EGC&IA 2016

Cette deuxième journée EXTRACTION ET GESTION DES CONNAISSANCES et INTELLIGENCE ARTIFICIELLE réunit les deux communautés autour du thème des « Données Participatives et Sociales ». Ces données sont au cœur de nouveaux défis tant au niveau de la fouille de données que de l'intelligence artificielle. Les travaux de la littérature sont généralement associés à l'une des deux communautés, sans montrer le lien entre elles. Cet atelier cherche particulièrement à focaliser sur ce lien du point de vue représentation qu'analyse.

Date et Lieu

- Date 19 janvier 2016
- Lieu : IUT de Reims-Chalons-Charleville, Chemin des Rouliers, 51100 Reims

Programme

- 08h45 Ouverture et mot d'introduction par Arnaud Martin *(Université Rennes 1)* et Engelbert Mephu Nguifo *(Université Clermont-Ferrand 2)*.
- 09h00 « A la recherche des mini-publics : un problème de communautés, de singularités et de sémantique » par Eric Leclercq *(Université de Bourgogne, Dijon)*, Sergey Kirgizov et Maximilien Danisch.
- 09h30 « Analyse d'activité et exposition de la vie privée sur les médias sociaux » par Younes Abid *(Inria, Nancy)*, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Marc Rigolot, et Michaël Rusinowitch.
- 10h00 « Étude de la perception de l'usage de matériaux composites pour les véhicules du futur : Détection de communautés » par Marouane Hachicha *(Université de Nantes)*, Nadine Cullot, Eric Leclercq, Philippe Castel, Marie-Françoise Lacassagne et Stéphane Fontaine.
- 10h30 Pause-café
- 11h00 « Analyse des images qui circulent sur Internet : un aperçu du projet ImagiWeb » par Julien Velcin *(Université Lyon 2)*.
- 11h30 « Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social » par Coriane Nana Jipmo *(Centrale Supélec, Paris)*, Gianluca Quercini et Nacéra Bennacer.
- 12h00 « Un système personnalisé de recommandation basé sur le profil des utilisateurs dans les folksonomies » par Mohamed Nader Jelassi *(Université Clermont-Ferrand 2),* Sadok Ben Yahia, et Engelbert Mephu Nguifo.
- 12h30 Pause-repas
- 14h00 Présentations de l'AFIA par Yves Demazeau *(Président de l'AFIA)* et Fabrice Guillet *(Président de EGC)*.
- 14h30 « Fabrique logicielle de réseaux sociaux spécialisés : Aspects fonctionnels » par David Fernandez *(Université Montpellier 2)*, Benjamin Billet et Didier Parigot.
- 15h00 « Worker-Centricity Could Be Today's Disruptive Innovation in Crowdsourcing » par Sihem Amer-Yahia *(CNRS Grenoble)* (conférence invitée).
- 16h00 Pause-café
- 16h30 « Connaissance et collaboration » par Jérôme Euzenat *(Inria Grenoble)* (conférence invitée).
- 17h30 Discussion animée par Arnaud Martin *(Université Rennes 1)* et Engelbert Mephu Nguifo *(Université Clermont-Ferrand 2)*.
- 18h00 Clôture.

Inscriptions

L'atelier se tient dans le cadre de la conférence EGC 2016. Les inscriptions se font sur le site de la conférence (http://egc2016.univ-reims.fr/index.php/Inscription). .

Worker-Centricity Could Be Today's Disruptive Innovation in Crowdsourcing

Sihem Amer-Yahia*

*LIG-CNRS, Grenoble

Summary

Organizational studies have been focusing on understanding human factors that influence the ability of an individual to perform a task, or a set of tasks, alone, or in collaboration with others, for over 40 years. The reason crowdsourcing platforms have been so successful is that tasks are small and simple, and do not require a long engagement from workers. The crowd is typically volatile, its arrival and departure asynchronous, and its levels of attention and accuracy diverse. Today, crowdsourcing platforms have plateaued and, despite a high demand, they are not adequate for emerging applications such as citizen science and disaster management. I will argue that workers need to be brought back into the loop by enabling worker-centric crowdsourcing. My current research seeks to verify how human factors such as skills, expected wage and motivation, contribute to making crowdsourcing kick-off again. In particular, I will discuss team formation for collaborative tasks, adaptive task assignment, and task composition to help workers find useful tasks.

This is joint work with Senjuti Basu Roy from the University of Washington and Dongwon Lee, Penn State University.

Bio : Sihem Amer-Yahia is DR1 CNRS at LIG in Grenoble where she leads the SLIDE team. Her interests are at the intersection of large-scale data management and data analytics. Before joining CNRS, she was Principal Scientist at QCRI, Senior Scientist at Yahoo! Research and at&t Labs. Sihem has served on the SIGMOD Executive Board, the VLDB Endowment, and the EDBT Board. She is the Editor-in-Chief of the VLDB Journal for Europe and Africa and is on the editorial boards of TODS and the Information Systems Journal. She was PC chair of BDA 2015 and SIGMOD Industrial 2015 and is currently chairing VLDB Workshops 2016. Sihem received her Ph.D. in CS from Paris-Orsay and INRIA in 1999, and her Diplôme d'Ingénieur from INI, Algeria.

À la recherche des *mini-publics* : un problème de communautés, de singularités et de sémantique

Éric Leclercq*, Sergey Kirgizov*, Maximilien Danisch**

*LE2I CNRS UMR 6306 - Université de Bourgogne Franche-Comté
Eric.Leclercq@u-bourgogne.Fr, Sergey.krigizov@u-bourgogne.fr
** Télécom ParisTech

Résumé. La notion de communauté est largement traitée dans l'analyse des réseaux complexes. Cet article propose de montrer comment une approche pluridisciplinaire revisite la notion de communautés en la connectant à une question de recherche en sciences de l'information et de la communication.

1 Introduction et contexte

Le travail que nous décrivons s'inscrit dans le projet international et pluri-disciplinaire TEE 2014 qui vise à étudier la structure de la communication politique sur Twitter durant les élections européennes de 2014, dans 6 pays (France, Allemagne, Italie, Espagne, Belgique, Royaume-Uni). Le corpus généré comporte plus de 50M de tweets sur la période de la campagne des élections. Les données issues des tweets, par les opérateurs (reply, retweet RT, mention @, hashtag # et URL) prennent la forme d'un graphe hétérogène (réseau complexe multi-relationnel) de 50Go environ. Parmi les questions scientifiques abordées par les chercheurs en sciences de la communication (SC), la recherche de communautés (de petite taille), d'utilisateurs échangeant autour de hashtags spécifiques est un enjeu pour la compréhension des formes de communication. À partir des travaux de Goodin et Dryzek (2006), les chercheurs en SC impliqués dans le projet ont proposé une première définition : « Mini-Publics are defined as groups small enough to be genuinely deliberative. Mini publics serve as smaller circles of (better) informed groups, which engage in important information exchange processes and discourses ».

Cette définition ne permet pas directement d'établir un lien avec les algorithmes existant. Une analyse exploratoire a été conduite, en appliquant les principaux algorithmes de détection de communautés, mais aucun n'a permis de faire apparaître de manière pertinente les petites communautés recherchées. Seul l'algorithme Walktrap de Pons et Latapy (2005) appliqué sur le réseau multi-relationnel (RT, #, @), associé à une visualisation enrichie avec la connaissance du domaine a mis en évidence des éléments singuliers. En effet, disposant d'une ontologie de domaine constituée par la description des partis politiques, des candidats et de règles logiques, nous avons découvert quelques candidats singuliers, appartenant à un parti mais classés dans une communauté qui rassemble exclusivement des candidats d'un autre parti. À partir de ces premiers éléments et avec l'expérience de l'étude des tweets, la définition des mini-publics a été précisée.

2 Modèle et algorithme pour la détection des mini-publics

L'équipe Media-studies de l'Université de Bonn a proposé une définition empirique des mini-publics comportant 10 critères que nous avons partiellement traduit dans un modèle exécutable. Les critères mesurables retenus sont relatifs à la structure du graphe : utilisation intensive de hashtags spécifiques, des opérateurs retweets et reply par un ensemble d'utilisateurs *a priori* non borné. Un mini-public correspond donc à une communauté ego-centrée également désignée par communauté locale à un nœud, Bagrow et Bollt (2005), Danisch et al. (2013). Le critère d'intensité relève à la fois de la temporalité et de la structure. Il est évalué lors de la constitution des jeux de données sous la forme d'un facteur de vieillissement (par exemple proportionnel à la fréquence d'utilisation d'un hashtag par un utilisateur) appliqué sur le poids des liens.

Pour détecter les mini-publics, nous utilisons comme point de départ l'algorithme proposé par Danisch et al. (2013). À partir d'un ensemble de nœuds d'intérêt, une valeur est propagée aux autres nœuds du graphe afin d'en mesurer la proximité. Ainsi, chaque nœud possède une valeur qu'il transmet à ses voisins proportionnellement aux poids des liens. À l'issue d'itérations de ce processus, l'ensemble des valeurs des nœuds doit être stable pour établir un ranking de tous les nœuds du graphe par rapport aux nœuds d'intérêt. L'étude de l'allure de la courbe représentant la proximité en fonction du rang fait apparaître des plateaux qui correspondent à des structures communautaires. La convergence de l'algorithme a été prouvée en utilisant la théorie des chaînes de Markov.

3 Résultats, discussion et conclusion

Avec les chercheurs en SC, nous avons déterminé un ensemble d'expériences, dont quelques résultats sont visibles à l'adresse (`http://eric-leclercq.fr/minipublics/`). Nous avons appliqué l'algorithme sur des graphes user/hashtag non dirigés (pondérés), sur des graphes dirigés (user/hashtag/mention/RT) pondérés avec facteur de vieillissement pour des nœuds d'intérêt spécifiques. Les mini-publics obtenus, soumis à l'interprétation des chercheurs en SC, ont validé globalement l'algorithme et montré sa capacité à détecter des mini-publics pertinents, imbriqués dans certains cas, ainsi que les frontières entre plusieurs mini-public. Les travaux actuels se concentrent : 1) sur la modélisation de la notion de conversation de manière à mieux prendre en compte la définition originale 2) sur la modélisation des relations complexes par des hypergraphes.

Références

Bagrow, J. et E. Bollt (2005). A local method for detecting communities. *Phys. Rev.*.

Danisch, M., J.-L. Guillaume, et B. Le Grand (2013). Towards multi-ego-centred communities : a node similarity approach. *Int. Journal of Web Based Communities 9*(3), 299–322.

Goodin, R. et J. S. Dryzek (2006). Deliberative impacts : the macro-political uptake of mini-publics. *Politics & Society 34*, 219–244.

Pons, P. et M. Latapy (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pp. 284–293. Springer.

Analyse d'activité et exposition de la vie privée sur les médias sociaux

Younes Abid*, Abdessamad Imine*, Amedeo Napoli*,
Chedy Raïssi*, Marc Rigolot**, Michaël Rusinowitch *

*INRIA-Nancy, 54600 Villers-lès-Nancy
prenom.nom@inria.fr
** Fondation MAIF, 50 avenue Salvador Allende , 79000 Niort
prenom.nom@fondation.maif.fr

Résumé. L'anonymat sur les réseaux sociaux ne supprime pas les risques sur la vie privée des utilisateurs découlant du recoupement des informations personnelles publiées par ceux-ci ou par leurs relations en ligne. Dans cette optique, nous avons mené une enquête par questionnaire pour mesurer la sensibilité des données personnelles publiées sur les médias sociaux et analysé les pratiques des utilisateurs. Nous montrons ainsi que plus de 76 % des internautes sondés sont vulnérables aux attaques de révélation d'identité et d'inférence d'information sensible. L'étude est complétée par la description d'une procédure automatique qui montre que ces vulnérabilités sont simples à exploiter en pratique et doivent donc être prises en compte dans un système de protection. Ce travail est réalisé dans le cadre d'un projet financé par la Fondation MAIF.[1].

Summary

Anonymous use of social networks do not prevent users from privacy risks arising from infering and cross-checking information published by themselves or their relationhips. With this in mind we have conducted a survey in order to measure sensitiveness of personal data published on social media and to analyze the users behaviors. We have shown that 76 % of internet users that have answered the survey are vulnerable to identity or sensitive data disclosure. Our study is completed by the description of an automatic procedure that shows how easily these vulnerabilities can be exploited and motivates the need for more advanced protection mechanisms.

[1] http://www.fondation-maif.fr/

Étude de la perception de l'usage de matériaux composites pour les véhicules du futur : Détection de communautés

Marouane Hachicha*, Nadine Cullot*, Éric Leclercq*
Philippe Castel**, Marie-Françoise Lacassagne**, Stéphane Fontaine***

*LE2I UMR6306, CNRS, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France
**SPMS EA4180, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France
***DRIVE EA1859, Univ. Bourgogne Franche-Comté, F-58000 Nevers, France
prénom.nom@u-bourgogne.fr

1 Introduction : contexte et objectifs

Le travail proposé s'inscrit dans le cadre d'un projet interdisciplinaire mené par des psychosociologues, des informaticiens et des chercheurs spécialistes en ingénierie des véhicules pour l'environnement (matériaux, énergie, communications), pour l'étude de la perception par des usagers, de l'utilisation de matériaux composites (Belaid et al., 2016) pour la construction de véhicules du futur. Le projet dans son ensemble concerne l'étude de groupes de populations pour permettre le repérage de minorités actives dans la promulgation des véhicules du futur ; ces minorités étant vues comme des groupes composés de membres partageant la même identification sociale et jouant le même rôle (Castel et Lacassagne, 2015). Une première étude comparative a été menée visant à confronter les notions centrales d'identification de communautés développées par les psychosociologues, à savoir l'identité catégorielle définie par le sentiment d'appartenance à un groupe objectivement défini et la découverte dynamique de communautés virtuelles proposées par les informaticiens. Le travail proposé dans cet article s'inscrit dans le cadre de cette étude et concerne la détection des communautés par un algorithme spécifique adapté et adaptable au contexte de l'étude.

Données collectées : Le travail réalisé s'appuie sur des données collectées via un questionnaire REPMUT [1]. Dans un premier temps, les personnes interrogées se définissent comme faisant partie d'un groupe que l'on peut qualifier d'*écomobilistes* c'est-à-dire comme des personnes prêtes à acquérir une voiture avec des matériaux composites ou d'un groupe d'*automobilistes* préférant les voitures classiques. Dans un deuxième temps, les personnes caractérisent "les personnes de leur groupe", "de l'autre groupe" et "les voitures à matériaux composites" à l'aide de cinq adjectifs libres pour chaque caractérisation et elles accordent à chaque adjectif une valeur permettant de lui donner un "poids" par rapport à certains critères (par exemple, indiquer à quel point l'adjectif donné pour caractériser les automobilistes est lié à la reconnaissance sociale, ou à quel point les automobilistes possèdent-ils cette caractéristique, etc.). Ces valeurs numériques liées à la valence, au statut, à l'homogénéité, à la typicalité et l'entitativité, ont été traitées pour dégager cinq indicateurs caractérisant les relations entre les groupes

1. www.repmut.com

de personnes interrogées (écomobilistes et automobilistes). Ce sont ces indicateurs qui sont utilisés dans le travail proposé.

2 Méthodologie de détection des communautés

Comme décrit précédemment, après le traitement des données collectées, chaque individu est représenté par les cinq indicateurs numériques qui le caractérisent. Ces valeurs ne sont pas directement prises en compte mais considérées selon trois classes `N` (négative), `P` (positive), ou `Z` (zéro). Chaque individu est donc identifié par une chaîne ordonnée qui peut être de la forme `PNZPN`, par exemple.

L'algorithme de détection de communautés proposé considère en entrée les individus et les chaînes de caractères représentatives de leurs caractérisations. L'algorithme s'appuie sur l'utilisation de la distance de Hamming (Steane, 1996) comme mesure de similarité (Fortunato, 2010) des caractéristiques de deux individus et met en œuvre un ensemble de règles spécifiques pour construire les communautés. Initialement, les individus identiques, c'est-à-dire avec une distance de Hamming égale à 0, sont regroupés dans des clusters appelés "clusters singletons" qui représentent les communautés initiales. Puis chaque cluster singleton est traité et peut être regroupé dans une communauté avec d'autres clusters singletons en application des deux règles suivantes : (1) le cluster singleton en cours de traitement rejoint la communauté qui contient le maximum de clusters singletons qui lui sont les plus proches avec la distance de Hamming, mais (2) il ne rejoint jamais une communauté comportant au moins un cluster singleton complètement différent (c'est-à-dire avec une distance de Hamming maximale).

Résultats et perspectives : L'algorithme a été appliqué avec, en entrée, les données de 109 individus (76 écomobilistes et 33 automobilistes) donnant naissance à 52 clusters singletons. En sortie, il produit 8 communautés qui regroupent de façon assez significative des individus ayant des caractérisations plutôt négatives comme `NNNNN`, `NNNPN` et `NNNPP` ou plutôt positives comme `PPNPP`, `PPPPP` et `PPPPZ` ou neutres comme `NNPZP` mais également quelques communautés plus partagées. Il est à noter que ces communautés comportent à la fois des écomobilistes et des automobilistes. Les résultats de l'algorithme sont en cours d'analyse de façon plus précise par les psychosociologues et des améliorations sont en cours de discussion pour affiner les règles de l'algorithme et les prétraitements des données collectées. Après cette étape de détection des communautés, une recherche de minorités pertinentes à l'intérieur de ces communautés est envisagée.

Références

Belaid, M., S. Fontaine, A. El-Hafidi, B. Piezel, et B. Gning (2016). Prediction of dissipative properties of flax fibers reinforced laminates by vibration analysis. *Applied Mechanics and Materials 822*, 411–417.

Castel, P. et M.-F. Lacassagne (2015). Theory of social partitions and identity dynamics. *B. Mohan (Ed.), Construction of Social Psychology*.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3), 75–174.

Steane, A. M. (1996). Error correcting codes in quantum theory. *Phys. Rev. Lett. 77*(5), 793.

Analyse des images qui circulent sur Internet : un aperçu du projet ImagiWeb

Julien Velcin*

*Université de Lyon (ERIC, Lyon 2)
julien.velcin@univ-lyon2.fr,
http://mediamining.univ-lyon2.fr/velcin/

Résumé. Dans cette présentation, je donne un aperçu du projet ImagiWeb. L'objectif du projet consiste à capturer l'image d'une entité, au sens de sa représentation, qui circule sur Internet et dans les médias sociaux. Le projet a mobilisé six équipes de chercheurs entre informatique et sciences sociales (ERIC à Lyon, LIA à Avignon, CEPEL à Montpellier, Xerox à Grenoble, EDF à Paris et AMI Software à Montpellier) pendant trois ans et demi.

1 Introduction

L'image de nombreuses entités (par ex. célébrités, entreprises, marques) nous parvient principalement par l'intermédiaire de l'existence virtuelle qu'elles mènent sur le Web et dans les nouveaux médias. L'objectif du projet ImagiWeb est d'analyser l'opinion exprimée dans les messages postés sur Internet au sujet de ces entités, à l'aide de techniques informatiques et statistiques, et de la relier aux caractéristiques sociales des individus qui les ont produits en suivant une logique de panélisation.

A cette approche résolument pluridisciplinaire s'ajoute la volonté d'apprécier l'opinion en regard de cibles qui décrivent l'entité (par ex. : les soutiens de l'homme politique ou la politique tarifaire de l'entreprise) et de la suivre de manière dynamique. Pour cela, il est nécessaire d'aller au-delà des méthodes habituelles de classification d'opinion (Liu, 2015). Ce type de travaux peut avoir un impact aussi bien technique (développer des nouveaux algorithmes et logiciels), stratégique (apporter une nouvelle solution pour la gestion de la réputation en ligne) que sociétal (mieux comprendre la naissance et la diffusion des représentations).

2 Méthodes et technologies utilisées

Pour résoudre cette problématique, après avoir mis en place un cadre complet d'acquisition et d'annotation des données (Velcin et al., 2014), nous avons choisi de combiner des outils avancés de traitement automatique de la langue (approche linguistique), de fouille de textes (approche statistique) et d'apprentissage automatique (supervisé et non supervisé).

La prédiction des cibles et des polarités d'opinion est obtenue à l'aide d'une méthode hybride et active de classification supervisée (Stavrianou et al., 2014; Cossu et al., 2015) tandis

que le regroupement des groupes d'opinion homogène est construit à l'aide de clustering probabiliste évolutionnaire (Kim et al., 2015).

Les producteurs des messages analysés sont identifiés à l'aide d'une stratégie originale de panélisation des internautes mise en place en modernisant l'approche traditionnellement employée en sociologie (Dormagen et al., 2014).

Enfin, ces outils sont intégrés dans un prototype permettant de démontrer l'intérêt de l'approche sur deux cas d'étude et selon plusieurs scénarios d'usage envisagés, tels que la navigation dans les données et les annotations ou la visualisation temporelle des groupes d'opinion (Khouas et al., 2015).

3 Résultats obtenus dans le projet

Le projet a principalement permis de montrer qu'il était possible de capturer les opinions fines au sujet d'une entité en utilisant des outils d'analyse automatique des messages d'expression sur Internet.

Sur le cas des hommes politiques, il a été possible de comparer cette opinion en ligne aux baromètres habituels de l'opinion et d'en tirer des conclusions sur certaines convergences observées mais surtout sur des différences très marquées (Velcin et Boyadjian, 2016).

Sur le cas de l'entreprise EDF, le prototype mis en place a permis de confirmer des conclusions tirées par les sémiologues, et ce de manière exhaustive, mais également de proposer des informations nouvelles.

D'un point de vue industriel, ce projet a permis l'élaboration d'une méthode générale pour étudier l'image de marque (réputation) sur Internet. Celle-ci a été d'ores et déjà intégrée à la plateforme de veille éditée par l'entreprise partenaire.

4 Conclusion

Après trois ans et demi, le projet ImagiWeb a permis de montrer qu'il était possible de capturer l'opinion des groupes d'individus et, ce, à un degré fin d'analyse. Contrairement à l'idée reçue sur l'anonymat des internautes, une logique de panélisation est possible afin d'identifier les producteurs d'opinion mais il ne faut pas se tromper sur la valeur de représentativité des sources d'information, telle que Twitter. Bien sûr, il s'agit d'être vigilant sur les aspects liés à la vie privée des internautes en intégrant des mécanismes d'anonymisation.

L'image des entités peut être ainsi étudiée via l'utilisation de logiciels, comme nous l'avons montré avec le prototype de démonstration sur plusieurs scénarios d'analyse. L'évaluation par un sémiologue de l'apport de ce type d'outil affiche clairement ses avantages (navigation facilitée dans les données, résumé d'un vaste corpus de documents, capacité d'innovation) ainsi que des pistes d'amélioration (par exemple la gestion dynamique des cibles).

Références

Cossu, J.-V., E. SanJuan, J.-M. Torres-Moreno, et M. El-Bèze (2015). Multi-dimensional reputation modeling using micro-blog contents. In *Proceedings of the 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS)*, pp. 452–457. Springer.

Dormagen, J.-Y., J. Boyadjian, et M. Neihouser (2014). Hybrid method for measuring opinion. In *Proceedings of Asia Conference for e-Democracy and Open Government (CeDEM)*, Hong-Kong.

Khouas, L., C. Brun, A. Peradotto, J.-V. Cossu, J. Boyadjian, et J. Velcin (2015). Étude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le web social. In *Actes de la 22ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Caen.

Kim, Y.-M., J. Velcin, S. Bonnevay, et M.-A. Rizoiu (2015). Temporal multinomial mixture for instance-oriented evolutionary clustering. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pp. 593–604. Springer.

Liu, B. (2015). *Sentiment Analysis : Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Stavrianou, A., C. Brun, T. Silander, et C. Roux (2014). NLP-based feature extraction for automated tweet classification. In *Workshop on Interactions between Data Mining and Natural Language (DMNLP), in collocation with ECML-PKDD*, Nancy, France, pp. 15–19.

Velcin, J. et J. Boyadjian (2016). De l' "opinion mining" à la sociologie des opinions en ligne. pour une approche interdisciplinaire de l'étude du web politique. *Question de communication*. Article en cours d'évaluation.

Velcin, J., Y. Kim, C. Brun, J. Dormagen, E. SanJuan, L. Khouas, A. Peradotto, S. Bonnevay, C. Roux, J. Boyadjian, A. Molina, et M. Neihouser (2014). Investigating the image of entities in social media : Dataset design and first results. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp. 818–822.

Summary

In this talk, I will give an overview of the ImagiWeb project. The objectif of this project consists in capturing an entity's image (that is, its representation) that circulate through Internet and the social media. The project involved six research teams inbetween computer science and social sciences and humanities (ERIC at Lyon, LIA at Avignon, CEPEL at Montpellier, Xerox at Grenoble, EDF at Paris et AMI Software at Montpellier) for three years and a half.

Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social

Coriane Nana Jipmo*, Gianluca Quercini*
Nacéra Bennacer*

*Laboratoire de Recherche en Informatique (LRI)
CentraleSupélec, Univ. Paris-Saclay
Gif-sur-Yvette, France
{coriane.nanajipmo, gianluca.quercini, nacera.bennacer}@lri.fr

Résumé. Cet article présente une approche pour la catégorisation et la désambiguïsation des intérêts que les individus renseignent sur les réseaux sociaux en utilisant Wikipédia.

1 Introduction

Dans cet article, nous présentons une étude préliminaire sur le problème de caractérisation et de catégorisation des intérêts que les individus renseignent sur les réseaux sociaux, comme *reading*, *jogging*, *java*, etc. Ces intérêts étant exprimés en langage naturel, nous sommes confrontés au problème de la désambiguïsation dans un contexte limité, compte tenu du peu des informations accessibles dans les profils des individus. Les approches visant la désambiguïsation des *tags* dans les *folksonomies* sont confrontées aux même problème, bien qu'elles peuvent s'appuyer sur les ressources faisant l'objet des tags pour avoir un contexte plus riche [Garcia-Silva et al. (2012)]. Nous explorons une approche permettant de désambiguïser un intérêt d'un individu par la détermination d'un article Wikipédia qui contient la description de celui-ci. La désambiguïsation d'un intérêt se fera en utilisant les autres intérêts renseignés par l'individu comme contexte. Les résultats que nous avons obtenus sur 392 intérêts issus de 50 profils utilisateurs du réseau social *LiveJournal* sont encourageants.

2 Désambiguïsation des intérêts

Soit $\mathcal{I}_u = \{I_1, I_2, ..., I_n\}$ l'ensemble des intérêts qu'un individu α renseigne sur son profil, exprimés sous la forme d'une chaîne de caractères en Anglais (*Computer*, *Music*, ...). Notre objectif est d'associer à chaque intérêt I_j, $j = 1, ..., n$, un article Wikipédia décrivant I_j ; à cet effet, pour chaque intérêt I_j notre approche sélectionne la page Wikipédia P ayant par titre I_j, si elle existe. Deux cas peuvent se présenter : (i) P est un article décrivant I_j, ou (ii) P est une page de désambiguïsation. Dans le premier cas, le mot décrivant l'intérêt I_j n'est pas ambigu (par ex., *Music*), au point que la majorité des utilisateurs Wikipédia ont trouvé un accord quant à son *interprétation* (ou, signification) par défaut (*Music* désigne une forme d'art) et lui ont associé un article. La page P sera alors choisie comme la seule interprétation de l'intérêt I_j.

Dans le deuxième cas, le mot décrivant l'intérêt I_j est ambigu et la page de désambiguïsation P permet d'avoir la liste des articles Wikipédia représentant les interprétations possibles de I_j. Afin de choisir une interprétation pour les intérêts ambigus, notre approche construit un *graphe des interprétations* $\mathcal{G}$ comme suit. Pour chaque intérêt I_j, on ajoute un nœud dans $\mathcal{G}$ pour chaque interprétation de I_j ; un arc est établi entre deux nœuds correspondant à des interprétations de deux intérêts différents dont la similarité dépasse un seul fixé τ. La similarité de deux nœuds est calculée dans le graphe de Wikipédia en utilisant la mesure de similarité WLM [Milne et Witten (2008)]. Ensuite, l'algorithme *PageRank* est utilisé pour affecter un score d'importance à chaque nœud de $\mathcal{G}$; à chaque intérêt on affecte son interprétation ayant le score le plus élevé. L'intuition derrière l'utilisation de *PageRank* sur le graphe des interprétations est que l'interprétation de chaque intérêt vote pour les interprétations similaires des autres intérêts ; la co-occurrence de deux intérêts similaires (par ex., *c++* et *Java*) est donc prise en compte pour choisir leurs interprétations correctes.

3 Evaluation

Nous avons collecté 50 profils du réseau social *LiveJournal* avec un total de 392 intérêts ; 257 intérêts distincts ont un article Wikipédia (page par défaut et/ou page de désambiguïsation), 36 n'ont aucun article. Les interprétations correctes de chaque intérêt ont été déterminées manuellement par deux évaluateurs. Les premiers résultats montrent que la page par défaut correspondant à un certain intérêt est très souvent l'interprétation correcte de l'intérêt et l'interprétation correcte d'un intérêt ambigu figure normalement parmi les trois meilleurs résultats proposés par notre approche.

Références

Garcia-Silva, A., O. Corcho, H. Alani, et A. Gomez-Perez (2012). Review of the State of the Art : Discovering and Associating Semantics to Tags in Folksonomies. *The Knowledge Engineering Review 27*(01), 57–85.

Milne, D. et I. H. Witten (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *In Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence : an Evolving Synergy*, pp. 25–30. AAAI Press.

Summary

In this paper, we present a preliminary study of the problem of characterizing and categorizing the interests (e.g., *reading*, *jogging*, *java*) that the individuals disclose on social networks. As these interests are expressed in natural language, we are confronted with a disambiguation problem in a limited context, as social network profiles have usually limited textual content. The approach we present here disambiguates an interest of an individual by determining a Wikipedia article that describes it; the other interests disclosed by the individual form the context. The results that we obtained on 392 interests of 50 user profiles of the social network *LiveJournal* are encouraging.

Un système personnalisé de recommandation basé sur le profil des utilisateurs dans les folksonomies

Mohamed Nader JELASSI[*] [**], Sadok BEN YAHIA[**], Engelbert MEPHU NGUIFO[*]

[*] Université Blaise Pascal, Clermont Ferrand, France.
nader.jelassi, mephu@isima.fr
[**]Faculté des Sciences de Tunis, Tunis, Tunisia.
nader.jelassi, sadok.benyahia@fst.rnu.tn

Résumé. Dans ce papier, nous présentons un système personnalisé de recommandation qui se base à la fois sur le profil des utilisateurs et des tags et ressources qu'ils ont partagé dans les folksonomies.

1 Introduction

Une *folksonomie* désigne un système de classification collaborative par les internautes (Mika, 2005). L'idée est de permettre à des utilisateurs de partager et de décrire des objets via des mots-clés (tags) librement choisis. Formellement, une *folksonomie* est composée de trois ensembles : un ensemble d'utilisateurs, un ensemble de tags (ou étiquettes) et un ensemble de ressources (films, livres, sites web, photos, etc.). Les utilisateurs sont les acteurs principaux du système et contribuent au contenu par l'ajout de ressources et l'affectation de tags. Cependant, il s'avère que le choix de tags et de ressources partagées par un utilisateur d'une *folksonomie* varie selon plusieurs critères : le genre, l'âge ou encore la profession de celui qui partage l'information. Cela a incité les chercheurs à proposer des systèmes de recommandation personnalisés afin de suggérer les tags et ressources les plus appropriés aux utilisateurs et de répondre aux besoins de chaque utilisateur. En effet, le domaine de personnalisation tente de fournir des solutions afin d'aider les utilisateurs à partager les bons tags et les bonnes ressources parmi le très grand nombre de données dans les folksonomies. Ainsi, un système de recommandation offre à l'utilisateur une liste de tags ou de ressources recommandés qu'il est susceptible d'aimer et lui permet de trouver plus facilement ses tags et ressources préférés dans la *folksonomie*. De plus, la personnalisation tente d'aider les utilisateurs à aborder le problème de surcharge d'information. Pour atteindre cet objectif, nous considérons le profil des utilisateurs comme une nouvelle dimension dans une *folksonomie*, classiquement composée de trois dimensions (utilisateurs, tags et ressources), et nous proposons une approche de regroupement des utilisateurs aux intérêts et profils équivalents sous forme de structure appelées concepts quadratiques (ou quadri-concepts) (Jelassi et al., 2015) (Jelassi et al., 2014).

2 Les données participatives et personnelles des utilisateurs

Les quadri-concepts que nous proposons d'extraire à partir des folksonomies afin de personnaliser les recommandations sont composées de deux parties :

Les données participatives des utilisateurs Il s'agit des tags et ressources partagées par les utilisateurs dans le passé. Chaque utilisateur est libre d'ajouter les tags et ressources de son choix dans les folksonomies. Les quadri-concepts vont ainsi permettre de regrouper les utilisateurs ayant partagé des tags et ressources en commun sous des structures quadratiques.

Le profil des utilisateurs En plus de l'historique de tagging, chaque quadri-concept représente une conceptualisation (Jäschke et al., 2008) entre utilisateurs ayant des profils équivalents. Par exemple, un quadri-concept peut être : *"Jack et Kate qui sont agés entre* 18 *et* 25 *ans ont utilisé les tags 'action' et 'aventure', parmi d'autres, pour annoter des films comme 'Indiana Jones' et 'Star Wars'"*.

3 Résultats expérimentaux et conclusion

Les résultats expérimentaux que nous avons menées sur des jeux de données du monde réel (MOVIELENS et BOOKCROSSING) ont montré que notre approche, qui prend en compte à la fois les données participatives et sociales des utilisateurs, permet d'améliorer la qualité des recommandations (en termes de précision et de rappel) par rapport aux approches de la littérature (Jelassi et al., 2014) (Jelassi et al., 2015). Parmi nos perspectives de travaux, nous voulons mettre à jour les données participatives des utilisateurs par le biais d'une approche incrémentale afin de recommander les données les plus récentes partagées par les utilisateurs.

Références

Jäschke, R., A. Hotho, C. Schmitz, B. Ganter, et G. Stumme (2008). Discovering shared conceptualizations in folksonomies. *Web Semantics. 6*, 38–53.

Jelassi, M. N., S. Ben Yahia, et E. Mephu Nguifo (2014). Vers des recommandations plus personnalisées dans les folksonomies. In *IC 2014*, Clermont-Ferrand, France.

Jelassi, M. N., S. Ben Yahia, et E. Mephu Nguifo (2015). Towards more targeted recommendations in folksonomies. *Journal of Social Network Analysis and Mining (SNAM).* Accepted..

Mika, P. (2005). Ontologies are us : A unified model of social networks and semantics. In *Proc. of ISWC 2005*, Volume 3729 of *LNCS*, pp. 522–536. Springer-Verlag.

Summary

In this paper, we present a personalized recommender system based on both users' profile and tagging history, *i.e.,* tags and resources shared by users in folksonomies.

Fabrique logicielle de réseaux sociaux spécialisés : Aspects fonctionnels

Benjamin Billet*, David Fernandez*, Didier Parigot*

*Équipe-projet ZENITH, Inria
{prenom.nom}@inria.fr

Introduction En partenariat avec la startup BEEPEERS [1], nous concevons une fabrique logicielle (Greenfield et Short, 2003) pour le développement de réseaux sociaux spécialisés à destination de communautés ciblées. L'objectif de cette fabrique est de minimiser les coûts de conception et de production de ces réseaux. Concrètement, cette fabrique opère par spécialisation d'un réseau social abstrait, au moyen d'un mécanisme de sous-typage pour obtenir les réseaux sociaux spécialisés.

Un **réseau social spécialisé** est un réseau à destination d'une communauté spécifique (p. ex. les exposants et visiteurs d'un salon, ou encore, les licenciés d'un club sportif) dont le vocabulaire et les fonctionalités sont conditionnés par les besoins de cette communauté. À l'aide de l'expertise de BEEPEERS, nous avons réalisé un **réseau social abstrait** [2] regroupant tous les concepts utilisés par leurs différents réseaux sociaux. L'obtention d'un nouveau réseau social se fait en spécialisant ce réseau abstrait au moyen de fichiers de configuration fournis à la fabrique. D'autre part, compte tenu de la forte connectivité des données manipulées par ces réseaux, ils sont conçus au dessus des concepts des **bases de données graphe**.

Plus précisément, un mécanisme de **typage** et de **sous-typage** construit au dessus des bases de données graphes permet d'hériter des concepts du réseau social abstrait pour définir un réseau social spécialisé. Concrètement, un sous-type héritant d'un type de nœud du réseau abstrait peut (i) renommer ce concept, et (ii) réduire la liste des fonctionnalités qui lui sont associées. Les types de nœud ainsi redéfinis pour un réseau social spécialisé, forment une hiérarchie de types qui définit le schéma de la base de données graphe pour ce réseau.
Pour éviter un travail de développement spécifique à chaque réseau, il est important de disposer de requêtes exécutables sur tous les réseaux. Nous avons donc élaboré un moteur de requête qui supporte ce mécanisme de typage. Ainsi, les **requêtes génériques** définis avec les types du réseau social abstrait fonctionneront pour tous les réseaux.

Nous avons établi que pour mettre en œuvre les algorithmes classiques de la **recommandation** (Bobadilla et al., 2013), p. ex. filtrage collaboratif, sous la forme de **data flow**, seuls

1. www.beepeers.com (07/10/2015)
2. Schéma du modèle de données du réseau social abstrait : https://huit.re/eiZDMvh7 (07/10/2015).

quatre opérateurs sont nécessaires : *PatternFinder*, *Product*, *Map* et *Aggregator*. Un système de data flow muni de ces quatre opérateurs et son langage dédié ont donc été élaboré. Comme ces algorithmes utilisent des requêtes génériques, ils sont exploitables par tous les réseaux sociaux spécialisés.

Conclusion Nous venons de présenter les fonctionnalités requises par une fabrique logicielle pour la production de réseaux sociaux spécialisés : mécanismes de typage, moteur de requêtes et système de recommandation génériques. À l'heure actuelle, un prototype de notre fabrique implémente ces fonctionnalités au dessus de Blueprints (Apache TinkerPop, 2015) : un ensemble d'interfaces de manipulation de graphes supporté par les principales bases de données graphe. Le système de recommandation, quant à lui, a été conçu au moyen de SON (Lahcen et Parigot, 2012) un intergiciel pour le développement d'applications orientées service à base de composants. Cela permet une exécution en parallèle et en pipeline des dataflow. Des tests effectués sur l'ensemble des données des applications de BEEPEERS nous ont permis de valider cette approche préliminaire.

Remerciements Nous remercions Mickaël Jurret de BEEPEERS pour sa collaboration active.

Références

Apache TinkerPop (2015). Tinkerpop3 Webpage. `http://tinkerpop.incubator.apache.org`.

Bobadilla, J., F. Ortega, A. Hernando, et A. Gutiérrez (2013). Recommender systems survey. *Knowledge-Based Systems 46*.

Greenfield, J. et K. Short (2003). Software factories : Assembling applications with patterns, models, frameworks and tools. In *Companion of the 18th Annual ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*.

Lahcen, A. A. et D. Parigot (2012). A lightweight middleware for developing P2P applications with component and service-based principles. In *15th IEEE International Conference on Computational Science and Engineering*.

Summary

This paper introduces a software factory for developing social networks. This factory takes an abstract social network and creates a concrete one, using mechanism of sub-typing.