Extraction et Analyse de réseaux sociaux issus de Bases de Données Relationnelles

Rania Soussi *, Amine Louati**
Marie-Aude Aufaure* Hajer Baazaoui**,Yves Lechevallier***,Henda Ben Ghézala**

*Laboratoire MAS Ecole Centrale de Paris Grande Voie des Vignes
Chatenay-Malabry, France
{Rania.Soussi,Marie-Aude.Aufaure}@ecp.fr,
**Laboratoire RIADI-GDL, Ecole Nationale des Sciences de l'Informatique,
Compus Universitaire de la manouba La Manouba 2010
{Amine.louati,hajer.baazaouizghal,henda.BenGhézala}@riadi.rnu.tn
***INRIA-Rocquencourt,Domaine de Voluceau 78150 Rocquencourt
Yves.Lechevallier@inria.fr

Résumé. Dans un contexte d'entreprise, beaucoup d'informations importantes restent stockées dans des bases de données relationnelles, constituant une source riche pour construire des réseaux sociaux. Le réseau, ainsi extrait, a souvent une taille importante ce qui rend son analyse et sa visualisation difficiles. Dans ce travail, nous proposons une étape d'extraction suivie d'une étape d'agrégation des réseaux sociaux à partir des bases de données relationnelles. L'étape d'extraction ou de construction transforme une base de données relationnelle en base de données graphe, puis le réseau social est extrait. L'étape d'agrégation, qui est basée sur l'algorithme k-SNAP, produit un graphe résumé.

1 Introduction

Les réseaux sociaux jouent un rôle important dans le partage et la recherche d'information. Un réseau social (RS) est généralement représenté par une structure de graphe. Les sommets désignent des individus, des groupes ou des organisations et sont reliés entre eux par des interactions ou des liaisons qui forment les arêtes de ce graphe. Ceci permet de décrire d'une façon naturelle différents types de collaboration et d'échange entre des individus, les liaisons entre des laboratoires etc. Dans un contexte d'entreprise, l'objectif d'un RS est d'informer sur les rôles des personnes (par exemple : qui détient l'information, qui est le responsable, qui est l'expert) et leurs interactions (par exemple : qui collabore avec qui). Les techniques actuelles de construction des RS se basent sur des données extraites à partir des documents du web. Cependant, les données d'entreprise sont stockées dans des fichiers (XML, Excel,..) et plus particuliérement dans des bases de données relationnelles (BDR). Nous proposons une approche de construction automatique d'un RS à partir de BDR d'entreprise (Soussi et al., 2010). Le réseau ainsi construit sous la forme d'un graphe peut avoir une taille très importante. Par conséquent, il devient difficile d'exploiter et surtout d'interpréter d'une manière significative l'information de ce graphe par une simple visualisation. D'où la nécessité de disposer de

méthodes efficaces d'agrégation d'information permettant de produire un graphe résumé qui conserve non seulement les principales caractéristiques structurelles mais surtout améliore les performances d'analyse et d'interprétation. Ce papier est organisé comme suit : dans la section 2, nous présentons les différentes approches existantes d'extraction des RS. La section 3 introduit notre approche qui transforme une BDR en base de données graphe (Angles et Gutierrez, 2008), puis un RS est extrait. Dans la section 4, nous présentons l'étape d'agrégation basée sur l'algorithme k-SNAP qui produit un graphe résumé grâce à un groupement des nœuds en fonction des attributs et des relations sélectionnés par l'utilisateur. La section 5 conclut et aborde nos principales perspectives.

2 Approches d'extraction des réseaux sociaux

Plusieurs approches d'extraction des RS ont été proposées. Kautz et son équipe proposent l'outil Refferal Web (Kautz et al., 1997) qui permet d'extraire un RS en utilisant des données issues d'un site web telle que : les liens dans les pages personnelles, les listes des co-auteurs dans les papiers techniques et les citations, les communications entre des individus enregistrés dans les archives de Netnews et les chartes des organisations. La même méthode a été utilisée dans Flink(Mika, 2005). McCallum et al. (Bekkerman et McCallum, 2005) présentent un système qui extrait un réseau social d'utilisateurs à partir d'une boite email. Ce système identifie les personnes dans les messages dans une boite email, trouve leurs pages web personnelles et remplit les champs d'un carnet d'adresses. ArnetMiner(Tang et al., 2008) est un autre système d'extraction des RS académiques. Il est basé sur l'extraction automatique des profils des chercheurs. Les approches existantes d'extraction des RS utilisent toutes des données extraites du web. Ces approches utilisent la cooccurrence des noms dans le web pour déterminer les relations entre les individus (Mika, 2005). Dans ce travail, nous allons extraire le RS à partir d'une BDR. Cette approche doit distinguer les tables qui représentent des individus de celle d'autres entités et leurs relations. La section suivante décrit l'approche proposée.

3 Approche de construction d'un réseau social à partir d'une base de données relationnelle

Notre approche d'extraction est basée sur deux étapes : (1) la transformation de la BDR en base de donnée graphe et (2) la construction du RS à partir de la base de données graphe.

3.1 Transformation de la base de données relationnelle en base de données graphe

Cette étape de transformation permet d'avoir un graphe contenant tous les objets stockés dans la BDR et les relations entre eux. Ce graphe d'objets va faciliter l'extraction des entités du RS. Les bases de données graphe (BDG)(Angles et Gutierrez, 2008) représentent une solution intéressante et naturelle pour modéliser les RS. L'avantage de ce type de structure est sa dynamique et la possibilité de représenter des relations, éventuellement multiples, entre les objets. Elle peut aussi représenter des objets complexes ayant des multiples attributs. Nous avons choisi de travailler avec le modèle hypernode (Levene et Loizou, 1995) permettant la

modélisation de graphes complexes. L'intérêt de ce modèle est de pouvoir représenter et visualiser de manière explicite les relations entre les entités, alors que ces mêmes relations sont exprimées dans les BDR à travers les clés primaires et étrangères, donc moins visibles. La BDG est composée de deux niveaux : le niveau schéma qui décrit les métadonnées de la base (les types des objets et les relations, les attributs de chaque objet et leurs types) et le niveau instance (Fig 1). La transformation de la BDR en une BDG inclut la translation du schéma et la conversion des données (Maatuk et al., 2008). Nous avons détaillé dans (Soussi et al., 2010) cette phase de transformation en BDG (ici base de données hypernode). *Á* partir des sept tables de la BDR présentée dans (Soussi et al., 2010) nous avons six hypernodes (les tables $Thesis$, $Laboratory$, $Thesis\_hasStudent$, $Student$, $Director\_thesis$ et $Foreign\_Student$) et une table de liaison (la table $Thesis\_hasLab$ car elle ne contient que des clés) qui sera une relation dans la BDG. L'hypernode $Foreign\_Student$ est composé d'un nœud $(country)$ avec un type prédéfini et un nœud St_id relié à l'hypernode $Student$ de $\mathcal{H}$.

3.2 Transformation de la base de données graphe en réseau social

Á partir de la BDG, nous appliquons des règles de transformation de cette base de données hypernode (BDH) en un RS modélisé par un graphe où les nœuds sont des personnes et les arcs sont les relations entre ces personnes appelées *entités*. Le processus d'extraction du RS est réalisé en deux étapes : l'identification des entités puis la construction des relations entre ces entités.

L'identification des entités est réalisée en deux étapes :(1) le schéma de la BDG est utilisé pour extraire les hypernodes candidats (ceux représentant des individus), puis (2) les hypernodes instances sont utilisés pour détecter ceux représentant réellement des entités.

Détection des hypernodes candidats. Ainsi, nous utilisons une ontologie de personnes , contenant les caractéristiques d'une personne (nom, prénom,...) et construite manuellement pour analyser les hypernodes. En effet, l'ensemble des nœuds de chaque hypernode *h* de la BDH est analysé avec les concepts de l'ontologie.

Analyse des hypernodes candidats. Chaque hypernode candidat h a un ensemble d'hypernodes instances h_i. Pour analyser h, chaque nom trouvé dans chaque h_i (nous prenons les 10 premières instances de h) est envoyé à un moteur de recherche. Seul les 10 premiers documents sont traités et analysés avec NER (Named entity Recognition), un outil de reconnaissance des entités nommées proposé par Stanford [1]. Nous attribuons à chaque document une valeur *rd*. Si le nom est étiqueté dans le document par *Personne*, nous attribuons au document la valeur *rd*=1 sinon *rd*=0. La moyenne assignée au nom trouvé dans *hi* (*avghi*) représente le nombre de fois *où* ce nom est étiqueté par le tag *Personne* (le nom considéré est ainsi un nom de personne) dans les documents : $avghi = \frac{\sum rd}{nombre-de-documents}$. La moyenne assignée à l'hypernode (*avgH*) donne la moyenne de nombre de fois o*ù* les noms trouvés dans ses instances sont considérés comme des noms de personnes : $avgH = \frac{\sum avghi}{nombre-de-h_i}$.

Lorsqu'un hypernode est identifié comme une entité alors toutes ses instances sont ajoutées au réseau.

L'étape suivante est celle de la *construction des relations*. Au cours de l'étape de l'extraction de graphe, plusieurs relations entre les différents objets ont été identifiées. Les relations

1. http ://nlp.stanford.edu/ner/index.shtml

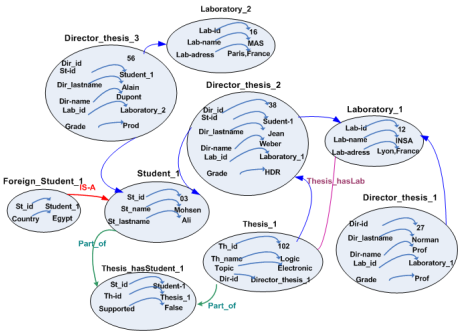

FIG. 1 – *une partie de la BDH instance*

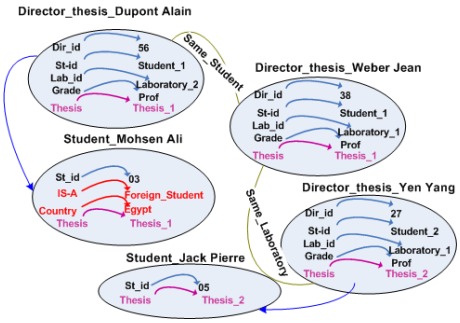

FIG. 2 – *Le RS correspondant*

existantes entre les entités dans la BDG sont maintenues dans le réseau social et nous cherchons à extraire des relations cachées. Pour faciliter cette tâche, un ensemble de patrons de relations a été proposé en utilisant le schéma de la BDH. Ces patrons permettent soit de trouver des relations existantes (dans la BDG) entre des hypernodes instances soit de créer des nouvelles relations. En appliquant les patrons identifiés sur la BDH instance (Fig 1), un RS est extrait (Fig 2). Cependant, les BDR réelles contiennent des centaines voire des milliers d'enregistrements et nous obtenons un RS large. Les outils disponibles ainsi que les méthodes traditionnelles pour appréhender ce type de graphe, demeurent impuissants pour comprendre et surtout pour interpréter d'une manière significative l'information codée visuellement. *À* partir du réseau résultant (Fig 2), Un graphe constitué d' hypernodes homogènes est extrait en fonction du centre d'intérêt de l'utilisateur (les directeurs de thèse par exemple) afin d'appliquer l'algorithme k-SNAP permettant d'obtenir une vue agrégée de ce graphe et de l'analyser plus finement. Le graphe, sur lequel K-SNAP est exécuté contient les hypernodes instances de "Director_thesis" comme sommets et les relations entre eux comme arrêtes.

4 Agrégation des réseaux sociaux à l'aide du *k-SNAP*

L'agrégation des graphes est une méthode qui permet d'alléger la visualisation et de mettre en évidence les communautés présentes dans le réseau, ce qui facilite énormément l'interprétation. La plupart des travaux existants utilisent des procédés statistiques, tels que *degree distributions*, *hop-plots* (Chakrabarti et al., 2007) et *clustering coefficients* ; les résultats obtenus sont souvent utiles mais difficiles à contrôler et surtout à exploiter, d'autres emploient des algorithmes de partitionnement hiérarchique de graphe comme *superGraph* (Rodrigues Jr. et al., 2006) pour visualiser les graphes larges, cependant ces techniques ignorent totalement les attributs associés aux nœuds ce qui rend l'interprétation très relative. Enfin, certains algorithmes utilisent la notion de similarité tel que *k-SNAP* qui se base sur les attributs des sommets et les relations choisies par l'utilisateur pour agréger le graphe. Nous avons choisi d'aggréger nos graphes avec *k-SNAP* (Tian et al., 2008) puisqu'il permet de conserver les principales caractéristiques structurelles avec une perte minimale d'information et surtout qu'il donne comme résultat un graphe plus petit donc moins encombré ce qui améliore les performances d'analyse et d'interprétation. Dans cette section, nous allons appliquer l'algorithme sur le graphe extrait

à partir du RS des directeurs de thèse issu de la section précédente. La visualisation directe du graphe (image de gauche de la figure 3) met en évidence notre incapacité à interpréter ce graphe sans traitement supplémentaire. L'opération d'analyse portera sur l'attribut *grade* et les relations *Same_Laboratory* et *Same_student*.

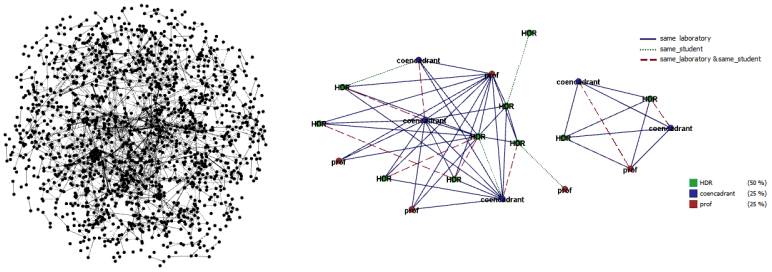

FIG. 3 – *Le graphe complet et une partie de ce graphe*

Au début *k-SNAP* génère un graphe formé par trois groupements A-compatible HDR, $coencadrant$ et $prof$ à partir de l'attribut *grade* comme le montre le graphe de droite de la figure 3 qui est une partie du graphe réel. La première itération (Fig 4) a conduit à la subdivision du groupe HDR de l'attribut *grade* en deux groupes HDR_1 et HDR_2 par rapport à la relation *Same_Student* car elle maximise notre critère d'évaluation. Cette itération donne naissance à deux groupes : Le groupe HDR_1 qui est constitué par les HDRs qui encadrent un étudiant avec au moins un professeur ou un coencadrant. Le groupe HDR_2 est constitué par les *HDR*s qui encadrent des étudiants ayant des directeurs de thèse uniquement *HDR*.

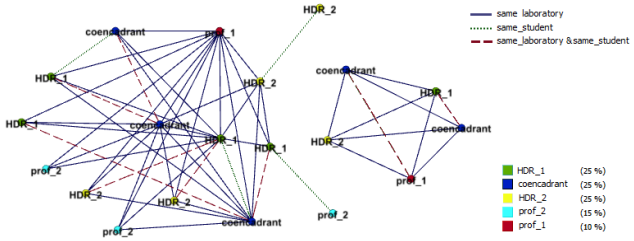

FIG. 4 – *Aperçu du graphe après la deuxième itération*

Au cours de la deuxième itération, le groupe $prof$ est découpé en deux sous-groupes par rapport à la relation *Same_laboratory*. Les professeurs du groupe $prof\_1$ partagent le laboratoire avec au moins une personne des autres groupes (HDR ou $coencadrant$) c'est à dire, que le laboratoire est constitué par des personnes ayant divers grades.

5 Conclusion

Dans ce travail, nous avons présenté une approche d'extraction des RS à partir de BDR suivie d'une étape d'agrégation à l'aide de k-SNAP. Nous avons présenté l'application de notre

approche sur une BDR réelle. Une perspective de ce travail est de rendre plus générique l'étape de transformation de la BDG selon des entités d'intérêt sélectionnées par l'utilisateur. La détection d'entités manipulées dans les entreprises comme des projets par exemple s'avère plus complexe. Pour ce faire, nous comptons définir une ontologie d'entreprise générique et nous appuyer sur celle-ci pour appliquer les patrons de transformation. L'autre perspective de ce travail est d'utiliser d'autre critère d'évaluation que celui proposé par *k-SNAP* et de relaxer la propriété de *A-compatible* en utilisant une méthode de classification.

Références

Angles, R. et C. Gutierrez (2008). Survey of graph database models. *ACM Comput. Surv. 40*(1), 1–39.

Bekkerman, R. et A. McCallum (2005). Disambiguating web appearances of people in a social network. In *WWW '05*, New York, NY, USA, pp. 463–470. ACM.

Chakrabarti, D., C. Faloutsos, et Y. Zhan (2007). Visualization of large networks with min-cut plots, a-plots and r-mat. *Int. J. Hum.-Comput. Stud. 65*(5), 434–445.

Kautz, H., B. Selman, et M. Shah (1997). The hidden web. *AI magazine 18*, 27–35.

Levene, M. et G. Loizou (1995). A graph-based data model and its ramifications. *IEEE Trans. on Knowl. and Data Eng. 7*(5), 809–823.

Maatuk, M. A., M. A. Ali, et B. N. Rossiter (2008). Relational database migration : A perspective. In *DEXA*, pp. 676–683.

Mika, P. (2005). Flink : Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics 3*(2-3), 211–223.

Rodrigues Jr., J. F., A. J. M. Traina, C. Faloutsos, et C. Traina Jr. (2006). Supergraph visualization. In *ISM '06*, Washington, DC, USA, pp. 227–234. IEEE Computer Society.

Soussi, R., M.-A. Aufaure, et H. B. Zghal (2010). Towards social network extraction using a graph database. In *DBKDA'10*, pp. 28–34.

Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang, et Z. Su (2008). Arnetminer : extraction and mining of academic social networks. In *KDD '08*, New York, NY, USA, pp. 990–998. ACM.

Tian, Y., R. A. Hankins, et J. M. Patel (2008). Efficient aggregation for graph summarization. In *SIGMOD '08*, New York, NY, USA, pp. 567–580. ACM.

Summary

In the enterprise context, a considerable amount of information is stored in relational databases. Therefore, relational database can be a rich source to extract social network. The extracted network has in general a huge size which makes its analyses and visualization difficult tasks. An aggregation step is needed in order to have more understandable graphs. In this work, we propose a social network extraction approach from relational database then we aggregate the resulting network using the k-SNAP algorithm which produces a resumed graph.