Fouille sémantique de motifs fréquents probabilistes sur des données liées

Fabrice Guillet et Mounira Harzallah, DUKe, LS2N

Contexte et problématique

Cette thèse se situe à la confluence de deux domaines de recherches complémentaires : la fouille de motifs et le web sémantique. La fouille de motifs a pour objectif d'extraire des connaissances à partir de données décrites par des variables et souvent définies dans des modèles relationnels, cette connaissance s'exprimant à l'aide de modèles statistiques sous la forme de combinaisons de variables répétitives et pertinentes. Le web sémantique procède d'une démarche inverse, puisqu'il s'attache à la conception de modèles sémantiques de connaissance, souvent dénommés modèles de domaine ou ontologies, à partir d'une expertise humaine, puis à l'usage de ces modèles afin d'annoter des données/entités. Bien qu'il existe de nombreux travaux de recherche à la frontière de ces deux domaines, par exemple en extraction automatique de concept, en alignement d'ontologie, le cloisonnement thématique accroît le degré de difficulté des travaux conjoints, et en particulier la prise en compte de la sémantique des données liées en fouille de données.
Avec l'émergence du web des données (ou linked data), se renforce la tendance à disposer de données "mixtes". Cette mixité se traduit d'une part à un niveau syntaxique où les données comportant fréquemment une composante textuelle sont décrites simultanément par des variables et des entités; d'autre part à un niveau sémantique par des propriétés logiques définies dans des ontologies annotant des entités. En complément, ces données sont stockées de manière croissante dans des formats non relationnels (stockages verticaux, nosql, rdf)

Objectifs et verrous scientifiques

L'objectif de cette thèse est de concevoir de nouveaux modèles de fouille de motifs enrichis, mieux adaptés à la complexité inhérente à la mixité de ce nouveau type de données, selon deux axes complémentaires :
(1) les *motifs fréquents probabilistes* : le croisement de modèles probabilistes de "topic modeling" et de modèles statistiques de motifs fréquents et de règles d'association, afin de mieux prendre en compte la mixité des données au niveau syntaxique et de contourner les limites des deux modèles;
(2) les *motifs sémantiques* : l'enrichissement sémantique des motifs fréquents probabilistes afin d'y intégrer les concepts et propriétés définies dans des ontologies au niveau syntaxique.

Domaine d'application

Les motifs ainsi enrichis permettront de découvrir de nouveaux liens entre données et ontologie sous la forme de nouveaux clusters/concepts, et de nouvelles associations/propriétés. L'usage cible sera la recommandation, mais les motifs enrichis pourront aussi permettre la découverte de nouvelles annotations, la révision et enrichissement d'ontologies.
Afin de disposer d'un corpus accessible, facilement compréhensible, permettant la réplication, et servant de référentiel, cette thèse utilisera les pages wikipedia et leurs annotations dérivées dans dbpedia, l'objectif in fine étant d'effectuer de la recommandation de pages. On se placera également dans le cadre des formats de stockage non relationnels imposés par les données liées.

Profil du candidat

Le ou la candidat(e) devra avoir des bonnes connaissances en fouille de données (en particulier, les techniques statistiques et probabilistes). Des connaissances en web sémantique sont souhaitables.

Candidature

Les candidats intéressés sont invités à envoyer, **avant le 16/05/2017**, à Fabrice Guillet (Fabrice.Guillet@univ-nantes.fr) et à Mounira Harzallah (mounira.harzallah@univ-nantes.fr) : un CV (avec les coordonnées de deux contacts pour une éventuelle recommandation), une lettre de motivation et les notes des deux dernières années d'étude ainsi que le classement de l'étudiant dans sa promotion.

Environnement de Travail

La thèse de doctorat se déroulera au sein de l'équipe DUKe (Data User Knowledge) du laboratoire LS2N de Université de Nantes à Polytech'Nantes.
Durée 3 ans, début de thèse en septembre 2017

Plan de travail

Dans la première année:

- le ou la candidat(e) effectuera une étude bibliographique sur les méthodes de fouille de motifs fréquents, et les méthodes probabilistes de « topic modeling » appliquées aux données liées sur le web. Ensuite, il (ou elle) étudiera les travaux qui ont appliqué ces méthodes sur des données du web annotées avec une ontologie et ceux qui ont couplé ces méthodes. Enfin, une étude bibliographique des méthodes de fouille pour l'enrichissement d'ontologie sera également réalisée.
- le ou la candidat(e) testera des méthodes de fouille de motifs fréquents et des méthodes « LDA » sur un corpus restreint de textes annotés avec DBpédia.

Dans la deuxième année, le ou la candidat(e) proposera une nouvelle approche de fouille sémantique à partir des données du web qui combine les méthodes de fouille de motifs fréquents, les règles d'association et les méthodes LDA, en intégrant les connaissances de l'ontologie qui annote ces données (i.e. ses concepts, ses relations et leurs propriétés et ses axiomes). Il (ou elle) complétera cette approche par une méthode d'enrichissement de cette

ontologie et de l'annotation de ces données, en utilisant ses résultats (i.e. les résultats de cette nouvelle approche de fouille sémantique).

Dans la troisième année, le ou la candidat(e) expérimentera cette nouvelle approche sur des corpus de données de Wikipédia annotés par DBpédia. Il (ou elle) l'évaluera en la comparant à d'autres approches d'annotation de textes avec DBpédia ou d'enrichissement d'ontologie.

Etat de l'art

Née de l'analyse du panier de la ménagère, la fouille de motifs fréquents s'intéresse à la découverte de modèles statistiques sous la forme de combinaisons pertinentes de produits. Ces modèles ont fait l'objet de nombreuses publications (*Han et al. 2007*), et ont été étendus aux différents types de données disponibles, tels que les séquences, les graphes, etc… L'avantage de ces modèles réside dans leur caractère non supervisé et leur interprétabilité, ce qui les rend aisés à comprendre par un expert et plus aptes à découvrir des nouveautés. Leur inconvénient dérive de leur manque de synthèse, lié au très grand nombre de nouveautés, qui les rend impraticable sans post-traitement à l'aide de mesures de qualité, de contraintes, de supervision par un décideur.

Dans l'optique d'étendre la fouille de motif par des représentations du web sémantique, on trouve des travaux permettant d'intégrer des ontologies dans les règles d'association, afin de faciliter la recommandation (*Missaoui et al. 2007*) ou de produire des motifs généralisés (*Kwuida et al. 2009*), ou encore de permettre à un utilisateur de diriger le processus de fouille (*Marinica & Guillet 2010*). Mais la mixité des données n'est pas prise en compte et les modèles des deux premières approches restent "très bavards" dans leurs résultats.

Sous l'angle du web sémantique, de nombreux travaux proposent d'intégrer des modèles de motifs fréquents. L'une des problématiques majeures, L'extraction d'ontologies (ontology learning) à partir de texte, voir (*Buitelaar et al. 2005, Poon & Domingos 2010, Petasis et al. 2011, Gherasim et al. 2013*) pour une synthèse, fait appel à de nombreux modèles statistiques, non supervisés, généralement basés sur les sacs de mots, qui se restreignent à des données textuelles. *Rettinger et al. 2012* évaluent l'intérêt des modèles statistiques à base de similarité en web sémantique dans le cadre des bases de connaissances comme Yago (*Suchanek et al. 2007*), mais les données étudiées se limitent aux entités annotées.
En considérant plus particulièrement l'usage des règles d'association en web sémantique, on trouve des travaux sur l'alignement d'ontologie (*David 2006, 2007*), mais aussi sur la découverte de règles entre propriétés dans des bases de connaissances en logique "monde ouvert" (*Galárraga et al. 2013, 2015*). Notons que les règles d'association sont aussi un cas particulier de dépendances fonctionnelles, dont on trouvera une synthèse récente en *Caruccio et al. 2016*.

Enfin, au-delà de leurs applications au web sémantique, les modèles probabilistes non supervisés comme le LDA (*Blei 2012*), sont très utilisés en "topic modeling" pour découvrir des thèmes dans des données textuelles. Par ailleurs, ces modèles peuvent être étendus (*Han et al. 2014*) à des données catégorielles, non textuelles. Mais, a contrario des motifs fréquents, l'interprétation des topics probabilistes reste difficile car les liens effectifs entre termes d'un même topic ou entre les topics ne sont pas fournis par le modèle.

Dans la prolongation de *Pépin et al. 2015*, la première proposition de cette thèse consistant à coupler topic modeling et motifs fréquents, vise à réduire le nombre de motifs fréquents candidats grâce aux topics LDA, et d'améliorer la compréhension des topics en caractérisant les liens intra/inter topics grâce aux règles d'association. Les contraintes taxonomiques de l'ontologie d'annotation et ses propriétés d'héritage vont aider à interpréter ces liens, voire à les enrichir.

Références

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, *15*(1), 55-86.

Missaoui, R., Valtchev, P., Djeraba, C., & Adda, M. (2007). Toward recommendation based on ontology-powered web-usage mining. *IEEE Internet Computing*, *11*(4).

Kwuida, L., Missaoui, R., Boumedjout, L., & Vaillancourt, J. (2009). Mining generalized patterns from large databases using ontologies. *arXiv preprint arXiv:0905.4713*.

Marinica, C., & Guillet, F. (2010). Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 784-797.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, *123*, 3-12.

Poon, H., & Domingos, P. (2010, July). Unsupervised ontology induction from text. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 296-305). Association for Computational Linguistics.

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., & Zavitsanos, E. (2011). Ontology population and enrichment: State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution* (pp. 134-166). Springer-Verlag.

Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., & Fanizzi, N. (2012). Mining the semantic web - Statistical learning for next generation knowledge bases. *Data Mining and Knowledge Discovery*, *24*(3), 613-662.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007, May). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* (pp. 697-706). ACM.

David, J., Guillet, F., & Briand, H. (2006, November). Matching directories and OWL ontologies with AROMA. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 830-831). ACM.

David, J. (2007). Association rule ontology matching approach. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *3*(2), 27-49.

Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013, May). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 413-422). ACM.

Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. *The VLDB Journal*, *24*(6), 707-730.

Gherasim,T., Harzallah, M., Berio, G. Kuntz, P. (2013) Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application. Pages 177–201 of: Advances in knowledge discovery and management, vol. 471. Springer.

Caruccio, L., Deufemia, V., & Polese, G. (2016). Relaxed Functional Dependencies—A Survey of Approaches. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 147-165.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Han, J., Wang, C., & El-Kishky, A. (2014, August). Bringing structure to text: mining phrases, entities, topics, and hierarchies. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1968-1968). ACM.

Pépin, L., Blanchard, J., Guillet, F., Kuntz, P., & Suignard, P. (2015). Visual Analysis of Topics in Twitter Based on Co-evolution of Terms. In *Data Science, Learning by Latent Structures, and Knowledge Discovery* (pp. 169-178). Springer Berlin Heidelberg.