Sujet de Maîtrise

Titre : « *Extraction de patrons dans la rédaction collaborative sur les Wikis et les Wikiblogs* »

Direction :

Esma Aïmeur, prof. Aggrégée DIRO (co-directrice)

Petko Valtchev, prof. Adjoint DIRO (directeur)

Problématique

Un wiki est un site Web qui permet à tout utilisateur d'éditer directement son contenu sur les pages même du site. Le concept de wiki a été introduit par Ward Cunningham en 1995. L'idée a pris son essor au début de l'année 2000, et plusieurs outils de création de wikis ont été développés. De nos jours, le plus grand wiki est Wikipedia[1] [Remy 2002], une encyclopédie universelle collaborative multilingue contenant plus d'un million de pages.

Les informations publiées sur les wikis sont fréquemment mises à jour par la communauté. C'est cet aspect communautaire et collaboratif qui fait la spécificité des wikis [Leuf et Cunningham 2001]. Cependant les mécanismes de collaboration sont encore mal compris.

Cadre

Dans le cadre d'un projet conjoint avec le département de communication, notre équipe cherche à mettre en évidence des manifestations de ces mécanismes à partir des journaux ou logs des sites wiki. Ceux-ci documentent les interventions des divers participants dans le processus de rédaction collaborative. Par la suite, les résultats de la recherche alimenteront une réflexion conjointe avec les chercheurs en communication quant aux mécanismes effectifs et leur interprétation.

Approche

Dans un premier temps, il s'agit d'organiser les informations des logs dans des structures explicitant les rapports entre les interventions individuelles des participants, puis d'analyser ces dernières pour en extraire des *régularités* à divers niveaux de granularité. Sur la base de l'ensemble de régularités découvertes, nous formulerons des hypothèses sur les modes de

[1] http://www.wikipedia.org

collaboration propres au wikis.

Plus spécifiquement, il faudra examiner les historiques des sites et extraire les interventions de chaque participant, puis constituer des structures de graphes des interventions reliées. Dans un tel graphe, les nœuds contiendront les textes des contributions et seront annotés par divers types de méta-données comme la date et l'auteur (alias). Les nœuds seront connectés par des liens de type ajout/suppression/reformulation. Les données factuelles, organisées dans de tels graphes, constitueront l'entrée du processus d'analyse dont le but sera la découverte de nouvelles régularités observables.

Dans un deuxième temps, les données seront étudiées selon plusieurs axes d'analyse complémentaires et à l'aide de méthodes issues du *information extraction* [Jacobs 1992] et du *data mining* [Han et Kamber 2001] :

1- Tout d'abord, les divers types d'interventions des participants seront catégorisés par des algorithmes de *clustering* (*numérique* [Jain et Dubes 1988], *conceptuel* [Ganter et Wille 1999]). afin d'en extraire une taxonomie. Un espace de description doit être déterminé au préalable pour ces interventions

2- Les catégories identifiées seront utilisées pour annoter, de façon automatique ou semi-automatique, les structures d'interventions tirées des historiques. Le résultat sera ensuite analysé à l'aide de techniques de *association rule mining* [Agrawal *et al.* 1993] afin d'en extraire des patrons de collaboration, c'est-à-dire, des configurations locales de diverses catégories d'interventions qui apparaissent fréquemment dans les structures globales. Étant donné la structure graphique et donc relationnelle des données initiales, il s'agira d'appliquer des techniques récentes de fouille de graphes et d'arbres (*voir bibliographie*).

3- En parallèle, la description des structures ou des patrons qui en sont extraits pourrait être enrichie par d'autres informations. Par exemple, des renseignements supplémentaires à propos des auteurs des interventions (intérêts, compétences, etc.) pourront être injectées. À cette fin, on pourrait chercher à identifier la communauté autour d'un wiki à l'aide de techniques de *link analysis* [Thelwall 2004].

Résultats attendus

- Une taxonomie des interventions sur les outils étudiés
- Une liste de patrons de collaboration inhérents à l'usage des wiki et/ou wikiblogs
- Une méthodologie pour l'analyse des collaborations sur un wiki.

Exigences

- Bonnes connaissances en programmation en Java et/ou C++
- Autonomie de travail
- Motivation pour le développement d'un système de *data mining*

Des connaissances en traitement de langue naturelle et en analyse de données (ou *data mining*) seraient un atout. La durée *maximale* de la collaboration est du 1er septembre 2005 au 31 décembre 2006. Financement possible pour deux à trois sessions.

Les étudiant(e)s intéressé(e)s doivent envoyer leur curriculum vitae à l'un des co-responsables. Il est très souhaitable que les candidats aient terminé leurs cours de Maîtrise.

Contact

Esma Aïmeur (esma.aimeur@umontreal.ca)

Date-limite du dépôt des candidatures : 8 août 2005

Bibliographie

Générale :

R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. *Proc. of the ACM SIGMOD'93*, pages 207– 216, 1993.

B. Ganter and R. Wille. *Formal Concept Analysis, Mathematical Foundations*. Springer-Verlag, 1999.

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

N. Jain and R. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, 1988.

B. Leuf and W. Cunningham, *The Wiki Way: Quick Collaboration on the Web*. Boston: Addison-Wesley, 2001

M. Remy, Wikipedia: The Free Encyclopedia. *Online Information Review*. v.26, n.6, p.434.

M. Thelwall, *Link Analysis: An Information Science Approach*, Academic Press, 2004.

Techniques de fouille de graphes et d'arbres :

Y. Chi, Y. Yang, and R. R. Muntz. Indexing and mining free trees. *Proc. of the International Conference on Data Mining (ICDM'03)*, 2003.

Y. Chi, Y. Yang, and R. R. Muntz. Cmtreeminer: Mining both closed and maximal frequent subtrees. *Proc. of the 8th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'04)*, 2004.

M. Cohen and E. Gudes. Diagonally subgraphs pattern mining. *Proc. of the 9th ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge Discovery*, 2004.

J. Huan, W. Wang, J. Prins, and J. Yang. Spin: Mining maximal frequent subgraphs from graph databases. *Proc. of the 10th Intl. Conf. on Knowledge Discovery and Data Mining (KDD'04)*, pp. 581–586, 2004.

A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. *Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD '00)*, pages 13–23, 2000.

X. Yan and J. Han. gspan: Graph-based substructure pattern mining. *Proc. of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, page 721, 2002.

X. Yan and J. Han. Closegraph: Mining closed frequent graph patterns. *Proc. of the 9th Intl. Conf. on Knowledge Discovery and Data Mining (KDD'03*), pages 286–295, 2003.

M. J. Zaki. Efficiently mining frequent trees in a forest. *Proc. of the 8th Intl. Conf. on Knowledge Discovery and Data Mining (KDD'03*), Edmonton, Alberta, Canada, pages 71–80, 2002.