Forum Jeunes Chercheuses Jeunes Chercheurs : Actes de la 10e edition

Pierre-Emmanuel Arduin

FORUM JCJC 2020

Forum Jeunes Chercheuses Jeunes Chercheurs

Actes de la 10e édition

INFORSID 2020

Contenu des actes

Introduction au Forum Jeunes Chercheuses Jeunes Chercheurs d'INFORSID 2020

Pierre-Emmanuel Arduin

Université Paris-Dauphine, PSL, DRM UMR CNRS 7088
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France
pierre-emmanuel.arduin@dauphine.psl.eu

Tous les deux ans, le Forum jeunes chercheuses jeunes chercheurs (JCJC) d'INFORSID permet à des doctorants en première ou deuxième année de thèse de confronter leurs travaux à la communauté INFORSID. Il prodigue aussi – et surtout – un instantané du paysage de la recherche française en systèmes d'information. Quelles préoccupations portent ces collègues en devenir ? Quels sont leurs objets d'étude ? Quelles motivations scientifiques, industrielles ou sociétales ont les doctorants français d'aujourd'hui en systèmes d'information ? Le Forum JCJC d'INFORSID permet d'en apprendre davantage sur ces questions, tout en transmettant à des doctorants les conseils, la bienveillance et la rigueur qui font la force de la communauté scientifique française en systèmes d'information.

Tous les deux ans, donc, le Forum JCJC se tient pendant le congrès INFORSID et les doctorants ayant soumis un article établissent un premier contact avec la communauté. Tous les deux ans sauf... en 2020 ! La crise épidémique mondiale a entraîné l'annulation du congrès INFORSID 2020 à Dijon, mais ce n'est que partie remise puisque les auteurs d'un article du Forum JCJC 2020 pourront venir le présenter à Dijon pendant le congrès INFORSID 2021.

Sept auteurs ont contribué aux articles présents dans ces actes (voir biographies page 33). Il conviendra de les remercier – ainsi que leurs directeurs de thèse – pour leur contribution, leur ponctualité et leur capacité à intégrer les remarques des relecteurs dans un temps record et un contexte de travail à distance tendu que nous avons tous traversé.

Dans le premier article, Abou-Jamra (2020) présente une approche pour détecter des signaux faibles dans les réseaux sociaux. L'approche est basée sur l'analyse de la structure latente du réseau et se fait par le biais de graphlets. Une série temporelle est construite et les intervalles précédents et pendant un évènement significatif sont analysés en terme de vitesse de croissance/décroissance du nombre de graphlets.

Dans le second article, Andriamampianina (2020) s'intéresse à la dimension temporelle des informations au travers de graphes temporels. Un état de l'art des études existantes ainsi que de leurs limites y est présenté. Notamment, la conception des concepts et des formalismes pour le développement d'applications basées sur des graphes temporels est envisagée.

Dans le troisième article, Cayèré (2020) propose une approche de traitement et d'exploitation de traces spatio-temporelles numériques dans le domaine du tourisme. L'objectif de cette recherche est de concevoir une plateforme modulaire permettant de créer des chaînes de traitement personnalisées afin de faciliter l'analyse de ces traces.

Dans le quatrième article, Chasseray (2020) tente de réduire au maximum le nombre de modules d'importation de données brutes dans un processus de peuplement d'une base de connaissances. Une telle base est souhaitée stable même lors de l'intégration d'une nouvelle source de données brutes, ce que l'auteur propose de discuter.

Dans le cinquième article, Guennouni (2020) introduit une nouvelle méthodologie d'explication automatisée d'évènements survenant au sein de systèmes d'information. La problématique de recherche s'articule autour du couplage d'informations fournies par des capteurs et d'un corpus documentaire.

Dans le sixième article, Ramdani (2020) propose d'optimiser le processus de e-recrument en concevant un système de recommandation capable de cibler des candidats potentiels à moindre coût. Cette recherche vise à concevoir une aide au recrutement et à l'optimisation du budget du recruteur via l'identification automatique des canaux pour maximiser les chances d'obtenir les meilleurs candidats avec un coût maîtrisé.

Dans le septième article, Tang (2020) tente de construire un système de recommandation de ressources pertinentes dans le domaine de l'apprentissage en ligne. Les informations sur le statut de l'apprenant doivent être obtenues et l'auteur propose d'utiliser le standard de suivi de réseau xAPI pour aider à enregistrer les traces d'apprentissage générées par les apprenants dans un environnement d'apprentissage collaboratif.

Il est important de remercier ici les membres du comité de relecture pour leurs conseils, suggestions et avis précieux pour les auteurs. Chaque auteur a reçu deux relectures et une méta-relecture sur sa soumission. La bienveillance était de mise et les auteurs ont intégré les remarques pour préparer une version finale de leur soumission. Ainsi, merci à :

- *Marie-Hélène Abel*, Université de Technologie de Compiègne, HEUDIASYC,
- *Armelle Brun*, Université de Lorraine, LORIA,
- *Benjamin Costé*, Airbus Cybersecurity,
- *Cyril Faucher*, Université de La Rochelle, L3i,
- *Thomas Hujsa*, LAAS-CNRS,
- *Sébastien Laborie*, Université de Pau et des Pays de l'Adour, LIUPPA,

– *Éric Leclercq*, Université de Bourgogne, Le2i,
– *Davy Monticolo*, Université de Lorraine, ERPI,
– *Elsa Negre*, Université Paris-Dauphine – PSL, LAMSADE,
– *Christian Sallaberry*, Université de Pau et des Pays de l'Adour, LIUPPA,
– *Marinette Savonnet*, Université de Bourgogne, Le2i.

Le bureau d'INFORSID a enfin toute ma reconnaissance pour l'honneur qui m'a été fait de me confier ce 10^e Forum JCJC, gérer les contributions des doctorants et la préparation de ces actes. L'organisation n'en est pas complètement terminée avec ce document puisque les auteurs pourront venir s'ils le souhaitent présenter leurs travaux l'année prochaine pendant INFORSID 2021.

Le lecteur pourra apprécier le caractère temporel, applicatif, collecte et analyse de traces, peuplement de bases de connaissances, explicabilité, recommandation et apprentissage en ligne des soumissions au Forum JCJC d'INFORSID 2020. Il y a là un paysage de la recherche française en systèmes d'information clairement symptomatique de l'actualité que nous traversons : une recherche dynamique d'interactions traçables, recommandables et explicables.

Pierre-Emmanuel Arduin
Organisateur du Forum JCJC 2020

Bibliographie

Abou-Jamra H. (2020). Analyse de la structure latente des réseaux sociaux par graphlets. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 5–8.

Andriamampianina L. (2020). Toward a generic approach to capture the temporal evolution in graphs. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 9–12.

Cayèré C. (2020). Plateforme etl dédiée à l'analyse de la mobilité touristique dans une ville. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 13–16.

Chasseray Y. (2020). Un méta-modèle pour la population d'ontologie indépendamment du domaine. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 17–20.

Guennouni N. (2020). Interprétation d'événement dans un système d'information hétérogène – Analyse croisée de données issues de capteurs et de corpus documentaires. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 21–24.

Ramdani H. (2020). Un système intelligent pour l'optimisation du e-recrutement. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 25–28.

Tang Q. (2020). Towards pedagogical resources recommender system based on collaboration's tracks. In *Actes du Forum jeunes chercheuses jeunes chercheurs d'INFORSID 2020*, p. 29–32.

Analyse de la structure latente des réseaux sociaux par graphlets

Hiba Abou Jamra

Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté - 9, Avenue Alain Savary
F-21078 Dijon - France
Hiba_Abou-Jamra@etu.u-bourgogne.fr

MOTS-CLÉS : *Signaux faibles, Graphlets, Structure des réseaux, Twitter, Détection d'évènements.*

KEYWORDS: *Weak signals, Graphlets, Network structure, Twitter, Event detection.*

ENCADREMENT : *Marinette Savonnet et Éric Leclercq*

1. Introduction et Problématique

La détection de signaux faibles à partir d'informations cachées dans la masse de données produites quotidiennement sur les réseaux sociaux est un enjeu important puisqu'elle permet d'anticiper des prises de décision en matière de politique industrielle, commerciale, sanitaire et de stratégie de communication tout en projetant des scénarios d'avenir.

La première théorisation des signaux faibles a été proposée par Ansoff (Ansoff, 1975) qui place son étude dans le contexte de la planification et de la gestion des enjeux stratégiques des entreprises. Nous avons adopté sa définition dans laquelle les signaux faibles sont vus comme les premiers symptômes de discontinuités stratégiques qui agissent comme une information d'alerte précoce, de faible intensité, pouvant être annonciatrice d'une tendance ou d'un évènement important. Lesca and Blanco (2002) présentent des caractéristiques permettant d'identifier un signal faible dont : **fragmentaire**, **visibilité faible**, **peu ou pas familier**, **utilité faible** et **fiabilité faible**.

La quantité des données produite par les réseaux sociaux est si importante que les méthodes classiques s'appuyant sur des statistiques simples ne permettent pas d'extraire les signaux faibles. La construction d'outils algorithmiques travaillant plus localement est nécessaire.

Nous avons alors envisagé trois approches possibles parmi les outils du traitement des signaux comme le débruitage et la décomposition en ondelettes (Ranta *et*

al., 2003), les graphons (Glasscock, 2016) et les graphlets. Cependant, une décomposition en ondelettes sur une matrice d'adjacence ou de distance entre nœuds est difficilement interprétable. Les graphons s'appliquent plutôt sur des graphes denses, ce qui n'est pas le cas des graphes générés par les données des réseaux sociaux. Nous avons donc choisi de privilégier la piste des graphlets puisqu'ils n'ont pas besoin d'accéder à toute la structure du graphe durant la recherche. Notre hypothèse de travail est qu'ils permettent de déterminer des structures récurrentes ou patterns (petits graphes) qui se révèlent être des précurseurs d'évènements. Afin de tester notre hypothèse, nous avons réalisé des expérimentations sur les données du projet ISITE (*Initiatives Science Innovation Territoires Économie en Bourgogne-Franche-Comté*) Cocktail [1] dont le but est de créer un observatoire en temps réel des tendances, des innovations et des signaux faibles circulant dans les discours des contextes métiers alimentaire et santé sur Twitter.

Dans ce qui suit, nous présentons et discutons notre méthode et les premiers résultats obtenus à partir des expériences mises en œuvre sur des données réelles. Puis, nous présentons nos perspectives.

2. Approche et Expérimentation

Nous faisons comme hypothèse qu'il existe des patterns caractéristiques des signaux faibles, les graphlets, que l'on peut trouver à partir de la topologie du réseau.

Les graphlets ont été introduits pour la première fois par Pržulj *et al*. (2004). Un graphlet est un sous-graphe non isomorphe induit connecté (2 à 5 nœuds) choisi parmi les nœuds d'un large graphe. La figure 1 montre les 9 graphlets de 2 à 4 nœuds.

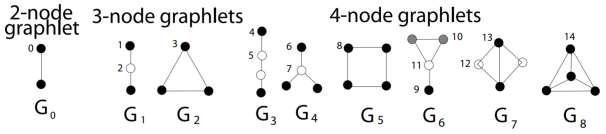


Figure 1. Représentation des 9 graphlets de 2 jusqu'à 4 nœuds

Il est ainsi possible de calculer pour un graphe une signature en terme de graphlets, c'est-à-dire en comptant le nombre de graphlets qui apparaissent dans le graphe. Plusieurs méthodes de décomposition en graphlet, c'est-à-dire d'énumération des graphlets, ont été proposées comme celle de Hočevar and Demšar (2014) que nous utilisons dans notre étude [2].

Cas d'étude : visite du Président Macron à Rouen le 30 octobre 2019 dans le cadre de l'incendie à l'usine Lubrizol

1. Ce travail est soutenu par le programme « Investissements d'Avenir », projet ISITE-BFC (contrat ANR-15-IDEX-0003), https://projet-cocktail.fr/

2. https://github.com/thocevar/orca

Le 26 septembre 2019, une partie de l'usine Lubrizol à Rouen a été ravagée par un incendie, l'intérêt de l'évènement nous a amené à lancer une collecte pour laquelle 47 hashtags et 111 comptes utilisateurs ont été fournis par les chercheurs en sciences humaines et sociales du projet Cocktail comme critères de collecte. Cette collecte a ramené environ 2 millions de tweets entre le 26 septembre et le 26 novembre 2019. Nous avons ensuite nettoyé ces données en filtrant les tweets avec les mot-clés `lubrizol` et `rouen` entre le 11 octobre et le 24 novembre. À l'issue de cette opération, le corpus est réduit à 137 561 tweets dont 16 100 sont des tweets originaux et 57 649 comportent des mentions. En construisant la série temporelle des tweets, nous avons vu que la venue du Président Macron à Rouen le 30 octobre au soir a provoqué une forte émission de tweets, indiquant que cette venue était un évènement. Nous avons alors décidé de vérifier notre hypothèse concernant les graphlets comme précurseurs d'évènement. Pour cela, nous avons découpé notre corpus en trois périodes autour de l'évènement de telle façon à avoir des graphes de taille comparable en terme de nœuds : P1 du 11 au 30 octobre midi (période avant l'évènement), P2 du 30 octobre midi au 3 novembre minuit (période à proximité de l'évènement), P3 du 4 au 24 novembre (période qui suit l'évènement). Afin de déterminer une signature topologique avant, pendant et après l'évènement, nous avons effectué les analyses suivantes :

1. une énumération des graphlets de 2 à 4 nœuds, en se basant sur la méthode de Hočevar and Demšar (2014) (G_0 à G_8 dans la figure 1) par pas de 24 heures. L'énumération est appliquée sur le graphe des mentions où les nœuds sont les comptes utilisateurs ;

2. un calcul des pentes normalisées comme suit, soit t représentant 24 heures durant une période P_i : $Pente\_graphlets_{P_i(t)} = (G_{X_{(t)}} - G_{X_{(t-1)}})/G_{X_{(t-1)}}$,
et $Pente\_mentions_{P_i(t)} = (G_{X_{(t)}} - G_{X_{(t-1)}})/n, avec\, i \in \{1, 2, 3\}, X \in \{0, \dots 8\}$
où $G_{X_{(t)}}$ est le nombre de graphlets de type X au temps t et n est le nombre des mentions dans le graphe durant la période P_i.

Pour comprendre l'évolution du nombre de graphlets, nous avons alors observé la vitesse de croissance/décroissance du nombre de graphlets donnée par les pentes normalisées. Sur les trois périodes, nous avons constaté que les signatures topologiques obtenues avec l'énumération en graphlets sont différentes. Le tableau 1 est un extrait des pentes normalisées par rapport au nombre de mentions entre le 23 et le 30 octobre (derniers jours de la période P1 et première demi-journée de P2).

Tableau 1. Pentes normalisées par mention (fin P1, début P2)

Jour	G0	G1	G2	G3	G4	G5	G6	G7
23/10	-0.550	-28.537	-0.041	-81.172	-1771.727	-0.562	-7.517	-0.631
24/10	-0.373	-23.305	0.020	-21.709	-1190.142	0.543	-9.415	-0.301
25/10	1.164	96.023	0.265	388.391	6704.945	48.029	52.198	6.501
26/10	-2.563	-207.338	-0.771	-875.825	-15158.076	-109.604	-122.724	-15.877
27/10	-1.976	-63.498	0.180	-124.576	-1915.135	-10.531	-2.220	1.151
28/10	0.301	1.837	-0.180	-15.277	15.017	3.419	-3.138	-1.183
29/10	0.195	-0.057	0.190	1.355	-13.057	-1.260	2.130	1.249
30/10	1.235	141.230	0.305	367.037	26899.298	6.197	152.724	5.765

Le calcul des pentes met en évidence que le 29 octobre (veille de la visite du Président Macron), la pente du graphlet de type G_6 devient positive et plus grande que

celle des autres graphlets (valeur 2.13 mise en évidence dans le tableau), le lendemain (date de l'évènement qui nous intéresse) les pentes sont toutes positives. Le graphlet G_6 peut être vu comme précurseur d'un évènement. Une deuxième singularité est aussi apparue, le 24 octobre la pente du graphlet de type G_5 devient positive (valeur 0.543 mise en évidence dans le tableau) puis le jour suivant les pentes sont toutes positives. Le graphlet G_5 peut être considéré ainsi comme un précurseur d'évènement car le 25 octobre la société Lubrizol a communiqué et cette communication a été considérée comme un évènement par nos collègues en science de la communication.

3. Conclusion et Perspectives

Cette étude liminaire a permis de conforter notre hypothèse qui considère que les graphlets sont des précurseurs d'évènements et qu'ils peuvent être vus comme des signaux faibles. En effet, les graphlets sont de petits patterns caractéristiques qui présentent des anomalies dans leur nombre avant et autour de l'apparition d'un évènement. L'étude des trois périodes montre bien des signatures en terme de nombre de graphlets différents, les données et les programmes de l'expérimentation sont disponibles (https://github.com/hibaaboujamra/GraphletLubrizol). Cette première étude doit être poursuivie avec l'étude d'autres évènements comme par exemple la diffusion de `#sansmoile7mai` entre les deux tours de l'élection présidentielle de 2017. Les graphlets ont été catégorisés en chemin, triangulés, troués, etc., à partir de cette catégorisation, nous voulons à partir du type de graphlet précurseur d'évènement connaître si un évènement aboutira à la construction d'une communauté et/ou à une diffusion virale.

Bibliographie

Ansoff H. I. (1975). Managing strategic surprise by response to weak signals. *California management review*, vol. 18, n° 2, p. 21–33.

Glasscock D. (2016). *What is a graphon?* Retrieved from https://arxiv.org/abs/1611.00718

Hočevar T., Demšar J. (2014). A combinatorial approach to graphlet counting. *Bioinformatics*, vol. 30, n° 4, p. 559–565.

Lesca H., Blanco S. (2002). Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles. In *6è congrès international francophone sur la pme*, p. 10–1.

Pržulj N., Corneil D. G., Jurisica I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, vol. 20, n° 18, p. 3508–3515.

Ranta R., Louis-Dorr V., Heinrich C., Wolf D., Guillemin F. (2003). Débruitage par ondelettes et segmentation de signaux non-stationnaires: réinterprétation d'un algorithme itératif et application à la phonoentérographie. *Traitement du signal*, vol. 20, n° 2, p. 119–135.

Toward a generic approach to capture the temporal evolution in graphs

Landy Andriamampianina [12]

1. *Institut de Recherche en Informatique de Toulouse (IRIT) - Université Toulouse 1 Capitole (UT1)*
   2 Rue du Doyen-Gabriel-Marty 31042 Toulouse
   landy.andria@irit.fr
2. *Activus Group*
   1 Chemin du Pigeonnier de la Cépière, 31100 Toulouse
   landy.andriamampianina@activus-group.fr

KEYWORDS: *Temporal graphs, historization mechanism, evolution types.*

MOTS-CLÉS : *Graphes temporels, mécanisme d'historisation, types d'évolution.*

ENCADREMENT : *Franck Ravat*

1. Introduction

Many domains, such as - computer networks, social networks, communication networks, biological networks etc. - are modelled as a graph as its flexible structure allows naturally representing complex relations among interconnected entities. In its most basic form, a graph is a collection of nodes with directed or undirected edges connecting them. The basic graph form is static, that is, it only reflects the snapshot[1] of a graph at a given time. However, real-world graph-based applications evolve constantly over time. Static graphs fail to represent, manage and trace their changes through time. Some studies have introduced the concept of temporal graphs to model the temporal evolution of graphs. Compared to static graphs, temporal graphs allow to model some evolution types in some specific applications. The purpose of our work is to propose a generic solution to manage any type of temporal evolution for any graph-based application.

1. an image of the entire graph at a given time.

2. State of the Art

A temporal graph is a graph representing nodes and edges, as well as all other parameters that the graph can have according to a model, that can be modified over time. The temporal evolution of a graph refers to the changes on its parameters over time, namely (i) *its topology*[2] and (ii) *data*[3] (Zaki *et al*., 2016).

The temporal evolution of graph topology is discussed in the literature through the addition and deletion of nodes and edges over time (Kostakos, 2009). Regarding the temporal evolution of graph data, it is discussed in the literature through changes in the attributes value of nodes (Desmier *et al*., 2012) or edges (Zhao *et al*., 2020) over time.

The standard temporal graph model in the literature is the sequence of graph snapshots[4]. It consists in splitting time into slices and then create a snapshot of the graph for each time slice (Fard *et al*., 2012). Other studies model the temporal evolution of the graph by attaching to nodes and edges their presence times instead of attaching time to the whole graph (Latapy *et al*., 2017).

In a nutshell, according to their applications, current modelling solutions include the temporal evolution of topology or data or both to track the temporal evolution of a graph. Most of them are based on the snapshot-based approach to keep temporal evolution traces.

3. Research questions

The first problem we want to address in our research is the representation of multiple types of the temporal evolution of a graph-based application. The state of the art shows that existing temporal graph models capture partially the temporal evolution of a graph. To the best of our knowledge, they do not capture the temporal evolution of graph data structure. Yet, this evolution type is a new information source that can enrich the evolution analysis of an application. As an example, we consider a social network where users and their relationships are respectively represented by nodes and edges in a temporal graph. Nodes and edges have attributes to describe users and their relationships characteristics. When a user adds a new friend, it creates a new edge in the graph modifying the graph topology and it updates the value of the attribute "friends list" of the user. When a user gets a job for the first time, a new attribute called "job status" should be added to its attributes set. This information can be new knowledge to understand evolution trends in the social network. For instance, the time when an important increase in the number of friends of a user may be correlated with the time when he added its job status, meaning that he possibly made new professional

2. the way in which the nodes and edges are arranged within a graph.

3. the attributes value of nodes and edges.

4. denoted $G_1, G_2, ..., G_T$ where G_i is an image of the entire graph at the time i and $[1; T]$ is the lifetime of the graph.

relationships. Therefore, some applications may require the complete aspect of temporal evolution i.e. the temporal evolution of graph topology, graph data and graph data structure. In this context, the first research question we want to answer is how to extend and generalize existing temporal graph concepts to be generic enough for capturing any evolution type and be reusable in any temporal graph based application.

The second problem we want to address in our research is the historization mechanism of temporal graph models. The snapshots-based model is the most used approach but it does not provide a straightforward solution to track and keep changes in a graph. To ensure the extraction of significant information on temporal evolution, it requires on the one hand to choose relevant time slice durations, which is a research topic itself (Ribeiro *et al*., 2013). On the other hand, it requires observing differences between two snapshots of the graph. Then it is difficult to track the individual temporal evolution of nodes and edges. The drawbacks of snapshots-based approaches are due to the fact that the temporal evolution is treated at the level of the entire graph. In this context, the second research question we want to answer is how to track and keep evolution traces at different graph levels (node, edge and graph) contrary to existing snapshot-based approaches.

4. Proposition and future actions

The objective of this thesis is to design concepts and formalisms for the development of applications based on temporal graphs. To do so, we study two research axis: the modelling and the manipulation of temporal graphs.

First of all, we focused on the modelling of temporal graphs. To meet the shortcomings of the state of the art, our objective is to propose a modelling solution based on the concepts of *temporal entities*, *temporal relationships*, *instances* and *states*. We give an outline of our model in the following paragraph.

We define a temporal graph as representing evolving entities and relationships in time called respectively *temporal entities* and *temporal relationships*. To track and keep evolution traces of the latter, we model them at two levels of abstraction: the *state level* to describe their attributes set called *schema*, and the *instance level* to describe their attributes value. At each time there is a change in the schema of a temporal entity or relationship, a new state is created. At each time there is change in the attributes value, a new instance of the state is created. We affect to each state and each instance a time value depicting the time when it is valid. Thus, a temporal entity or relationship is composed by a set of states and each state is composed by a set of instances.

Compared to existing temporal graph models, our model has several advantages. First, it manages temporal evolution at a finer granularity level by affecting time values to entities and relationships rather than the entire graph as in snapshot-based approaches. Then, it embeds a complete information about the temporal evolution of entities and relationships. Indeed, the affected time values to temporal entities and relationships capture when the addition of new entity or relationship and new attribute(s)

in the schema of an entity or relationship, when the deletion of an entity or relationship and attribute(s) in the schema of an entity or relationship and, when the change in the attributes value of an entity or relationship are made in the graph.

Our next plan is the manipulation of our previously defined temporal graph model. Our objective for this part is to define a high-level language for querying our temporal graph combining different principles: extension and adaptation to our context of (i) database query languages based on versions (Ozsoyoglu, Snodgrass, 1995), (ii) specific operators of temporal graphs (Moffitt, Stoyanovich, 2017) and (iii) inference of new components in the graph (Fang *et al*., 2011).

We will validate our modelling as well as its manipulation using data of real-world applications provided by the Activus Group company. We have already identified two implementation alternatives of our model in two data stores based on property graph model: Neo4j and OrientDB. We have defined translation rules to convert our model concepts into the property graph model.

References

Desmier E., Plantevit M., Robardet C., Boulicaut J.-F. (2012). Cohesive co-evolution patterns in dynamic attributed graphs. In *International Conference on Discovery Science*, pp. 110–124. Springer.

Fang C., Kohram M., Meng X., Ralescu A. (2011). Graph embedding framework for link prediction and vertex behavior modeling in temporal social networks. In *Proceddings of the SIGKDD Workshop on Social Network Mining and Analysis.*

Fard A., Abdolrashidi A., Ramaswamy L., Miller J. (2012). Towards Efficient Query Processing on Massive Time-Evolving Graphs. In *Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing.* IEEE.

Kostakos V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, Vol. 388, No. 6, pp. 1007–1023.

Latapy M., Viard T., Magnien C. (2017). Stream Graphs and Link Streams for the Modeling of Interactions over Time. *arXiv:1710.04073 [physics, stat].*

Moffitt V. Z., Stoyanovich J. (2017). Temporal graph algebra. In *Proceedings of The 16th International Symposium on Database Programming Languages*, pp. 1–12.

Ozsoyoglu G., Snodgrass R. (1995). Temporal and real-time databases: a survey. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 4, pp. 513–532.

Ribeiro B., Perra N., Baronchelli A. (2013, December). Quantifying the effect of temporal resolution on time-varying networks. *Scientific Reports*, Vol. 3, No. 1, pp. 3006.

Zaki A., Attia M., Hegazy D., Amin S. (2016). Comprehensive Survey on Dynamic Graph Models. *International Journal of Advanced Computer Science and Applications*, Vol. 7, No. 2.

Zhao A., Liu G., Zheng B., Zhao Y., Zheng K. (2020). Temporal paths discovery with multiple constraints in attributed dynamic graphs. *World Wide Web*, Vol. 23, pp. 313–336.

Plateforme ETL dédiée à l'analyse de la mobilité touristique dans une ville

Cécile Cayèré

La Rochelle Université,
23 Avenue Albert Einstein,
17000 La Rochelle, France
cecile.cayere1@univ-lr.fr

MOTS-CLÉS : *Données spatio-temporelles, trajectoire sémantique, plateforme modulaire, tourisme.*

KEYWORDS: *Spatio-temporal data, semantic trajectory, modular platform, tourism.*

ENCADREMENT : *Cyril Faucher, Christian Sallaberry, Marie-Noëlle Bessagnet et Philippe Roose*

1. Contexte

La traçabilité de la mobilité humaine est un phénomène qui prend beaucoup d'ampleur de par l'évolution des technologies GPS et l'augmentation des déplacements humains. Dans le projet régional Nouvelle-Aquitaine DA3T (i.e. Dispositif d'Analyse des Traces numériques pour la valorisation des Territoires Touristiques), nous faisons l'hypothèse que les traces numériques de mouvement laissées par les touristes peuvent aider les décideurs locaux dans la gestion et l'aménagement des territoires touristiques. Notre contribution vise des verrous de recherche *Mobilités et Trajectoires* [1] à des fins (i) d'extraction de sémantique à partir de données issue de capteur, (ii) de conception de méthodes de confrontation et de traitement conjoint de ces données avec des données issues d'enquêtes et d'analyse de l'environnement et (iii) de compréhension des comportements de mobilité. Nous utilisons la fouille de données et l'apprentissage automatique pour (i) calculer des indicateurs à partir de trajectoires brutes (e.g. objectif du déplacement, arrêts et déplacements, points chauds, etc.), (ii) créer différentes interprétations des trajectoires sur la base d'annotations et (iii) catégoriser les trajectoires de touriste.
Nous disposons actuellement des traces numériques de touristes volontaires issues de l'application mobile GeoLuciole. Elle a été développée dans le cadre du projet et

1. Website : http://gdr-magis.imag.fr/actions-prospectives/mobilites-et-trajectoires/

capture à intervalles de temps réguliers la position spatiale horodatée d'un téléphone mobile. Dans ce contexte, mon travail de thèse est de concevoir une plateforme modulaire permettant de créer des chaînes de traitement personnalisées afin de faciliter l'analyse de ces traces. Elle doit être simple d'utilisation et ergonomique car elle est destinée à des utilisateurs non-informaticiens. Elle doit proposer un ensemble d'outils modulaires permettant, par exemple, l'automatisation de la détection de zones d'intérêt (e.g. points chauds et froids etc.) ou la corrélation d'un certain comportement de mobilité avec le contexte courant (e.g. météo, période de l'année, etc.).

2. État de l'art

De la même manière que nous laissons des empreintes de notre passage dans un environnement physique, nous laissons également des empreintes numériques dans un environnement informatique. La trace est l'observation et l'interprétation d'un ensemble d'empreintes par un observateur (Mille, 2013). Lorsque l'empreinte possède une dimension spatiale et une dimension temporelle, on parle de **données spatio-temporelles**. (Flouvat, 2019) classe ces données dans quatre types : mobilité, événement, région et réseau. Les données qui nous intéressent sont de type mobilité et décrivent les positions spatiales d'objets mobiles dans le temps, nous parlons de **traces de mouvement**. (Parent *et al*., 2013) décrit une **trajectoire brute** comme une sous-partie d'intérêt de la trace de mouvement exhaustive. Pour transformer une trace de mouvement en une trajectoire brute, elle doit passer par un processus de reconstruction qui consiste à (i) nettoyer les données imprécises en les corrigeant ou les supprimant, (ii) faire de la cartospondance lorsque la trace de mouvement suit un réseau routier ou piétonnier (Newson, Krumm, 2016) et (iii) compresser les données afin de pouvoir les stocker efficacement. Afin de réaliser une analyse plus poussée, une trajectoire brute peut être enrichie avec des données provenant (i) de bases de données externes (e.g. bases de données ouvertes du tourisme, des collectivités, etc.), (ii) de capteurs externes, (iii) de calculs utilisant les données brutes (e.g. vitesse, orientation, etc), (iv) de l'avis d'experts ou (v) d'enquête ou d'interviews. (Spaccapietra, Parent, 2011) propose les **annotations** pour lier ces données additionnelles à la trajectoire ou à une partie de la trajectoire. Il introduit également le concept d'**épisode**. Il s'agit de la partie de trajectoire la plus longue répondant à un prédicat défini, basé sur les positions spatio-temporelles ou sur les annotations. Une **interprétation** de la trajectoire est une séquence d'épisode de cette trajectoire appartenant à une thématique donnée (e.g. modes de transport, activités, etc.), c'est-à-dire que tous les épisodes de cette interprétation appartiennent à cette thématique. Ainsi enrichie, la trajectoire brute devient une **trajectoire sémantique**. Les trajectoires peuvent être analysées de manière individuelle ou agrégée. Ces deux types d'analyse permettent de traiter la problématique de la mobilité touristique sous différents angles.

3. Problématique

Notre objectif est de proposer une plateforme modulaire permettant à un utilisateur de construire sa propre chaîne de traitement à partir des modules disponibles pour fa-

ciliter l'analyse des données de mobilité dont il dispose. Un module prend des données d'un type défini et un certain paramétrage en entrée, effectue un traitement sur ces données en tenant compte du paramétrage et renvoi le résultat. Les modules s'enchaînent dans la chaine de traitement afin de réaliser un traitement complet des données. Les modules peuvent être classés dans trois étapes inspirées des procédures ETL (i.e. EXTRACT, TRANSFORM, LOAD). La partie supérieure de la FIGURE 1 montre l'enchainement des différentes étapes et la partie inférieure illustre un exemple de chaîne de traitement. EXTRACT (c.f. FIGURE 1, bloc 1) regroupe les modules conçus pour extraire les données provenant de sources hétérogènes. TRANSFORM (c.f. FIGURE 1, bloc 2) peut être séparé en deux sous-étapes : (i) la reconstruction qui rassemble les modules permettant de transformer les traces de mouvement en trajectoires brutes et (ii) l'enrichissement qui rassemble les modules permettant de transformer les trajectoires brutes en trajectoires sémantiques. Enfin, LOAD (c.f. FIGURE 1, bloc 3) réunit les modules de visualisation et de stockage des résultats des étapes précédentes. Cette plateforme est destiné à des utilisateurs non-informaticiens, son utilisation doit donc être intuitive afin qu'elle soit facile à prendre en main. Ainsi, les types d'entrées et de sorties ainsi que les descriptions de chaque module doivent être documentées. L'ergonomie de la plateforme doit permettre à l'utilisateur de concevoir rapidement une chaîne de traitement avec les modules pour faire les traitements qu'il souhaite sur les données dont il dispose.

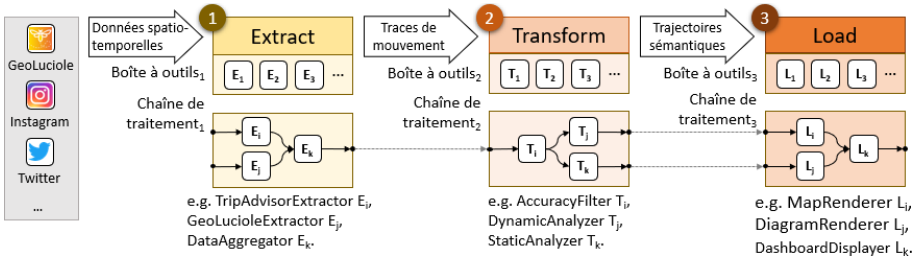


Figure 1. Chaîne de traitement des données inspirée des procédures ETL

Nous nous appuyons sur l'article de (Parent *et al.*, 2013) pour modéliser les données qui circuleront à travers la chaîne de traitement. Une trajectoire brute Tb d'identifiant k d'un objet mobile O est définie telle que :

$$Tb = (k, O, \langle (t_i, p_i, v_i) \rangle_{i \leq n})$$

Chaque tuple de la séquence contient la position p_i de l'objet mobile O capturée au temps t_i, où v_i représente l'ensemble possiblement vide des données brutes additionnelles (i.e. vitesse, orientation, précision, etc.).

La trajectoire sémantique Te d'identifiant k d'un objet mobile O enrichie par l'ensemble des interprétations I_{Σ_k} est définie telle que :

$$Te = (k, O, \langle (t_i, p_i, v_i) \rangle_{i \leq n}, \{I_{\Sigma_k}\})$$

Une interprétation est une séquence $\langle (t_{aj}, t_{bj}, a_j) \rangle_{j \leq m}$ où chaque tuple est un épisode, avec a_j l'annotation de l'épisode appartenant à l'alphabet Σ_k compris entre les temps t_{aj} et t_{bj}.

4. Actions réalisées

Cette section illustre l'état actuel de notre travail à travers quelques exemples de modules utilisant les données issues de GeoLuciole.

– EXTRACT : Le module GEOLUCIOLEEXTRACTOR exporte les traces de mouvement et les convertit au format GeoJSON.

– TRANSFORM : La position capturée n'est pas toujours fidèle à la position réelle du smartphone. Le module ACCURACYFILTER filtre les positions d'une trajectoire en fonction d'un seuil de précision donné. Nous testons également un algorithme replaçant les positions sur la route la plus proche mais cette technique entraîne des erreurs. Les modules d'enrichissement implémentés permettent d'analyser les trajectoires de manière statique (i.e. STATICANALYZER) ou dynamique (i.e. DYNAMICANALYZER). Ce dernier module prend en entrée un pas de temps (e.g. mois, jour, heure, minute) et utilise l'horodatage de chaque position pour échantillonner les trajectoires.

– LOAD : Le module MAPRENDERER prend en entrée une ou un ensemble de trajectoires et génère un rendu cartographique.

5. Actions futures

Nous avons travaillé sur l'étape EXTRACT et sur la reconstruction des données. En ce qui concerne la cartospondance, nous faisons des recherches sur des méthodes promettant des résultats plus fiables. D'importants travaux de recherche portent sur l'étape d'enrichissement et une grande partie de nos efforts est sur la création de modules d'annotation. Concernant l'étape LOAD, nous travaillons sur un module de type *tableau de bord* affichant différentes représentations paramétrables des données. L'efficacité de la plateforme sera évaluée par les géographes sur un nouveau jeu de traces de mouvement produit durant la période touristique 2020.

Bibliographie

Flouvat F. (2019). *Extraction de motifs spatio-temporels : co-localisations, séquences et graphes dynamiques attribués*. thesis, Université de la Nouvelle-Calédonie.

Mille A. (2013). De la trace à la connaissance à l'ère du Web. Introduction au dossier. *Intellectica*, vol. 59, n° 1, p. 7–28.

Newson P., Krumm J. (2016, December). Hidden Markov Map Matching Through Noise and Sparseness.

Parent C., Spaccapietra S., Renso C., Andrienko G. L., Andrienko N. V., Bogorny V. *et al.* (2013). Semantic trajectories modeling and analysis. *CSUR*.

Spaccapietra S., Parent C. (2011). Adding Meaning to Your Steps (Keynote Paper). In M. Jeusfeld, L. Delcambre, T.-W. Ling (Eds.), *Conceptual Modeling ? ER 2011*, p. 13–31. Berlin, Heidelberg, Springer.

Un méta-modèle pour la population d'ontologie indépendamment du domaine

Yohann Chasseray

Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France
4 allée Émile Monso, 31030 Toulouse, France
yohann.chasseray@ensiacet.fr

MOTS-CLÉS : *Bases de connaissances, extraction de connaissances, ingénierie dirigée par les modèles, transformation de modèle, interprétation du langage naturel.*

KEYWORDS: *Knowledge bases, knowledge extraction, model driven engineering, model transformation, natural language processing.*

ENCADREMENT : *Jean-Marc Le Lann, Anne-Marie Barthe-Delanoë et Stéphane Négny*

1. Contexte

Dans de nombreux domaines scientifiques et techniques, l'utilisation de systèmes experts permet d'offrir à l'humain une assistance pour des tâches allant du diagnostic à la prise de décision. De tels systèmes permettent de résoudre des problèmes complexes en se basant sur l'association d'éléments de connaissance et de règles de déduction simples. L'un des pré-requis pour l'utilisation de tels systèmes est l'existence de bases de connaissances permettant de transcrire l'état de la connaissance d'un domaine dans un format structuré et interprétable par ces systèmes experts. Le développement des ontologies répond en partie à ce besoin de structuration de la connaissance en fournissant à la fois la possibilité de définir les concepts clés d'un domaine et, via des moteurs d'inférence, de raisonner sur ces concepts. Depuis leur définition par Gruber (Gruber, 1993), les ontologies ont été largement développées dans des domaines tels que le domaine médical ou celui de la gestion de crise. Malheureusement ces ontologies ne peuvent pas être utilisées dans des cas d'application concrets sans qu'un expert du domaine ne les complète. Parallèlement, le nombre de sources de données renfermant de la connaissance ne cesse de se multiplier. Malheureusement, la connaissance existant dans ces sources de données (articles et ouvrages scientifiques, bases de données, compte rendu d'experts) est souvent non structurée. Il s'agit donc de s'intéresser à l'extraction automatique et générique de la connaissance contenue dans ces sources de données afin de peupler des ontologies, quel que soit le domaine associé.

2. État de l'art

Aujourd'hui, de nombreuses études s'intéressent à l'automatisation de la population d'ontologies. Les plateformes comme Datalift (Scharffe *et al*., 2012) ainsi que les modèles comme l'Ontology Definition Metamodel (ODM) développé par l'Object Management Group sont autant d'outils pour traiter différents formats de données structurées et les inter-relier sous un standard commun. Mais majoritairement, les systèmes d'extraction sont dédiés à des formats de données déjà structurés ou à une source unique de données non structurée (Faria *et al*., 2014). Certains frameworks présentent l'avantage de traiter une grande diversité de sources de données (Remolona *et al*., 2017). Néanmoins, ces approches utilisant de la connaissance du domaine déjà structurée sont limitées pour des domaines dans lesquels ces données n'existent pas encore. Enfin, les systèmes qui ont une approche générique vis-à-vis du domaine altèrent dans le même temps la structure de l'ontologie voire recréent celle-ci à partir des données (Hillairet *et al*., 2008) et perdent la structure préalablement établie par les experts du domaine. Ainsi, les systèmes de population automatique d'ontologies se classent en deux catégories : soit ils sont spécifiques à un domaine et/ou à une source de données, soit ils altèrent la structure de l'ontologie à peupler.

3. Problématique

Comme évoqué dans la section précédente, les limites que l'on retrouve dans les systèmes de population d'ontologie sont (i) la dépendance du système vis-à-vis du domaine décrit par l'ontologie, (ii) la limitation des chaînes d'extraction à certaines sources de données et (iii) la modification de la structure de l'ontologie lors du processus de population. L'objectif des travaux présentés ici, est de mettre en place un système de population d'ontologie (semi-)automatisé sans altérer cette dernière. Ce système se veut ainsi interopérable et indépendant à la fois des sources de données à traiter et de l'ontologie ciblée.

4. Actions réalisées

Pour faire le pont de manière générique entre les données hétérogènes et les domaines métiers pour lesquels des ontologies sont définies, un méta-modèle générique (FIGURE 1) pour l'extraction de données a été construit. D'une part, ce méta-modèle est défini dans un souci de généricité de manière à englober un large spectre de sources de données. D'autre part, la structure du méta-modèle défini se rapproche de la structure généralement adoptée par les ontologies. De cette manière, il est aisé de dériver des alignements entre ce méta-modèle et l'ontologie cible.

Dans le méta-modèle présenté, la classe *Entité* représente les objets extraits des données et pouvant être ou bien associés à des objets déjà présents dans l'ontologie cible, ou bien ajoutés à cette ontologie cible. La classe *Objet ontologique* représente les objets constitutifs de la connaissance (*Concept* ou *Instance*) que l'on retrouve habituellement dans une ontologie. La classe *Relation*, identifie les relations détectées

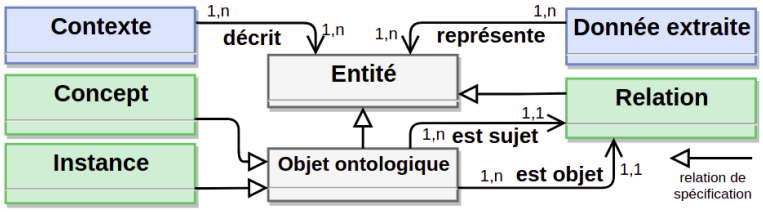


Figure 1. Méta-modèle générique pour l'extraction de données hétérogènes.

dans les données entre deux *Objets ontologiques*. Il n'est pas utile de représenter l'ensemble des données brutes dans ce métamodèle. Néanmoins, la classe *Donnée extraite* embarque une partie de cette donnée brute (phrase contenant une *Entité* par exemple) et permet de garder une trace de la donnée dont est issue une *Entité*. Enfin, la classe *Contexte* permet d'embarquer des éléments de contexte pour décrire plus précisément les *Entités* extraites. Ces derniers peuvent être de nature très variée en fonction du types de données exploitées (termes co-occurrents, titre de figure, méta-données).

Un cadre méthodologique pour la population générique d'ontologies (FIGURE 2) est construit autour du méta-modèle présenté précédemment. Dans la lignée de travaux antérieurs sur la représentation des ontologies (Gaševic *et al.*, 2009), ce cadre s'appuie sur les principes de l'Ingénierie Dirigée par les Modèles (Kent, 2002). Le méta-modèle défini précédemment y est utilisé comme un pivot entre les données et l'ontologie cible que l'on souhaite peupler. Cette ontologie peut varier en fonction du domaine qu'elle décrit. On définit alors la base de connaissances comme la réunion des instances et de l'ontologie cible que viennent alimenter ces instances.

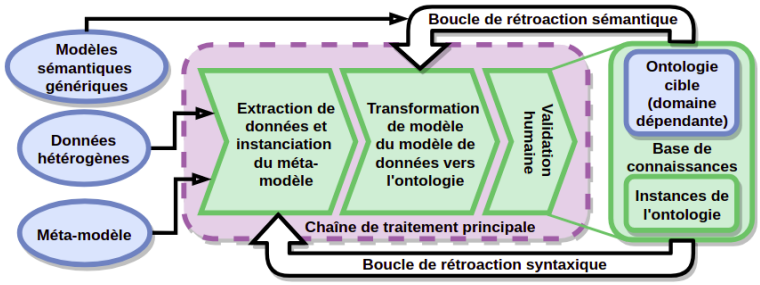


Figure 2. Cadre méthodologique pour la population d'ontologies.

Le cadre se structure en trois chaînes de traitement qui sont la *chaîne de traitement principale*, la *boucle de rétroaction sémantique* et la *boucle de rétroaction syntaxique*. *La chaîne de traitement* principale traite la donnée en deux étapes. La première étape extrait l'information de la donnée brute et l'organise en connaissance en s'appuyant sur les concepts du méta-modèle. Cela peut être effectué avec des méthodes d'extraction classiques (extraction par règles, traitement d'image, exploration de bases de données par exemple). De cette première étape naît un modèle de données qui contient des candidats potentiels (*Entités*) pour l'instanciation de l'ontologie. La

deuxième étape est une transformation de modèle du modèle de données vers l'ontologie cible. Cette transformation de modèle se base sur des règles d'alignement définies au niveau du méta-modèle et exprimées dans une logique d'ordre 1. Elle permet par exemple de transformer des instances issues du méta-modèle en instances de l'ontologie. Une étape de validation humaine est incluse à la chaîne de traitement principale afin de valider ou d'invalider les matchings effectués. Les deux boucles de rétroaction permettent de rendre le framework itératif et de limiter ainsi l'intervention humaine au fil des itérations. De nouvelles règles d'extraction ainsi que de nouveaux alignements peuvent de cette manière être déduits à partir respectivement des boucles de rétroaction syntaxiques et sémantiques.

5. Actions futures

Afin d'illustrer l'utilité du méta-modèle et l'applicabilité d'un tel cadre, une preuve de concept logicielle est en cours de développement. Le cas d'étude retenu s'intéresse dans un premier temps à traiter des données textuelles (pages Web, documents PDF) pour peupler l'ontologie CheBI (Hastings *et al*., 2016), issue du domaine de la chimie. Il est prévu de mettre en oeuvre des méthodes d'interprétation du langage naturel pour l'extraction de données et des méthodes de matching sémantique pour réaliser l'alignement entre le modèle de données et l'ontologie.

Bibliographie

Faria C., Serra I., Girardi R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming*, vol. 95, p. 26–43.

Gaševic D., Djuric D., Devedžic V. (2009). *Model driven engineering and ontology development*. Springer Science & Business Media.

Gruber T. (1993). Knowledge acquisition. *A translation approach to portable ontology specifications*, vol. 5, n° 199-220, p. 10–1006.

Hastings J., Owen G., Dekker A., Ennis M., Kale N., Muthukrishnan V. *et al*. (2016). Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, vol. 44, n° D1, p. D1214–D1219.

Hillairet G., Bertrand F., Lafaye J. Y. (2008). Un processus dirigé par les modèles pour la création de bases de connaissance ontologiques. *IDM*.

Kent S. (2002). Model driven engineering. In *International conference on integrated formal methods*, p. 286–298.

Remolona M. F. M., Conway M. F., Balasubramanian S., Fan L., Feng Z., Gu T. *et al*. (2017). Hybrid ontology-learning materials engineering system for pharmaceutical products: Multi-label entity recognition and concept detection. *Computers & Chemical Engineering*, vol. 107, p. 49–60.

Scharffe F., Bihanic L., Képéklian G., Atemezing G., Troncy R., Cotton F. *et al*. (2012). Enabling linked data publication with the datalift platform. In *Workshops at the twenty-sixth aaai conference on artificial intelligence.*

Interprétation d'événement dans un système d'information hétérogène

Analyse croisée de données issues de capteurs et de corpus documentaires

Nabila Guennouni

Université de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA
2 Allée du Parc Montaury
64600 Anglet
Nabila.Guennouni@univ-pau.fr

MOTS-CLÉS : *Évènement, Capteur, Sémantique, 5W1H, Système d'information.*

KEYWORDS: *Event, Sensor, Semantic, 5W1H, Information System.*

ENCADREMENT : *Christian Sallaberry et Sébastien Laborie*

1. Contexte

Les réseaux de capteurs font désormais partie de notre quotidien. Ainsi, ces dernières années de nombreux domaines d'applications ont pu émerger grâce à ces technologies, tels que les smart-homes ou plus généralement les smart-cities. Dans ce contexte, des systèmes connectés sont exploités afin de détecter des événements et fournir des informations à un utilisateur pour lui permettre de prendre des décisions (Chen *et al*., 2014). Pourtant lorsqu'un événement est détecté, nous constatons deux limites aux approches existantes.La première limite est que l'utilisateur doit chercher de lui-même l'explication d'un phénomène par de multiples requêtes successives : quelles observations particulières ont déclenché l'événement ? Où sont localisés les capteurs associés à cet événement ? Existe-t-il d'autres données intéressantes fournies par d'autres capteurs ?... Bien entendu, l'utilisateur doit ensuite faire un énorme travail de compilation de données pour construire l'interprétation de l'évènement par lui-même. La seconde limite est que ces systèmes ne sont généralement pas connectés à des corpus documentaires qui pourraient pourtant fournir aux utilisateurs des informations complémentaires.

Dans cet article, nous montrons qu'au sein de la communauté des systèmes d'information (SI), les travaux récents de (Hamborg *et al*., 2018) relatifs au 5W1H ("What? Who? Where? When? Why? How?") sont une piste intéressante à exploiter pour offrir un premier guide à l'utilisateur dans l'explication d'un événement issu d'un environnement connecté (EC). Nous démontrons que ce travail qui, au départ, se concentre exclusivement sur des données issues d'articles de journaux, peut s'appliquer aux données d'un réseau de capteurs pour expliquer un événement. De plus, nous montrons que des questionnements 5W1H successifs permettent de rapprocher des données hétérogènes à des fins d'explication d'événement.

Notre proposition s'intègre dans le cadre du projet de recherche BISE2 [1] (Business Information System for Energy and Environment) dont l'objectif est de proposer une plate-forme générique de gestion des EC (modélisation 2D/3D de l'environnement, détection des évènements, indexation et requêtage d'information composée notamment de données de capteurs et de corpus documentaires etc.).

2. État de l'art

Nombreux travaux de gestion et l'interrogation sémantique des données de capteurs et des corpus documentaires ont été proposés dans la littérature. (Perry *et al*., 2011; Grandi, 2010) proposent deux extensions assez complètes de SPARQL qui permettent la modélisation de tous les éléments de EC. Néanmoins, elles souffrent de quelques limites au niveau des requêtes spatio-temporelles complexes et de la modélisation de la dynamicité de l'environnement. (Mansour, 2019) proposent HSSN une ontologie pour la modélisation des EC, en prenant en compte la mobilité des capteurs, la variété des données capturées et la diversité des plates-formes. ainsi que le langage EQL- CE qui s'appuie sur HSSN et qui permet la gestion de données de capteurs et la définition d'événement. Ces approches de modélisation et d'interrogation partagent la même limite, celle de ne pas permettre l'intégration des données issues des corpus documentaires. Les systèmes de gestion et d'interrogation des corpus documentaires hétérogènes ont été également énormément exploré dans la littérature, notamment en se basant sur des modélisations sous forme de graphes de connaissances, comme dans (Vijayarajan *et al*., 2016). Néanmoins la majorité des travaux proposés se focalisent uniquement sur les documents Web et ne permettent pas l'intégration des données issues des réseaux de capteurs.

Dans un EC qui intègre à la fois des données de capteurs et un corpus de documents hétérogènes, deux approches peuvent permettre d'interpréter le déclenchement d'un événement. Une première approche consiste à rechercher en parallèle en envoyant des requêtes sur les deux SI. Des techniques de mise en correspondance (Christen, 2012) peuvent ensuite établir des liens entre les différents résultats de ces recherches. Néanmoins, ces liens sont construits a posteriori de la demande d'explication, ce qui implique un temps d'attente considérable. De plus, les liens peuvent ne pas

1. http://slaborie.perso.univ-pau.fr/index.php/fr/recherche-fr/projet-bise2

être pertinents à cause des recherches isolées des deux côtés. Une seconde approche consiste à définir, avant le déclenchement d'événements, des correspondances entre capteurs et documents (Euzenat, Shvaiko, 2013). L'utilisateur peut ensuite exploiter ces liens lors d'une demande d'explication d'un événement. Dans ce contexte, une mise à jour de l'environnement entraîne inévitablement une nouvelle reconfiguration des liens de correspondance. Nous proposons de tirer parti de la notion d'événement pour construire une troisième approche, à la croisée des deux premières. En effet, un événement se définit au préalable et des liens peuvent se construire sur la base de cette définition (e.g., entre entités de l'EC et/ou d'autres entités décrites dans des documentations techniques). Lorsqu'un événement se déclenche, d'autres liens entre entités peuvent émerger et s'affiner de manière itérative. Dans ce cadre, l'approche 5W1H appliquée à ces liens permettrait de fournir un cadre structuré d'explications d'événement.

3. Problématique

Notre approche [2], consiste à exploiter la sémantique issue de (i) la définition de l'EC (e.g., un parking, des automobiles, des capteurs de CO2), (ii) la définition des événements (e.g., "Taux de CO2 élevé") et (iii) le déclenchement des événements (e.g., CO2 élevé à 10:05 au bloc C du parking). À l'étape 1 (définition de l'EC), des représentations sémantiques de réseaux de capteurs et de domaines métiers spécifiques intègrent le SI. Pour ce faire, des instances sont créées dans le SI lors de l'installation d'un capteur ou de l'introduction d'une documentation technique. À l'étape 2 (définition d'événement), un premier réseau de relations sémantiques peut-être construit (e.g., pour l'évènement "Taux de CO2 élevé" des liens sont créés entre les concepts "CO2", "Pollution" et "Qualité de l'air" des ontologies dédiées aux capteurs , automobile et parking, respectivement). Enfin, à l'étape 3 (déclenchement d'un événement), un second réseau de relations sémantiques plus fins est construit, sur la base des données spatio-temporelles issus du déclenchement de l'événement. Nous faisons l'hypothèse qu'une telle analyse croisée de données de capteurs et de corpus documentaires couplé avec permet de construire dynamiquement une explication du déclenchement d'un événement. L'application adaptée de questionnements de type 5W1H nous permettra de sélectionner plus efficacement, à chaque étape, les éléments informationnels pertinents.

Les verrous de recherche associés visent l'appariement sémantique : (i) Comment exploiter pleinement les données issues de la définition et du déclenchement d'événements, afin d'établir des connexions avec des ontologies de domaine ? (ii) Comment prendre à la fois en compte l'évènement et les questions de type 5W1H lors de la connexion d'ontologies ? (iii) Dans un second temps, comment gérer le déclenchement simultané de plusieurs événements ou d'événements composites ? D'autres verrous, liés à ces derniers, visent la Recherche d'information (RI) sémantique et la présentation de résultats de RI : Comment analyser et exploiter (i) les connexions entre

2. Figure visible à cette adresse http://slaborie.perso.univ-pau.fr/images/Documents/Nabila-Approach.JPG

concepts , (ii) les connexions entre instances et (iii) les valeurs d'instances? (iv) Comment composer une explication d'événement à partir de résultats de questions de type 5W1H ?

4. Actions réalisées et action futures

En ce qui concerne les action réalisées: au sein du projet BISE2, comme mentionné dans l'état de l'art, (Mansour, 2019) a proposé l'ontologie HSSN, le langage EQL-CE et la machine virtuelle eVM pour la détection des évènements. Le développement d'une plate-forme de gestion des données dans les EC est également en cours (modélisation 2D/3D, simulation des données de capteurs, etc.). Dans le cadre de cette thèse, un travail bibliographique sur les systèmes de gestion des EC et les corpus documentaires hétérogènes, ainsi que les systèmes d'interrogation 5W1H a été réalisé. Un jeu de données composé à la fois de valeurs de capteurs et de documents a été élaboré.

En ce qui concerne les actions futures: au sein du projet BISE2, le développement de la plate-forme se poursuit (intégration de la machine virtuelle eVM, module d'insertion des données des corpus documentaires, module de requêtage pour les non-experts, etc.). Dans cette thèse, je vais proposer une architecture fonctionnelle avec : (a) un module de gestion sémantique de capteurs, de documents ainsi que d'événements, (b) un module d'interconnexion de concepts/instances d'ontologies grâce à aux définitions/déclenchements d'événements, et (c) un module d'interrogation guidé par l'approche 5W1H permettant de fournir des explications pour un événement déclenché.

Bibliographie

Chen C. Y., Fu J. H., Sung T., Wang P.-F., Jou E., Feng M.-W. (2014). Complex event processing for the internet of things and its applications. In *Case 2014*, p. 1144–1149.

Christen P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Euzenat J., Shvaiko P. (2013). *Ontology matching*. 2nd edn, Springer Publishing Company.

Grandi F. (2010). T-sparql: A tsql2-like temporal query language for rdf. In *Adbis (local proceedings)*, p. 21–30.

Hamborg F., Lachnit S., Schubotz M., Hepp T., Gipp B. (2018). Giveme5w: main event retrieval from news articles by extraction of the five journalistic w questions. In *International conference on information*, p. 356–366.

Mansour E. (2019). *Event detection in connected environments*. Unpublished doctoral dissertation, Université de Pau et des Pays de l'Adour.

Perry M., Jain P., Sheth A. P. (2011). Sparql-st: Extending sparql to support spatiotemporal queries. In *Geospatial semantics and the semantic web*, p. 61–86. Springer.

Vijayarajan V., Dinakaran M., Tejaswin P., Lohani M. (2016). A generic framework for ontology-based information retrieval and image retrieval in web data. *Human-centric Computing and Information Sciences*, vol. 6, n° 1, p. 18.

Un système intelligent pour l'optimisation du e-recrutement

Halima Ramdani [1,2]

1. Equipe de Recherche sur les Processus Innovatifs, Université de Lorraine, Nancy, France

2. Laboratoire lorrain de recherche en informatique et ses applications, Université de Lorraine, Nancy, France

MOTS-CLÉS : *E-recrutement, système de recommandation, apprentissage automatique.*

KEYWORDS: *E-recruitment, recommander system, machine learning.*

ENCADREMENT : *Armelle Brun, Eric Bonjour et Davy Monticolo*

1. Introduction

Avec la révolution digitale, les canaux de diffusion d'offres d'emploi, de typologies différentes (réseaux sociaux, jobs boards, sites publicitaires), se multiplient. Chaque canal suit sa propre stratégie financière et cible des profils candidats particuliers. De ce fait, la tâche de recrutement en ligne devient de plus en plus difficile pour le recruteur. En effet, pour optimiser le budget de recrutement il est nécessaire d'analyser précisément les caractéristiques des canaux de diffusion et les offres d'emploi proposées afin de choisir les canaux qui offriront les meilleures chances de recrutement. Face à cette difficulté, l'objectif de nos travaux de recherche est de pouvoir aider le recruteur à optimiser son budget en identifiant automatiquement les canaux permettant de maximiser les chances d'obtenir les meilleurs candidats, tout en minimisant le coût dépensé. Cet article décrit notre problématique relative à l'optimisation du choix des canaux en minimisant le coût et maximisant le nombre de CV pertinents reçus. Nous présenterons un état de l'art du domaine du e-recrutement et nous exposerons les verrous scientifiques liés à la problématique.

2. Positionnement et problématique

Dans le cadre de notre recherche, nous nous plaçons en tant qu'intermédiaire entre le recruteur et les candidats en utilisant les canaux de diffusion pour optimiser le e-recrutement. Le e-recrutement est un processus qui intègre plusieurs acteurs : 1) Le recruteur, qui est à la recherche de candidats pertinents répondant au profil recherché dans l'offre d'emploi et à un coût optimal (prix auquel le recruteur est susceptible de maximiser le nombre de candidats pertinents, qui répondent aux profil recherché, en minimisant son budget). Le recruteur bien que spécialiste dans son domaine, n'a pas connaissance de l'ensemble des canaux de diffusions qui s'offrent à lui pour recruter. 2) L'intermédiaire du marché du travail (les cabinets de recrutement, les agences d'intérim etc.), qui intervient de deux manières différentes (Fondeur, Lhermitte, 2006) : soit en tant que support d'informations totalement neutre, soit en tant que vesteur de relation entre les candidats et les recruteurs. Ces deux dernières décennies, un autre type d'intermédiaire est apparu : les jobs boards (ou sites web de recherche d'emploi). Il existe donc deux types d'intermédiaires : ceux dits traditionnels et les canaux de diffusion qui regroupent les job boards, les réseaux sociaux, les médias, etc. 3) Le candidat, qui est présent sur le web est est à la recherche (ou non) d'un emploi adapté à ses compétences techniques et générales. L'ensemble de ces acteurs intervient dans le processus de e-recrutement. Les trois premières phases du e-recrutement global (Dhamija, 2012) se décomposent de cette façon: (1) Sourcing et identification qui consiste à rédiger et diffuser les offres d'emploi sur les canaux de diffusion. L'évaluation de cette phase se fait grâce à différents indicateurs de performance : le nombre de clics que l'annonce a reçu, est un indicateur qui permet au recruteur d'évaluer si son annonce est visible par les candidats. Le taux de conversion qui correspond au rapport entre le nombre de candidats qui ont cliqué pour visualiser une annonce et les candidats qui ont envoyé leur CV. Cet indicateur permet de déduire l'efficacité d'une campagne de recrutement, mais pas de sa rentabilité. Le coût par candidat correspond au budget dépensé pour recevoir un CV. Le coût par candidat pertinent qui correspond au budget dépensé pour recevoir un CV pertinent. Cet indicateur permet de déduire la rentabilité d'une campagne de recrutement. (2) Traitement et analyse des candidatures : consiste à recevoir, classifier et qualifier la véracité des informations transmises par le candidat au regard du poste proposé. (3) Évaluation et vérification des aptitudes : consiste à analyser la qualité des candidatures. Elle est souvent traitée manuellement par les recruteurs.

Les différents travaux de la littérature traitent les deux premières phases séparément. Néanmoins, la séparation de ces deux étapes a pour conséquence une allocation de temps supplémentaire pour le recruteur, ainsi qu'un budget de recrutement plus conséquent puisque les recruteurs n'ont pas connaissance des caractéristiques de chaque canal de diffusion. Par conséquent, la pertinence des CVs reçus n'est pas certaine entraînant ainsi une charge de travail importante lors de la phase d'évaluation des aptitudes. De ce fait, l'association de ces deux phases semble une piste pertinente pour l'amélioration du e-recrutement. Elle permettra ainsi d'optimiser le budget du recruteur en évaluant les candidats pendant la phase de sourcing. De ce fait, le recruteur

gagnera du temps pendant la phase d'évaluation des aptitudes puisqu'il ne recevra que des candidats en adéquation avec le profil recherché. Notre projet de recherche vise donc à fusionner les deux phases "Sourcing et identification" et "Évaluation et vérification des aptitudes" pour optimiser le e-recrutement, en particulier le coût lié à celui-ci.

3. Etat de l'art

Dans le domaine de l'optimisation du e-recrutement, (Séguéla, 2012) a proposé un système de recommandation des canaux de diffusions basé sur le contenu de l'offre d'emploi. Ce travail utilise un corpus de données contenant les offres d'emploi diffusées dans le passé et leurs données statistiques sur chaque canal. L'indicateur de performance utilisé est le taux de conversion. Nous constatons que la dimension temporelle de l'information relative au canal de diffusion n'est pas prise en compte, alors qu'il existe des périodes où les candidats sont plus actifs que d'autres (Hamel, 2015). Les travaux de (Benabderrahmane *et al*., 2018) ont permis de vérifier que la prise en compte de la temporalité permet d'avoir de meilleures performances. Dans ce travail, l'utilisation d'un système hybride (basé sur le contenu et collaboratif) pour recommander les canaux de diffusion s'est avérée pertinente. Ce système recommande les canaux en fonction du nombre de clics qui est considéré comme seul indicateur de performance d'un canal. Nous notons plusieurs limites dans les travaux existants : L1) L'absence de l'identification du profil recherché à partir de l'offre d'emploi pour la recommandation des canaux. L2) La prise en compte d'indicateurs de performance constitués par le nombre de clics et le taux de conversion. En effet, ces indicateurs ne permettent pas d'optimiser le budget et de maximiser le nombre de CV pertinents. Pourtant, la prise en compte du nombre de CV pertinents est fondamentale pour l'optimisation du budget du recruteur sur les canaux de diffusion puisque ce critère est indispensable pour évaluer la qualité d'une campagne de recrutement. L3) La caractérisation des canaux et des profils candidats que ces canaux véhiculent n'est pas étudiée. Plusieurs facteurs externes (marché du travail, secteur spécifique etc.) ou internes (diffusion plus fréquente sur certains canaux, biais humain sur le choix des canaux dans les données historiques) peuvent influencer la réception de clics ou de CV et donc la performance d'un canal. Ces facteurs reflètent un environnement mouvant et incertain, caractérisé par des paramètres difficilement identifiables (marché du travail, politique, secteur, etc.). Le caractère incertain de l'environnement n'est aujourd'hui pas pris en compte dans la littérature.

4. Verrous et problématiques scientifiques

L'objectif de notre recherche est de lever les limites des travaux de la littérature dans le domaine du e-recrutement en répondant aux problématiques suivantes : P1- a) Comment analyser et extraire de l'information à partir de texte semi-structuré et rédigé en langage naturel? (problématique liée à la limite L1) P1-b) Comment mettre en correspondance deux textes rédigés en langage naturel et semi-structurés? (lien avec

L2) P2- Comment concevoir un système de recommandation dans un environnement incertain? (lien avec L3) Nos travaux de recherche soulève la problématique générale suivante: Comment optimiser le choix des canaux de recrutement en fonction du profil recherché dans l'offre et de la pertinence des CVs reçues?

5. Conclusion et avancement

Nos premiers travaux ont eu pour objectif de répondre au premier verrou scientifique. Nous avons proposé une méthodologie d'indexation automatique de documents, reposant sur l'étiquetage de séquences. Cette méthodologie est validée au travers d'algorithmes d'apprentissage supervisé sur un corpus réel d'offres d'emploi. Nos travaux actuels concernent l'élaboration d'une méthode de mise en correspondance entre deux textes et la conception d'un système de recommandation dans un environnement incertain.

Bibliographie

Benabderrahmane S., Mellouli N., Lamolle M. (2018). On the predictive analysis of behavioral massive job data using embedded clustering and deep recurrent neural networks. *Knowledge-Based Systems*, vol. 151, p. 95–113.

Dhamija P. (2012). E-recruitment: a roadmap towards e-human resource management. *Researchers World*, vol. 3, n° 3, p. 33.

Fondeur Y., Lhermitte F. (2006). Réseaux sociaux numériques et marché du travail. *La Revue de l'IRES*, n° 3, p. 101–131.

Séguéla J. (2012). *Fouille de données textuelles et systèmes de recommandation appliqués aux offres d'emploi diffusées sur le web*. Unpublished doctoral dissertation.

Towards pedagogical resources recommender system based on collaboration's tracks

Qing Tang

Sorbonne universités, Université de technologie de Compiègne, CNRS UMR 7253, HEUDIASYC, Compiègne, France

qing.tang@utc.fr

MOTS-CLÉS : Apprentissage en ligne, Environnement d'apprentissage collaboratif, Système de recommandation, Traces, Modèle des apprenants

KEYWORDS: Online-learning, Collaborative learning environment, Recommender system, Learner tracks, Learner models

ENCADREMENT : Marie-Hélène Abel et Elsa Negre

1. Introduction

With the rapid and diversified development of education methods, online education is getting more and more attention. Online education is convenient, fast and efficient, but it faces a problem: Information overload, how to help learner choose the right learning resources and improve their learning efficiency? Research shows that collaborative learning can greatly improve learning quality (Zambrano, Kirschner & Sweller, 2019). Although learning efficiency has improved, many learners still cannot find suitable resources. Recommender systems (RSs) facilitate the exploration of resources and decrease learners' information load. The more the RS knows about learners, the more accurate and targeted the recommendations will be. Learner's information can be obtained explicitly and implicitly. For the latter, in collaborative learning environment (CLE), when learners generate learning activities, they leave digital historical messages, called 'tracks' (Li, Abel, Paul & Barthès, 2014). This paper is organized as follow: Section 2 describes the related work. Section 3 presents the main contribution, which is mainly about track model, learner model and rules of recommendation. Section 4 is discussion and future work.

2. Related work

Online learning platforms help individuals share information, participate and collaborate to learn from communities (Rennie and Morrison, 2013). 'Collaborative learning' means learning with someone to pursue the same learning goal, and learner in CLE is more productive than individual (Zambrano, Kirschner & Sweller, 2019). A group has individuals who usually possess different knowledge and skills. A CLE support group members' coordination so that they complete the task more efficiently (Kirchner, 2014). But some group members also face the problem of resource overload, RSs can handle the problem that learners are experiencing difficulties in retrieving useful and relevant learning resources. Existing RSs have recorded significant success in online-commerce domain, but most of them do not consider differences in learner characteristics (e.g., knowledge level, status information, etc.), which are necessary for improving the performance of recommender. Almost all the past interactions represent tracks that can be regarded as the learner's study experience. According to Clauzel, Sehaba and Prié (2009), a track is the history of users' actions collected from their interaction with the software. These researches emphasize the personal aspect, but provide little for answering the question on "how to share and reuse the users' experiences in a group" and do not provide an effective method to deal with tracks.

3. Contribution

3.1. Collect tracks

We consider these tracks as another kind of resource, which can help us analyze learner's status to improve the quality of recommendations, instead of simple records of learner's behaviors. It is necessary to build a model to analyze and utilize tracks data. xAPI[1] is a standard for learning technology that makes it possible to collect tracks data about the wide range of experiences a person has (online and offline) and can help simplify learner features representation. Via xAPI, every activity generated by learners in the CLE will be recorded. We define every activity as a track:

$$T = \left\{u_i, g_j, t, v, R_z\right\} \quad i,j,z \in \mathbb{N}^+ \tag{1}$$

In the above equation, T represents single track record, u_i represents the i^{th} learner, g_j represents the j^{th} group, t represents time, v represents action, and R_z represents the z^{th} resource.

3.2. Learner model

We propose a learner model consists of basic learner information (e.g., language, preference, etc.), learner contribution, learner credibility, and learner competency.

[1]. https://xapi.com/

Learner contribution represents a measure of the valuable activities made by learners in the learning group.

$$Con_{u_i}^{G_j}=\frac{n_u}{N}\times w_1+(n_g-n_b)\times w_2+n_s\times w_3+(n_d+n_v+n_a)\times w_4 \quad (2)$$

$Con_{u_i}^{G_j}$ represents the contribution of the i^{th} learner u_i to the j^{th} group G_j. n_u represents the amount of resources uploaded by learner u_i on group G_j, and N indicates the total amount of resources in this group. n_g and n_b represent the number of the resources vote status, (where vote status is 'goodvote' if resource vote ≥ 3, classified into n_g; and 'badvote' otherwise (vote$\in$[0,5]), classified into n_b); n_s represents the search times of the learner in this group; n_d represents the discussion number; n_v represents the vote number; n_a represents the annotate number.

Learner credibility is to what extent a learner can be trusted in the group.

$$Cre_{u_i}^{G_j}=Con_{u_i}^{G_j}\times w_5+V_k\times w_6+(t_c-t_s)\times w_7 \quad (3)$$

$Cre_{u_i}^{G_j}$ represents the credibility of the i^{th} learner u_i to the j^{th} group G_j. V_k represents the job title of the i^{th} learner u_i. t_c represents the current time and t_s represents the time of i^{th} learner joined this group, which is counted in year.

Learner competency indicates how familiar a learner is in a certain area.

$$Com_{u_i}^{G_j}=Cre_{u_i}^{G_j}\times w_8 \quad (4)$$

$Com_{u_i}^{G_j}$ represents the competency of the i^{th} learner u_i to the j^{th} group G_j. In previous equations, w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8 represent the weights.

Finally, a learner model is composed of 8 elements: Learner age, job title, learner contribution, learner credibility, learner competency, first language, second language and learning preference (theoretical or practical).

3.3. Recommendations

As indicated above, learners must be compared (similarity) to obtain relevant recommendations. The calculation rule is the cosine similarity using two 8-dimensional learner vectors (corresponding to the 8 elements of the learner model). The closer similarity between the two learners indicating that the learning ability and learning goals of the two are more similar. According to that, we can recommend resources to learners who study in the same group.

1. If Learners A and B have a highest similarity, from the resources B has used recently, select the resource with the highest vote (vote$\in$[3,5]) given by B, and recommend that resource to A.

2. If none of the resources Learner B has used has been given a vote 3 points or more, it means that he or she has not used high-quality resources recently.

3. From the other Learner C in the group with the second highest similarity with A, choose the resources that C has used recently, and select the resource that C gave the highest vote (vote$\in$[3,5]) to recommend to A.

4. If there is still no result, continue from step 2, until find a qualified resource.

4. Discussion and future work

The most used recommendation algorithm is collaborative filtering, but it has data correlation problem (Najafabadi, Mohamed & Choo, 2019). There is a tendency for learners to consume similar resources, the ratings for these resources will be similar. To avoid such problem and improve the recommendation accuracy, we must establish the learner's personalized features and define the learner model according to the features, so as to maximize the personalized recommendation. We choose to expand another direction (feature extraction), which means that we will make use of personalized behavioral features extracted from all the tracks generated by learners in the system. In the future, we will conduct experiments with the above technology on the MEMORAe[2] platform to compare the detection results.

References

Clauzel D., Sehaba K. and Prié Y. (2009). Modelling and visualising traces for reflexivity in synchronous collaborative systems, *Intelligent Networking and Collaborative Systems*, International Conference on. IEEE, pp. 16–23.

Kirchner F. (2014). Collaboration principle 3: Multimedia should facilitate effective and efficient communication and regulation of actions, *The Cambridge Handbook of Multimedia Learning*, pp. 561-566.

Li Q., Abel M.H., Paul J. and Barthès A. (2014). Facilitating collaboration and information retrieval: Collaborative traces based SWOT analysis and implications, *Distributed Systems and Applications of Information Filtering and Retrieval*, Springer Berlin Heidelberg, pp.65-78.

Najafabadi M.K., Mohamed A. and Choo W.O. (2019). An impact of time and item influencer in collaborative filtering recommendations using graph-based model. *Information Processing and Management*. Vol. 56, pp. 526-540, 2019.

Rennie F. and Morrison T. (2013). *E-learning and social networking handbook: resources for higher education Routledge.*

Zambrano J., Kirschner F. and Sweller J. (2019). Effects of prior knowledge on collaborative and individual learning, *Learning and Instruction*, Volume 63.

[2]. http://memorae.hds.utc.fr/

Biographies des auteurs

Hiba Abou Jamra est une doctorante en première année de thèse dans l'équipe Sciences des Données au Laboratoire d'informatique de Bourgogne (LIB EA 7534). Sa thèse intitulée « détection des signaux faibles dans les réseaux sociaux », est financée par le projet ISITE Cocktail qui vise à créer un observatoire en temps réel des tendances et des signaux faibles circulant dans le discours sur Twitter. Ses recherches abordent plusieurs thématiques scientifiques dont l'analyse et le traitement des données massives issues de Twitter, avec la détection des évènements et leur contextualisation au sein des réseaux multicouches. Elle a obtenu les diplômes de Licence et de Master en informatique au Liban. Son expérience professionnelle inclut l'analyse et le traitement des données massives.

Landy Andriamampianina a reçu son diplôme de Master MIAGE parcours Ingénierie des Processus Métiers de l'Université Toulouse 1 Capitole en 2019. Elle travaille actuellement en tant que doctorante en CIFRE dans l'équipe SIG de l'Institut de Recherche en Informatique de Toulouse (IRIT) et l'entreprise Activus Group, sous l'encadrement de Franck Ravat, Nathalie Vallés-Parlangeau et Jiefu Song. Ses recherches se concentrent sur la modélisation et la manipulation de graphes temporels.

Cécile Cayèré est doctorante en première année au laboratoire d'informatique L3i de La Rochelle Université. Après avoir terminé son stage de fin d'études pour le projet FEDER TCVPyr – dont l'objectif était de valoriser le patrimoine pyrénéen – et obtenu son Master Technologies de l'Internet à l'Université de Pau et des Pays de l'Adour, elle a commencé une thèse dans le cadre du projet régional Nouvelle-Aquitaine DA3T (*i.e.* Dispositif d'Analyse des Traces numériques pour la valorisation des Territoires Touristiques). L'objectif de cette thèse est de concevoir des méthodes et développer des outils de traitement permettant de faciliter l'analyse des données de mobilité touristique.

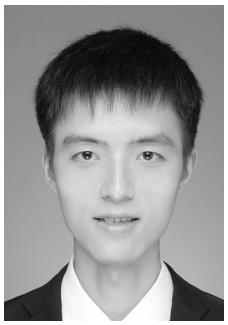
Yohann Chasseray est un doctorant de deuxième année dans le département procédés et systèmes industriels au Laboratoire de Génie Chimique (LGC) de l'École Nationale Supérieure des Ingénieurs en Arts Chimiques et Technologiques (INP-ENSIACET). Il a également reçu en 2018 le diplôme d'ingénieur en génie des systèmes d'information de l'Institut Mines Telecom Albi-Carmaux (IMT Albi-Carmaux). Au travers de sa thèse, il s'intéresse aux challenges liés à la population d'ontologies, incluant notamment des problématiques liées à la variété des sources disponibles et au besoin de généricité.

Nabila Guennouni a complété ses trois premières années de cursus universitaire au Maroc à l'Université Hassan II de Casablanca (UH2 C), puis elle a décroché son master à l'École Nationale Supérieure d'Informatique et d'Analyse des Systèmes de Rabat (ENSIAS). elle continue actuellement son cursus de doctorat à l'université de Pau et Pays de l'Adour. Sa thèse s'intègre dans le cadre du projet de recherche BISE2 (*Business Information System for Energy and Environnement*), et porte sur l'interprétation des événements dans un système d'information hétérogène composé de données de réseaux de capteurs et de corpus de document.

Halima Ramdani est une doctorante CIFRE en intelligence artificielle dans le R&D lab de l'entreprise Xtramile ainsi qu'aux laboratoires de recherche ERPI (Équipe de Recherche sur les Processus Innovatifs) et LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications) à Nancy. Elle travaille sous la supervision de Davy Monticolo et Éric Bonjour du laboratoire ERPI et Armelle Brun du laboratoire LORIA.

Sa thèse s'intitule « un système intelligent pour l'optimisation du e-recrutement » et est spécialisée sur les systèmes de recommandation et apprentissage automatique appliqués aux ressources humaines.

Après des classes préparatoires au Lycée Descartes, elle se spécialise au sein de l'école d'ingénieurs Télécom Nancy, notamment dans le cadre d'un Master des processus de Big Data et de Machine Learning.



Qing Tang est doctorant à l'Université de Technologie de Compiègne (Sorbonne universités), France. En tant que membre du laboratoire HEUDIASYC, ses intérêts de recherche sont l'informatique, les environnements d'apprentissage collaboratif, les traces d'apprentissage et les modèles d'apprenants. Les encadrants de Qing Tang sont Pr. Marie-Hélène Abel du laboratoire HEUDIASYC et Dr. Elsa Negre du laboratoire LAMSADE (Université Paris-Dauphine).

Résumés des articles

Les résumés des articles présents dans les actes du 10^e Forum jeunes chercheuses jeunes chercheurs sont exposés dans les pages qui suivent.

Analyse de la structure latente des réseaux sociaux par graphlets

Hiba Abou Jamra

Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté - 9, Avenue Alain Savary
F-21078 Dijon - France
Hiba_Abou-Jamra@etu.u-bourgogne.fr

RÉSUMÉ. *L'exploitation des données des réseaux sociaux peut révéler des structures latentes qui peuvent être des précurseurs de changements de la structure du réseau ou de phénomènes de diffusion importants. Contrairement aux phénomènes de viralité qui atteignent rapidement des niveaux de diffusion importants, les signaux faibles ne sont pas détectables facilement par des outils statistiques simples. Dans cet article, nous présentons une approche pour détecter des signaux faibles par énumération de graphlets. À partir du cas concret de l'incendie de l'usine Lubrizol et des tweets qui en ont découlé, nous construisons une série temporelle extraite à partir des données Twitter, nous analysons les intervalles précédents et pendant un évènement significatif en terme de vitesse de croissance/décroissance du nombre de graphlets.*

ABSTRACT. *The exploitation of data from social networks reveals latent structures which can be precursors to changes in the structure of the network or to important diffusion phenomena. Unlike virality phenomena which reach quickly important diffusion levels, weak signals are not easily detectable by simple statistical tools. In this article, we present an approach to detect weak signals by enumerating graphlets. Based on the concrete case of the Lubrizol factory fire and the tweets that ensued, we build a time series extracted from Twitter data, we analyze the previous intervals, during a significant event in terms of rate growth / decrease in the number of graphlets.*

MOTS-CLÉS : *Signaux faibles, Graphlets, Structure des réseaux, Twitter, Détection d'évènements.*

KEYWORDS: *Weak signals, Graphlets, Network structure, Twitter, Event detection.*

ENCADREMENT : *Marinette Savonnet et Éric Leclercq*

→ Article présent dans les actes à la page 5.

Toward a generic approach to capture the temporal evolution in graphs

Landy Andriamampianina [12]

1. *Institut de Recherche en Informatique de Toulouse (IRIT) - Université Toulouse 1 Capitole (UT1)*
   2 Rue du Doyen-Gabriel-Marty 31042 Toulouse
   landy.andria@irit.fr

2. *Activus Group*
   1 Chemin du Pigeonnier de la Cépière, 31100 Toulouse
   landy.andriamampianina@activus-group.fr

ABSTRACT. *Graphs have been widely used to represent interconnected entities in real-word applications. However, their representation is static and does not consider the changes that can occur over time. To accommodate to the time dimension, temporal graphs allow to manage different evolution types in order to understand the past, the present and the future of graph-based applications. Nevertheless, temporal graphs impose a set of challenges that are not being sufficiently addressed by the current works. This paper introduces a state of art of existing studies on temporal graphs and our proposition to address their limits.*

RÉSUMÉ. *Les graphes ont été largement utilisés pour représenter des entités interconnectées du monde réel. Cependant, leur représentation est statique et ne prend pas en compte les changements qui peuvent survenir au cours du temps. Pour s'adapter à la dimension temporelle, les graphes temporels permettent de gérer différents types d'évolution pour comprendre aussi bien le passé, le présent que le futur des applications basées sur les graphes. Néanmoins, les graphes temporels imposent un ensemble de défis qui ne sont pas suffisamment relevés par les travaux en cours. Cet article présente un état de l'art des études existantes sur les graphes temporels et notre proposition pour aborder leurs limites.*

KEYWORDS: *Temporal graphs, historization mechanism, evolution types.*

MOTS-CLÉS : *Graphes temporels, mécanisme d'historisation, types d'évolution.*

ENCADREMENT : *Franck Ravat*

→ Article présent dans les actes à la page 9.

Plateforme ETL dédiée à l'analyse de la mobilité touristique dans une ville

Cécile Cayèré

La Rochelle Université,
23 Avenue Albert Einstein,
17000 La Rochelle, France
cecile.cayere1@univ-lr.fr

RÉSUMÉ. *Dans un monde où la plupart de nos déplacements sont enregistrés, nous faisons l'hypothèse que les traces spatio-temporelles numériques provenant de touristes peuvent aider les collectivités et les professionnels du tourisme à gérer et à valoriser le territoire touristique. Nous visons à concevoir une plateforme modulaire dédiée à des utilisateurs non-informaticiens permettant de concevoir des chaînes de traitement personnalisé pour faciliter l'analyse des données de mobilité. Nous cherchons à extraire de la sémantique à partir de données issues de capteur, confronter ces données brutes avec des données issues d'enquêtes et d'analyse de l'environnement et comprendre les comportements de mobilité touristique.*

ABSTRACT. *In a world where most of our movements are recorded, we assume that digital spatio-temporal traces from tourists can help communities and tourism professionals to manage and enhance the tourism territory. We aim to design a modular platform dedicated to non-computer users allowing the design of customized pipelines to facilitate the analysis of mobility data. We want to extract semantics from sensor data, to compare this raw data with data from surveys and environmental analysis and to understand tourist mobility behaviours.*

MOTS-CLÉS : *Données spatio-temporelles, trajectoire sémantique, plateforme modulaire, tourisme.*

KEYWORDS: *Spatio-temporal data, semantic trajectory, modular platform, tourism.*

ENCADREMENT : *Cyril Faucher, Christian Sallaberry, Marie-Noëlle Bessagnet et Philippe Roose*

→ Article présent dans les actes à la page 13.

Un méta-modèle pour la population d'ontologie indépendamment du domaine

Yohann Chasseray

Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France
4 allée Émile Monso, 31030 Toulouse, France
yohann.chasseray@ensiacet.fr

RÉSUMÉ. *L'ajout de l'expertise et des facultés de raisonnement humaines constitue un des prochains défis dans la construction de systèmes experts. En ce sens, la mise en place de bases de connaissances robustes et interprétables devient nécessaire. Les ontologies et le formalisme qu'elles définissent pour la représentation de la connaissance sont autant de ressources utiles pour atteindre ce but. Malheureusement, la majorité des ontologies disponibles contiennent peu de connaissances concrètes et se révèlent difficilement applicables. Comme la population manuelle d'ontologies est une tâche chronophage, le but du travail présenté est d'automatiser cette population en fournissant un cadre méthodologique qui garantisse une généricité relative au format des données qui renferment la connaissance et au domaine décrit par l'ontologie.*

ABSTRACT. *The upcoming challenge in the elaboration of expert systems is the addition of human expertise and reasoning. It is then a priority to build strong computable knowledge bases. Ontologies and their formalisms for knowledge representation constitute an effective resource to achieve this goal. However, most of the existing ontologies remain unpopulated and can not be used in real-life problems. As the manual population of ontologies is hugely time-consuming, the goal of the presented work is to automate the ontology population process, providing a framework that guarantees genericity regarding both the data sources containing knowledge and the domain described in the ontology.*

MOTS-CLÉS : *Bases de connaissances, extraction de connaissances, ingénierie dirigée par les modèles, transformation de modèle, interprétation du langage naturel.*

KEYWORDS: *Knowledge bases, knowledge extraction, model driven engineering, model transformation, natural language processing.*

ENCADREMENT : *Jean-Marc Le Lann, Anne-Marie Barthe-Delanoë et Stéphane Négny*

→ Article présent dans les actes à la page 17.

Interprétation d'événement dans un système d'information hétérogène

Analyse croisée de données issues de capteurs et de corpus documentaires

Nabila Guennouni

Université de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA
2 Allée du Parc Montaury
64600 Anglet
Nabila.Guennouni@univ-pau.fr

RÉSUMÉ. *Dans cet article, nous montrons qu'au sein de la communauté des systèmes d'information, les travaux récents relatifs au 5W1H ("What? Who? Where? When? Why? How?") sont une piste intéressante à exploiter pour offrir un premier guide à l'utilisateur dans l'explication d'un événement issu d'un environnement connecté. Nous démontrons que ce travail qui, au départ, se concentre exclusivement sur les données d'articles de journaux, peut s'appliquer aux données d'un réseau de capteurs. De plus, dans un objectif de combiner données de capteurs et corpus documentaire, nous montrons que des questionnements 5W1H successifs permettent de rapprocher des données hétérogènes à des fins d'explication d'événement, en s'appuyant sur la connaissance de domaines spécifiques associée aux données.*

ABSTRACT. *In this article, we show that within the Information System community, recent researches on 5W1H ("What? Who? Where? When? Why? How?") are an interesting issue in order to provide end-users a preliminary event interpretation in a connected environment. We demonstrate that this work, which initially focuses exclusively on newspaper articles, can be applied to sensor network data in order to explain an event. Moreover, with the aim of combining sensor data and document corpus, we show that successive 5W1H queries can be used to bring together heterogeneous data for the purpose of explaining an event. This link is built dynamically based on the domain-specific knowledge associated with sensor data and documents.*

MOTS-CLÉS : *Évènement, Capteur, Sémantique, 5W1H, Système d'information.*

KEYWORDS: *Event, Sensor, Semantic, 5W1H, Information System.*

ENCADREMENT : *Christian Sallaberry et Sébastien Laborie*

→ Article présent dans les actes à la page 21.

Un système intelligent pour l'optimisation du e-recrutement

Halima Ramdani [1,2]

1. *Equipe de Recherche sur les Processus Innovatifs, Université de Lorraine, Nancy, France*

2. *Laboratoire lorrain de recherche en informatique et ses applications, Université de Lorraine, Nancy, France*

RÉSUMÉ. *Par le passé, les candidats potentiels pour une offre d'emploi se trouvaient dans des lieux physiques que l'on pouvait atteindre à travers les grands médias historiques, souvent fortement ancrés dans leur espace géographique local. Puis, avec l'ère d'Internet, l'e-recrutement est apparu, offrant aux annonceurs une portée géographique, en remplaçant les médias classiques par les médias numériques. Néanmoins ces médias numériques se sont multipliés entraînant ainsi une difficulté de cibler les candidats sur le web. L'objectif de notre projet de recherche est d'optimiser le processus de e-recrutement en concevant un système de recommandation capable de cibler des candidats potentiels à moindre coût.*

ABSTRACT. *In the past, potential candidates for a job offer were in physical locations that could be reached through the major historical media, often strongly rooted in their local geographic space. Today, digital media replaced traditional media, offering advertisers geographic reach. However digital media has multiplied, making it difficult to target candidates on the web. The objective of our research project is to optimize the e-recruitment process by designing a recommendation system capable of targeting potential candidates at a lower cost.*

MOTS-CLÉS : *E-recrutement, système de recommandation, apprentissage automatique.*

KEYWORDS: *E-recruitment, recommander system, machine learning.*

ENCADREMENT : *Armelle Brun, Eric Bonjour et Davy Monticolo*

→ Article présent dans les actes à la page 25.

Towards pedagogical resources recommender system based on collaboration's tracks

Qing Tang

Sorbonne universités, Université de technologie de Compiègne, CNRS UMR 7253, HEUDIASYC, Compiègne, France

qing.tang@utc.fr

RÉSUMÉ. *Un système de recommandation dans le domaine de l'apprentissage en ligne réduit la difficulté des apprenants à sélectionner des ressources pertinentes face à la grande quantité de ressources à leur disposition. Améliorer cette pertinence reste un défi. Afin de proposer des recommandations pertinentes, le système a besoin de bien connaître l'apprenant. Malheureusement, la plupart des apprenants peuvent rarement fournir des informations complémentaires afin de répondre aux besoins du système de recommandation. Nous proposons d'utiliser xAPI, reconnu comme un standard, pour aider à enregistrer les traces d'apprentissage générées par les apprenants dans un environnement d'apprentissage collaboratif ; extraire des informations sur les apprenants et les ajouter aux modèles d'apprenants; puis construire un système de recommandation de ressources pertinentes.*

ABSTRACT. *Learning recommender system reduces learners' burden of selecting when they face massive online learning resources, but how to improve the accuracy is still a tough challenge. In order to make targeted and efficient recommendation, recommender system needs comprehensive understanding about learner. But most learners cannot make additional actions to supply their information to fulfill the need of recommender system, so how to obtain the learner's status information has become a new challenge. We propose to use the network tracking standard xAPI to help record learning tracks generated by learners in a collaborative learning environment, to extract status information and add to learner models as pre-steps, then build recommender system to help learners improve their learning efficiency by returning resources.*

MOTS-CLÉS : *Apprentissage en ligne, Environnement d'apprentissage collaboratif, Système de recommandation, Traces, Modèle des apprenants.*

KEYWORDS: *Online-learning, Collaborative learning environment, Recommender system, Learner tracks, Learner models.*

ENCADREMENT : *Marie-Hélène Abel et Elsa Negre*

→ Article présent dans les actes à la page 29.

Appel à soumissions

L'appel à soumission pour le 10^{e} Forum jeunes chercheuses jeunes chercheurs est exposé dans les pages qui suivent.

Le tableau ci-dessous présente la chronologie du traitement des soumissions :

	Date
Première diffusion	**09/12/19**
Première relance	08/01/20
Deuxième relance	10/02/20
Troisème relance	05/03/20
Date limite de soumission	**22/03/20**
Extension	31/03/20
Envoi en relecture	01/04/20
Retour des relecteurs	20/04/20
Notification aux auteurs	21/04/20
Version finale	**03/05/20**
Publication des actes	[?]/06/20

INFORSID 2020 JCJC: INFORSID 2020 - Forum Jeunes Chercheuses
Jeunes Chercheurs (JCJC)
Université de Bourgogne
Dijon, France, June 2-4, 2020

Submission link	https://easychair.org/conferences/?conf=inforsid2020jcjc
Submission deadline	March 22, 2020

/* Toutes nos excuses en cas de réceptions multiples.
Merci de transférer cette information à toute personne susceptible d'être intéressée. */

INFORSID 2020, du 2 au 4 juin
Dijon, Université de Bourgogne

Appel à soumissions pour le :

Forum Jeunes Chercheuses Jeunes Chercheurs (JCJC)
https://inforsid2020.sciencesconf.org/

À l'attention des jeunes chercheuses et jeunes chercheurs en première ou deuxième année de doctorat, INFORSID 2020 organise la dixième édition du Forum Jeunes Chercheuses Jeunes Chercheurs !

L'association INFORSID prend en charge les frais d'inscription à la conférence*, l'hébergement en cité universitaire, les déjeuners, ainsi que le repas du gala pour les doctorants auteurs d'un article court retenu.

Les objectifs du Forum JCJC sont :
- de permettre aux jeunes chercheurs en première ou deuxième année de doctorat de présenter leur problématique de recherche,
- d'établir des contacts avec des équipes travaillant sur des domaines connexes,
- d'offrir un aperçu des axes de recherche actuels et ainsi élargir le champ des connaissances des jeunes chercheurs.

Les auteurs d'un article court retenu seront invités à :
- exposer synthétiquement leur travail en 180 secondes au cours d'une session spécifique du congrès,
- préparer un poster qui sera affiché pendant le congrès.

Les meilleures contributions seront réunies au sein d'un article publié avec les meilleurs articles sélectionnés d'INFORSID 2020 dans un numéro spécial de la revue ISI (Ingénierie des Systèmes d'Information).

Soumission

Le Forum est ouvert à la présentation de travaux de recherche originaux, de développements industriels et d'expériences significatives dans le domaine des systèmes d'information qui sont réalisés par des jeunes chercheuses et jeunes chercheurs en première ou deuxième année de doctorat.

Les articles courts (4 pages) devront se conformer au modèle Hermès-Lavoisier (feuille de style word[1], feuille de style LaTeX[2]). La langue officielle du congrès est le français, toutefois le Forum est ouvert aux contributions de langue anglaise.

Les soumissions devront être déposées avant le **22/03/20** via easychair[3].

Dates importantes

Date limite de soumission des articles :	**22/03/20**
Notification aux auteurs :	22/04/20
Réception des textes définitifs :	03/05/20
Congrès INFORSID 2020 :	02-04/06/20

Contact

Pierre-Emmanuel Arduin
PSL - Université Paris-Dauphine, Laboratoire DRM UMR CNRS 7088
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France
Tél. : +33.(0)1.44.05.48.19 - Fax : +33.(0)1.44.05.40.23
pierre-emmanuel.arduin@dauphine.psl.eu
http://arduinpe.free.fr/

P A R T I C I P E Z E T F A I T E S P A R T I C I P E R ! !

* : Les doctorants également auteurs (ou co-auteurs) d'un article accepté à la conférence INFORSID 2020 devront néanmoins régler les frais d'inscription à la conférence au titre de cet article.

[1] : http://arduinpe.free.fr/ForumJCJC/feuilledeStyle_ARL.doc
[2] : http://arduinpe.free.fr/ForumJCJC/Hermes-Journal_V4_2014.zip
[3] : https://easychair.org/conferences/?conf=inforsid2020jcjc