donc) sont capitales dans de nombreuses approches, et leur utilité dans un système de RI a été montrée à maintes reprises. Au niveau des données relatives à l'utilisateur, Ben Jabeur *et al.* (2011) se sont intéressés à l'influence du bloggeur et à l'importance de la structure du réseau social afin de dégager une pertinence thématique et sociale. D'autres travaux (Damak *et al.*, 2013) se sont concentrés entre autres sur la prise en compte primordiale d'URL dans le tweet. La structure même du tweet nous donne également accès à des formats très particuliers d'information au sein du document comme les *hashtags* ou les *mentions*. Certains travaux se sont concentrés sur cette particularité en montrant l'efficacité de leur prise en considération en recherche d'information de type hashtags (Efron, 2010) ou encore dans le domaine de la détection d'événements (Guille *et al.*, 2014). Cheng *et al.* (2013) ont combiné plusieurs de ces caractéristiques tout en mettant en œuvre un véritable traitement temps-réel du flux de tweets.

Par ailleurs, d'autres approches favorisent l'utilisation d'information externes notamment pour : l'expansion de requête (Liang *et al.*, 2012) afin de palier les problèmes de vocabulaire de Twitter, l'enrichissement de profils utilisateurs en amont mais aussi au fur et à mesure du processus (Gaglio *et al.*, 2015), ou encore l'expansion du document lui-même grâce au contenu du lien URL (Liang *et al.*, 2012).

3. Problématique

Notre problématique est de fournir aux utilisateurs un ensemble optimal de tweets répondant à leurs centres d'intérêt. Cette problématique est notamment celui de certains comptes Twitter, dont les administrateurs sont chargés de sélectionner de l'information à destination d'une population d'utilisateurs bien précise, sans jamais poster de messages personnels (Zhao *et al.*, 2014). Différents problèmes spécifiques apparaissent lorsqu'on cherche à automatiser ce principe. Tout d'abord, il faut relayer l'information « la plus pertinente » et donc fortement restreindre le nombre de tweets transmis pour ne pas surcharger l'utilisateur. La nouveauté de l'information doit être considérée, l'utilisateur ne souhaitant, en général, pas recevoir plusieurs fois la même information. L'obsolescence des messages transmis à l'utilisateur est également importante, le système doit donc identifier les tweets à relayer dans un temps très restreint. De plus, il peut arriver que pendant une certaine durée aucune notification ne doivent être transmise car aucune information pertinente n'est présente (notion de « période vide »). Dans ce cas, il faut être en mesure de déterminer si les tweets considérés pertinents le sont suffisamment pour être relayés. Enfin, comme tout système temps-réel, un problème récurrent est celui du démarrage à froid. Concrètement, le système prend du temps à intégrer certains aspects des profils utilisateurs s'il n'est pas entrainé en amont.

4. Actions réalisées

Pour répondre à cette problématique, nous avons proposé une approche prenant en compte le contexte d'un tweet, en plus du contenu du message lui-même, tout en privilégiant la vélocité du filtrage. L'approche cherche à exploiter les caractéristiques liées au tweet, ce que nous qualifions de contexte, pour améliorer les résultats retournés. Elle suppose également de disposer des centres d'intérêts utilisateurs produits en amont et qui vont représenter les profils utilisateurs. Cependant, cet aspect tout comme la présentation du résultat à l'utilisateur sortent du cadre des travaux réalisés.

Pour privilégier la vélocité du filtrage, nous avons choisi une approche incrémentale consistant à intégrer différents processus en veillant à conserver la dimension temps-réel du système global. L'approche proposée peut être divisée en deux parties. La première partie filtre les messages suivant la similarité entre leurs contenus textuels et les points d'intérêt utilisateur (profil utilisateur). La seconde partie repose sur un modèle de score intègre les différentes caractéristiques du contexte de tweet, relatives au contenu du message en lui-même (par exemple, qualité du langage et nombre de mots réels), aux données de l'utilisateur (par exemple, nombre de followers), et enfin aux entités du tweet (par exemple, hashtags et mentions). Le modèle calcule un score global pour chaque tweet en totalisant les points attribués pour chacune des caractéristiques. Un nombre de points est attribué à chaque caractéristique qui dépasse le seuil fixé pour celle-ci. Le démarrage à froid est ici corrigé par des expérimentations préalables visant à fixer les différents seuils. Les tweets dont le score global dépassent le seuil final sont notifiés à l'utilisateur.

Nous avons confronté notre approche à un jeu de données issu de la tâche TREC Microblog 2015 (Chellal *et al.*, 2015). Les différentes expérimentations menées montrent des résultats encourageants et des possibilités d'améliorations au regard des mesures d'évaluation définies pour cette tâche. Lors de ces expérimentations, nous avons entre autre fait varier les seuils de sélection. Nous avons pu constater une nette évolution des résultats, jusqu'à une amélioration de 30%.

5. Actions futures

Les actions futures seront, dans un premier temps, d'intégrer un processus d'enrichissement des profils utilisateurs par un ensemble de mots connexes. Ces mots pourront être extraits de ressources externes telles que les tops documents retournés par la Web API de Google pour chaque profil. Ceci vise à réduire davantage le démarrage à froid et ainsi augmenter la précision globale.

L'ajustement dynamique (Zhao *et al.*, 2014) des différents seuils utilisés, au fur et à mesure du processus, sera une des actions suivantes. Jusqu'à présent, ils étaient fixés en amont (issus d'expérimentations préalables) pour toute la durée de l'évaluation. Ils seront recalculés et mis à jour après un certain laps de temps (une journée par exemple) à partir de l'analyse des données des précédentes exécutions.

L'intégration d'un fenêtrage temporel (Gaglio *et al.*, 2015; Zhao *et al.*, 2014) est également projeté. Il s'agit de découper la période de sélection de tweets en fenêtres de durées régulières pour récupérer les tops tweets par fenêtre et ensuite sélectionner parmi eux ceux qui seront transmis aux utilisateurs.

Enfin, une des problématiques présentées, qui devra être abordée, concerne la nouveauté. Le but sera de vérifier que le document sélectionné fournit bien une information supplémentaire à ce qui a été renvoyé jusqu'à présent. Dans le même temps, une dernière problématique soulevée lors de la participation à TREC Microblog concerne les périodes vides, dont le bon traitement est indispensable à la satisfaction des utilisateurs.

6. Bibliographie

Ben Jabeur L., Tamine L., Boughanem M., « Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter », *Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI), Grenoble*, octobre, 2011.

Chellal A., Jabeur L. B., Soulier L., Moulahi B., Palmer T., Boughanem M., Pinel-Sauvagnat K., Tamine L., Hubert G., « IRIT at TREC Microblog 2015 », *34th Text REtrieval Conference (TREC 2015)*, 2015.

Cheng F., Zhang X., He B., Luo T., Wang W., « A Survey of Learning to Rank for Real-time Twitter Search », *Proceedings of the 2012 International Conference on Pervasive Computing and the Networked World*, ICPCA/SWS'12, p. 150–164, 2013.

Damak F., Pinel-Sauvagnat K., Boughanem M., Cabanac G., « Effectiveness of state-of-the-art features for microblog search », *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ACM, p. 914–919, 2013.

Efron M., « Hashtag retrieval in a microblogging environment », *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, p. 787–788, 2010.

Gaglio S., Lo Re G., Morana M., « Real-time detection of twitter social events from the user's perspective », *Communications (ICC), 2015 IEEE International Conference on*, IEEE, p. 1207–1212, 2015.

Guille A., Favre C., « Mention-anomaly-based event detection and tracking in twitter », *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, IEEE, p. 375–382, 2014.

Liang F., Qiang R., Yang J., « Exploiting Real-time Information Retrieval in the Microblogosphere », *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, ACM, New York, NY, USA, p. 267–276, 2012.

Zhao X., Tajima K., « Online Retweet Recommendation with Item Count Limits », *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*, WI-IAT '14, IEEE Computer Society, Washington, DC, USA, p. 282–289, 2014.