

# **Le web sémantique est-il soluble dans le web2.0 ?**

## **(Fouille de texte versus Fouille de communauté)**

Bernard Rothenburger, Nacim Chikhi, Nathalie Aussenac-Gilles

IRIT/INRIA, UPS 118 route de Narbonne, 31062 Toulouse Cedex  
{rothenbu,chikhi,nathalie.aussenac-gilles}@irit.fr

Si le Web2.0 est social il doit être possible de construire des structures sociales organisant des individus qui le composent. Le Web sémantique quant à lui, organise des descriptions du monde en structurant des concepts et des relations. Dans la lignée des travaux de Peter Mika (2005), nous nous interrogeons sur la possibilité d'associer des outils d'analyse de réseaux sociaux et des outils de structuration de concepts. Les premiers travaux que nous décrivons ici prennent en compte des liens sociaux dont les traces sont des liens entre documents, l'organisation des concepts se réduit à des clusters de termes. Les résultats obtenus montrent à la fois la dépendance entre communautés sociales et thématiques et l'intérêt à utiliser la combinaison des deux aspects.

Les comportements sociaux des producteurs de connaissances peuvent être tracés à partir des liens qu'ils expriment entre leur production. Concernant le domaine du Web les liens hypertextes entrent dans cette catégorie. Concernant les publications scientifiques les citations de travaux sont aussi des liens entre producteurs de connaissances.

La fouille de communauté est un moyen couramment utilisé pour caractériser la structure sociale de producteurs de connaissances en se basant sur l'analyse de ces liens. Elle a au moins deux buts. Le premier est construire des communautés de producteurs, c'est-à-dire en première hypothèse, des ensembles de producteurs qui se réfèrent plus à l'intérieur de la communauté qu'elle ne produit de référence à l'extérieur. Le second est de hiérarchiser les éléments d'une communauté par degrés d'autorité (ou de popularité). Mais, si les communautés sont composées par des individus elles sont aussi caractérisables par la production de ces individus c'est-à-dire par les ensembles des textes (des pages web, des blogs,...) qu'ils ont créés. On peut alors utiliser ces productions pour caractériser les préoccupations des communautés et donc les contenus sémantiques qu'ils véhiculent.

Le travail que nous allons présenter a d'abord consisté à définir des outils d'analyse de données adaptés à l'analyse de liens. Des éléments comme l'interprétabilité et la qualité des résultats sont à prendre en compte, la possibilité de s'affranchir du caractère exclusif de l'appartenance à une communauté est aussi

indispensable. Nous avons proposé un algorithme d'identification de communauté qui respecte au mieux ces contraintes : NHITS (Chikhi et al., 2008).

Pour valider ce travail nous avons confronté les communautés identifiées (décrites par leur textes les plus représentatifs) à des corpus classés par des experts sur des critères de contenus sémantiques. Cette validation montre que les classifications obtenues sont, dans certains cas, quasiment aussi bonnes que celle que l'on obtient en utilisant les méthodes classiques de fouilles de textes à partir des contenus.

L'idée que nous avons alors explorée est de prendre simultanément en compte les comportements sociaux des producteurs de connaissances et les contenus sémantiques de leur productions. Plusieurs manières de procéder seront présentées. Elles montrent que l'on peut ainsi obtenir des résultats à la fois supérieurs à ceux obtenus par la seule analyse de liens mais aussi à ceux obtenus par la seule fouille de contenus.

L'état d'avancement de travaux que nous avons menés donne clairement des arguments pour l'exploitation des caractéristiques sociales des communautés pour appuyer la caractérisation sémantique de leur production. Ce travail ouvre des perspectives dans les directions sociales et sémantiques. Dans la direction sociale, les liens que nous exploitons sont particuliers dans la mesure où les liens sous-tendent des rapports sémantiques. Il conviendra de prendre en compte d'autres réseaux sociaux pour lesquels les liens sont différents. Dans la direction sémantique les caractérisations aboutissent à des ensembles non-structurés de termes. Il conviendra d'aller plus loin.

## Références

MIKA, P. (2005). . "Ontologies are Us : a Unified Model of Social Networks and Semantics". In ISWC, volume 3729 of LNCS, p. 522–536 : Springer.

CHIKHI, N.F., ROTHENBURGER, B., AUSSENAC-GILLES, N. (à paraître). "*Authoritative Documents Identification based on Nonnegative Matrix Factorization*". Dans : *IEEE International Conference on Information Reuse and Integration (IRI 2008), Las Vegas (USA), 13/07/08-15/07/08*, IEEE.