Empreintes conceptuelles et spatiales pour la caractérisation des réseaux sociaux

Bénédicte Le Grand*, Marie-Aude Aufaure** and Michel Soto*

*Laboratoire d'Informatique de Paris 6 – UPMC
{Benedicte.Le-Grand, Michel.Soto}@lip6.fr
**Laboratoire MAS, Ecole Centrale Paris
Marie-Aude.Aufaure@ecp.fr

Résumé. Cet article propose une méthode reposant sur l'utilisation de l'Analyse Formelle de Concepts et des treillis de Galois pour l'analyse de systèmes complexes. Des statistiques reposant sur ces treillis permettent de calculer la *distribution conceptuelle* des objets classifiés par le treillis. L'expérimentation sur des échantillons de trois réseaux sociaux en ligne illustre l'utilisation de ces statistiques pour la caractérisation globale et pour le filtrage automatique de ces systèmes.

1 Introduction

L'objectif de ce papier est de proposer une méthode reposant sur l'utilisation de l'Analyse Formelle de Concepts et des treillis de Galois pour l'analyse de systèmes complexes. Cette technique fournit une caractérisation visuelle et intuitive de ces systèmes par le biais du calcul d'*empreintes conceptuelles* calculées à partir de treillis de Galois. Ces empreintes (définies dans la section 2.2) aident l'observateur à mieux comprendre la structure et les propriétés des données étudiées, et à identifier les éléments significatifs ou au contraire marginaux. Cette méthode permet également d'automatiser le processus de filtrage des éléments marginaux.

Bien que cette approche soit applicable à tout type de systèmes complexes, nous avons choisi de l'appliquer au contexte des réseaux sociaux. Les réseaux sociaux en ligne tels que Myspace, Facebook ou Flickr connaissent un succès grandissant ; ces sites permettent de construire des réseaux sociaux basés sur des relations professionnelles, des loisirs communs, etc. La recherche et la navigation dans ces réseaux, ainsi que leur visualisation, sont devenues des tâches ambitieuses.

L'analyse des réseaux sociaux (Wasserman et al., 1994) consiste à comprendre et interpréter le comportement d'un réseau. Cette analyse peut également fournir des informations sur la manière dont les communautés se forment et interagissent. Les réseaux sociaux ont été étudiés d'un point de vue mathématique et statistique (Newman, 2003), mais aussi en informatique pour les aspects recherche, navigation et visualisation sociale (Brusilovsky, 2008). Une manière intéressante de comprendre et interpréter les interactions dans les réseaux sociaux est de combiner des techniques d'analyse avec la visualisation, comme dans le logiciel Pajek (Batagelj et al., 2003). Les techniques proposées ici constituent une autre approche de ces réseaux, comme présenté dans la suite.

2 Empreintes conceptuelles de réseaux sociaux

2.1 Analyse Formelle de Concepts et treillis de Galois

L'Analyse Formelle de Concepts est une approche mathématique de l'analyse de données qui fournit une structure à l'information, utilisée par exemple pour le clustering conceptuel comme dans (Carpineto et al., 1993) et (Wille, 1984). La notion de treillis de Galois a été introduite par (Birkhoff, 1940) et par (Barbut et al., 1970). L'algorithme de Galois consiste à regrouper les objets étudiés dans des classes en fonction des propriétés qu'ils ont en commun.

Soient deux ensembles finis D (un ensemble d'*objets*) et M (l'ensemble des *propriétés* de ces objets), et une relation binaire $R \subseteq DxM$ entre ces deux ensembles. Soit *o* un objet de D et *p* une propriété de M. On a *o*R*p* si l'objet *o* possède la propriété *p*.

Soit P(D) une partition de D et P(M) une partition de M. Chaque élément du treillis est un couple, aussi appelé *concept*, noté (O, A). Un concept est composé des deux ensembles $O \in P(D)$ et $A \in P(M)$ qui satisfont les propriétés suivantes (1):

$A = f(O)$, où $f(O) = \{a \in M \mid o \in O, oRa\}$

$O = f'(A)$, où $f'(A) = \{o \in D \mid a \in A, oRa\}$

O est appelé l'extension du concept, et A en est l'intention. L'extension représente un sous-ensemble des objets de l'application et l'intention représente les propriétés communes à ces objets.

La complexité des treillis de Galois (liée en particulier à leur taille, liée au nombre de concepts qu'ils contiennent) peut les rendre extrêmement difficiles à interpréter avec des diagrammes de Hasse traditionnels. (Jay et al., 2008) ont défini des mesures intéressantes pour réduire la taille de grands treillis de concepts et appliquent leur méthode aux communautés de soins et de santé. Notre approche consiste à calculer des statistiques conduisant à la définition de ce que nous appelons la *distribution conceptuelle*, présentée dans la section suivante.

2.2 Distribution conceptuelle des objets d'un système complexe

Le treillis de Galois est utilisé pour calculer des statistiques sur chacun des objets du système complexe correspondant. Dans un réseau social, les *objets* peuvent être les membres du réseau. Dans ce cas, les *propriétés* d'un objet sont par exemple ses contacts. Dans la suite, on considère un objet *o*.

- **Calcul de la Relatedness**

Soit C l'ensemble des concepts du treillis contenant l'objet *o* dans leur extension. Soit C' l'ensemble des concepts de C contenant au moins un autre objet que *o* dans leur extension et au moins une propriété dans leur intention. La valeur de Relatedness de l'objet *o* (notée *Relatedness(o)*) correspond au nombre moyen d'objets avec lesquels *o* est regroupé dans les concepts de C', divisé par le nombre total d'objets du réseau social. La valeur de Relatedness indique si *o* est connecté à de nombreux autres objets.

- **Calcul de la Closeness**

Soit S l'ensemble des objets avec lesquels *o* est regroupé dans au moins un concept du treillis (i.e. l'ensemble des objets avec lesquels *o* est connecté) ; ces objets possèdent au moins une propriété commune avec *o* (par construction). La valeur de Closeness de l'objet *o*

(notée *Closeness(o)*) est le nombre moyen de propriétés que *o* partage avec les autres objets de S, divisé par le nombre total de propriétés de *o*. Ce paramètre permet d'indiquer si l'objet *o* a une ressemblance faible ou forte avec les autres objets de S.

- **Distribution conceptuelle**

Le couple *(Relatedness(o), Closeness(o))* constitue la *distribution conceptuelle* de l'objet *o*.

- **Empreinte conceptuelle**

La valeur moyenne des paramètres de Relatedness et de Closeness pour tous les objets d'un système donné constitue l'*empreinte conceptuelle* de ce système. Soit *syst* le système complexe étudié comportant N objets.

$$empr\_conc(syst) = (relatedness(syst), closeness(syst))$$

$$\text{où } relatedness(syst) = \frac{\sum_{i=1}^{N} relatedness(o_i)}{N}$$

$$\text{et } closeness(syst) = \frac{\sum_{i=1}^{N} closeness(o_i)}{N}$$

3 Expérimentation et résultats

3.1 Description des ensembles de données

Pour cette expérimentation, nous avons utilisé quatre échantillons de réseaux sociaux collectés à l'aide d'un "crawler" dédié permettant de parcourir ces réseaux à partir de points d'entrée donnés (c'est-à-dire des membres de ces réseaux sociaux). Le crawler récupère les contacts des membres choisis comme points d'entrée et recherche les contacts de ces contacts de manière récursive. L'extraction des contacts est bornée par deux paramètres qui définissent respectivement la profondeur de la recherche et le nombre maximum de contacts recherchés pour chaque membre. Chaque membre parcouru, c'est-à-dire dont les contacts ont été retrouvés, est appelé un objet et ses contacts correspondent à ses propriétés. Dans nos échantillons, le nombre de membres total de chacun des réseaux sociaux est beaucoup plus élevé que le nombre d'objets collectés car seule une fraction d'entre eux a été parcourue (à cause de la restriction en profondeur).

Echantillon	Réseau social en ligne	Nb d'objets	Nombre de points de départ	Profondeur	Nombre Max de contacts à chaque itération	Nombre total de contacts
Myspace	Myspace	20	2	2	50	726
Flickr5	Flickr	56	3	3	50	994
DailyMotion	DailyMotion	61	2	3	5	228
Flickr50	Flickr	217	2	3	5	1336

Tableau 1. Echantillons collectés sur différents réseaux sociaux en ligne

Ces échantillons représentent des vues partielles et biaisées des réseaux sociaux étudiés, dans la mesure où ils ne contiennent que des sous-ensembles des membres sélectionnés sur des critères différents en termes de point(s) de départ, de profondeur et de nombre maximum

de contacts pour chaque personne). Les interprétations sont valides localement, c'est-à-dire, pour les "régions" dont sont issues nos données. Comme nous le verrons dans la suite, il est donc possible d'obtenir des résultats différents avec les 2 échantillons collectés sur Flickr.

3.2 Résultats de l'analyse conceptuelle

La Figure 1 représente les empreintes conceptuelles des quatre échantillons de réseaux sociaux étudiés, constituées des paramètres de *Relatedness* et de *Closeness* de ces systèmes.

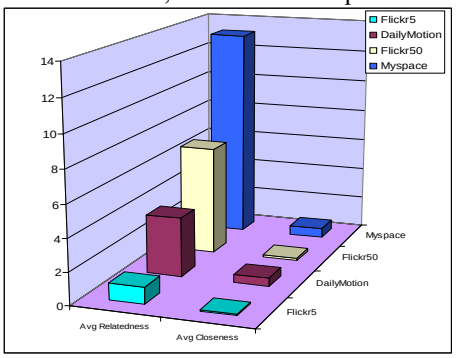

Fig. 1. Empreintes conceptuelles

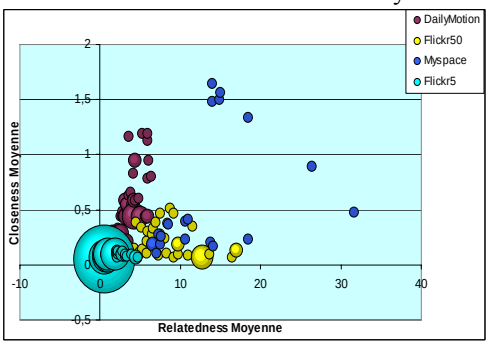

Fig. 2. Distributions conceptuelles des individus

La Figure 1 montre que l'échantillon *Myspace* a la valeur de Relatedness moyenne la plus élevée, devant *Flickr50, DailyMotion* et enfin *Flickr5*. Cela signifie que les membres de l'échantillon *Myspace* ont des contacts communs avec une plus grande proportion des autres membres de leur réseau que les membres des trois autres échantillons. Les faibles valeurs de Relatedness moyenne de *Flickr5* signifient que les membres de cet échantillon partagent peu de contacts communs. Les valeurs de Closeness permettent de savoir si les liens entre les membres sont forts ou non puisqu'elle indique s'il y a peu ou beaucoup de contacts communs. La Figure 1 montre que les valeurs de Closeness moyenne sont faibles.

Afin d'analyser plus précisément ces échantillons, nous étudions la distribution conceptuelle de chacun de leurs membres. Sur la Figure 2, chaque bulle représente un ensemble d'objets du réseau social dont les coordonnées du centre sont la Relatedness et la Closeness correspondantes. La distribution des membres individuels fournit une indication supplémentaire quant à l'homogénéité de ces valeurs. En effet, les valeurs de Relatedness et surtout de Closeness de l'échantillon *Flickr50* sont beaucoup plus homogènes (i.e. les bulles sont beaucoup plus regroupées dans une même zone du graphique) que celles de l'échantillon *Myspace* pour lequel on peut distinguer plusieurs clusters.

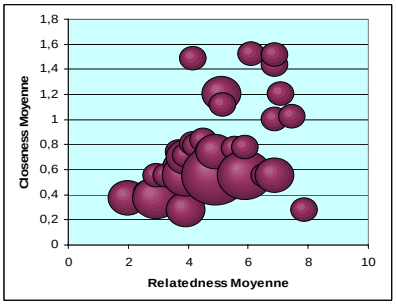

Fig. 3. Filtrage de *DailyMotion*

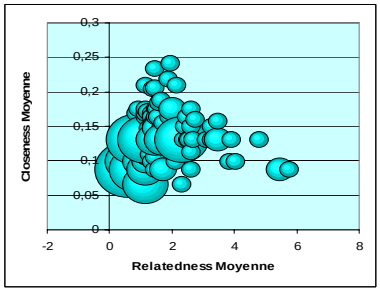

Fig. 4. Filtrage de *Flickr5*

Les distributions conceptuelles des membres des réseaux *Flickr5* et *DailyMotion* mettent en évidence des membres dont les valeurs de Relatedness et de Closeness sont très faibles (Fig. 2). Ces individus sont dits *marginaux* et il est intéressant de calculer les nouvelles distributions conceptuelles obtenues après les avoir éliminés (c'est-à-dire après avoir supprimé la bulle située en bas à gauche). Le résultat est présenté sur les Figures 3 et 4. Après filtrage, la distribution conceptuelle de *DailyMotion* (Fig. 3) ne contient plus la « grosse bulle » d'éléments marginaux, alors que la distribution conceptuelle de Flickr5 (Fig. 4) comporte à nouveau des éléments marginaux. Si l'on poursuit ce processus de filtrage sur *Flickr5* en recalculant la distribution conceptuelle après avoir éliminé la bulle en bas à gauche, il reste encore des éléments marginaux. L'échantillon *Flickr5* est intrinsèquement hétérogène et contiendra toujours des éléments marginaux, quel que soit le nombre de filtrages effectués.

Le processus de filtrage a été automatisé sur le critère des valeurs de Relatedness et de Closeness moyennes, en éliminant des ensembles de données d'origine, les objets dont les valeurs de Relatedness et de Closeness sont inférieures aux valeurs de Relatedness et de Closeness du système (moins α *écart type, où α peut varier).

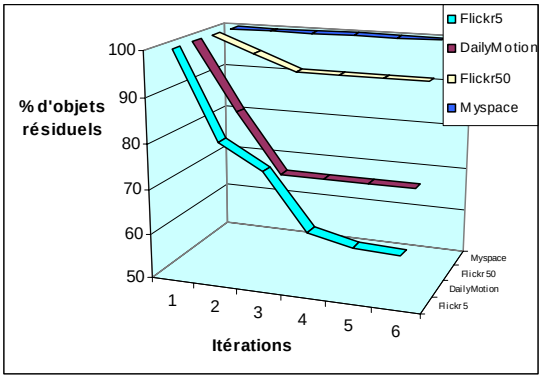

Fig. 5. Comparaison des résultats du filtrage conceptuel

Les paramètres de Relatedness et Closeness sont recalculés pour tous les objets du nouvel échantillon, et les éléments marginaux de cet ensemble sont éliminés, et ainsi de suite. L'algorithme de filtrage converge lorsque l'ensemble de données ne contient plus d'éléments marginaux, i.e. lorsque les valeurs de Relatedness et de Closeness de tous les objets résiduels sont suffisamment homogènes. La Figure 5 compare les résultats obtenus avec ce filtrage conceptuel automatique appliqué aux quatre échantillons étudiés, avec α=1,15. Comme l'on pouvait s'y attendre, le filtrage reposant sur des paramètres conceptuels affecte beaucoup l'échantillon *Flickr5* et confirme qu'il reste toujours des éléments marginaux après chaque étape de filtrage. A l'inverse, aucun membre du réseau *Myspace* n'est éliminé de l'échantillon initial.

4 Conclusion

Cet article a présenté une méthode conceptuelle pour l'analyse et l'exploitation de treillis de Galois. Cette caractérisation conceptuelle repose sur des statistiques calculées pour chaque objet en fonction des extensions et des intensions des concepts du treillis, sous la

forme de paramètres conceptuels appelés *Relatedness* et *Closeness*. Ces informations permettent de caractériser les ensembles de données en terme d'homogénéité ou d'hétérogénéité et de réaliser un filtrage automatique des objets dits marginaux. La méthode proposée a été expérimentée sur quatre échantillons de réseaux sociaux afin d'illustrer son fonctionnement et ses résultats.

Références

Barbut, M., Monjardet, B., Ordre et classification, Algebre et combinatoire, Tome 2, Hachette, 1970.

Batagelj, V., Mrvar, A.: Pajek - Analysis and Visualization of Large Networks. in Jünger, M., Mutzel, P., (Eds.) Graph Drawing Software. Springer, Berlin 2003. p. 77-103

Birkhoff, G., Lattice Theory, First Edition, Amer. Math. Soc. Pub. 25, Providence, R. I., 1940.

Brusilovsky P. Social Information Access: The Other Side of the Social Web, Proceedings of SOFSEM 2008, 34th International Conference on Current Trends in Theory and Practice of Computer Science, High Tatras, Slovakia, January 19-25, 2008 (Invited Talk).

Carpineto, C., Romano, G., Galois: An order-theoretic approach to conceptual clustering, Proc. Of the 10th Conference on Machine Learning, Amherst, MA, Kaufmann, pp. 33-40, 1993.

Godin, R, Chau, T.-T., Incremental concept formation algorithms based on Galois Lattices, Computational intelligence, 11, n° 2, p246 –267, 1998.

Jay, N., Kohler, F. and Napoli, A: Analysis of Social Communities with Iceberg and Stability-Based Concept Lattices. ICFCA 2008: 258-272.

Newman M.E.J., The structure and function of complex networks, SIAM Review 45, pp. 167-256, 2003.

Wasserman S., and Faust K., Social Network Analysis: Methods and applications. Cambridge University Press, Cambridge, UK, 1994.

Wille, R., Line diagrams of hierarchical concept systems, Int. Classif. 11, pp. 77-86, 1984.

Wille, R., Concept lattices and conceptual knowledge systems, Computers & Mathematics Applications, 23, n° 6-9, pp. 493-515, 1992.

Summary

In this paper, Formal Concept Analysis and Galois lattices are used for the analysis of complex datasets, online social networks in particular. Lattice-inspired statistics computed on the objects of the lattice provide their "conceptual distribution". An experimentation conducted on four social networks' samples shows how these statistics may be used to characterize these networks and filter them automatically.