Extraction de motifs séquentiels et connaissances pour une meilleure compréhension des usages d'un système d'information

- Sujet non encore financé.

- Candidatures en vue du montage d'un dossier de demande de bourse.

- Responsables de thèse : Florent Masseglia et Brigitte Trousse

Ce projet se situe dans le contexte d'un site Web qui met un moteur de recherche à la disposition de ses utilisateurs. Nous avons l'intention d'améliorer la compréhension des usages sur un tel site, à partir d'informations récoltées par différentes phases de data mining. Les algorithmes mis en œuvre reposeront sur des techniques d'extraction de motifs séquentiels (détection d'enchaînements fréquents d'événements, selon un support minimum, dans un ensemble d'enregistrements).

Ce sujet se situe au croisement de deux domaines extrêmement étudiés : le Web mining et le Web sémantique.

On peut distinguer dans le Web Mining trois domaines. Le Web Content Mining qui vise à extraire des informations relatives au contenu des pages d'un site Web (par exemple une classification des documents d'un site d'information qui aboutirait à des classes de documents sur la finance, sur les attentats, sur le chômage, etc.). Le Web Structure Mining qui consiste à analyser la façon dont les pages sont écrites (déterminer par exemple la fréquence d'utilisation de tels ou tels tags du langage HTML ou XML). Enfin le Web Usage Mining qui sert à l'analyse de comportement des utilisateurs d'un site Web (on peut par exemple déterminer les navigations les plus fréquentes afin d'améliorer le site ou le rendre adaptatif).

Le Web Sémantique est basé sur la vision de Tim Berners-Lee, l'inventeur du WWW. Le succès du Web apporte un nouveau challenge : un énorme quantité de données, interprétable par les humains seulement et qui rend le support machine limité. Berners-Lee suggère d'enrichir le Web avec des informations compréhensible par la machine afin d'aider l'utilisateur. Pour se convaincre de l'intérêt d'un tel domaine, il suffit de constater le taux de réponses inutiles renvoyées par un moteur de recherche pour la plupart des requêtes.

Dans [1, 11], nous trouvons un état de l'art intéressant sur les liens à explorer entre l'analyse des usages (Web Usage Mining) et le Web Sémantique. En particulier les interactions qui peuvent exister entre les deux domaines sont à double sens et méritent des travaux plus précis.

Extraction d'informations sur le site Web

Dans un premier temps, nous envisageons de travailler sur la découverte de structures dans divers types de documents. Notre objectif est de prendre en compte des formats de données différents comme les documents HTML ou XML, les fichiers propriétaires (log d'un système d'information, données générées par des capteurs, etc.) ou d'autres documents issus du Web.

Si des travaux existent pour analyser la structure d'un document (par l'inférence grammaticale [3, 4, 5] ou des algorithmes d'extraction [6]) peu se sont consacrés à l'analyse d'un ensemble de documents pour en extraire une structure commune avec un support minimum à partir de motifs séquentiels [7, 8]. Notre approche consistera de plus à utiliser ces structures pour proposer une classification des documents en fonction de leur organisation interne. Dans [9] nous avons en effet proposé une méthode de classification des utilisateurs d'un site Web en fonction de leurs navigations à partir des motifs séquentiels extraits sur le log. Une technique similaire doit être étudiée pour proposer une classification des documents en fonction de leur structure.

Dans la suite de ce document nous désignons par " ontologie " une hiérarchie de concepts sur les pages Web (par exemple la page " axis.html " peut être généralisée par le concept " analyse des usages " puis " systèmes d'information "). Dans [10] des pistes sont données pour l'aide à la construction d'ontologie des pages d'un site Web en se basant sur les usages et dans [11] les auteurs proposent une méthode pour améliorer un moteur de recherche à l'aide d'une ontologie. En fonction de l'avancement des travaux précédant, nous envisageons d'utiliser les usages d'un moteur de recherche afin de construire une ontologie sur le site. De tels travaux seront novateurs et les résultats permettront d'aider l'extraction d'ontologie à partir des usages.

Apport de ces informations sur une analyse des usages

Si des travaux [10, 12] existent sur la possibilité d'exploiter des informations sémantiques sur le site Web afin d'analyser les usages d'un site, il n'existe pas à notre connaissance, de travaux qui intègrent ensemble toutes les informations décrites dans ce projet, lors de cette analyse. A partir des éléments obtenus par les travaux décrits précédemment, nous envisageons donc de mener une étude sur l'intégration des connaissances du domaine (ontologie sur le site, structure des pages, contenu des pages, classification des documents contenu par le site) dans le processus d'extraction des motifs séquentiels destiné à la compréhencion des usages. Pour cela nous étudierons la possibilité de caractériser chaque page par des éléments de (entre autres) :

- sa structure,
- son contenu,
- sa position dans l'ontologie,

- sa classe parmi les autres pages.

Une fois cette caractérisation effectuée, nous proposerons des méthodes d'extraction de motifs séquentiels qui prennent en compte tous ces éléments afin de fournir des résultats qui soient compréhensibles par l'utilisateur. Nous espérons ainsi augmenter la pertinence et l'utilité des résultats de l'analyse des usages.

Bibliographie :

[1] Berendt, B., Stumme, G., & Hotho, A. (in press). *Usage mining for and on the Semantic Web*. In H. Kargupta, A. Joshi, K. Sivakumar, & Y. Yesha (Eds.), Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA: AAAI/MIT Press.

[2] Srikant, R., & Agrawal, R. (1996). *Mining sequential patterns: Generalizations and performance improvements*, Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT). Avignon, France.

[3] T. W. Hong and K. L. Clark. *Using grammatical inference to automate information extraction from the web*. In Principles of Data Mining and Knowledge Discovery, pages 216-227, 2001.

[4] V. Crescenzi, G. Mecca, and P. Merialdo. *Roadrunner: Towards automatic data extraction from large web sites.* Technical Report n. RT-DIA-64-2001, D.I.A., Universit a di Roma Tre, 2001.

[5] C. H. Chang, S.C. Lui. *IEPAD: Information Extraction Based on Pattern Discovery*. Proc.of 10 World Wide Web Conference, Hong Kong, pages 681-688, 2001.

[6] Arvind Arasu, Hector Garcia-Molina. *Extracting structured data from Web pages*. Proceedings of the 2003 ACM SIGMOD international conference on on Management of data table of contents, San Diego, California, pages 337-348, 2003.

[7] N. Vanetik, E. Gudes, S. E. Shimony. *Computing Frequent Graph Patterns from Semistructured Data.* 2002 IEEE International Conference on Data Mining (ICDM'02), December, 2002, Maebashi City, Japan

[8] Pierre-Alain Laur, Florent Masseglia, Pascal Poncelet. *Schema Mining: Finding Structural Regularity among Semistructured Data*. 498-503 Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France, September, 2000.

[9] Florent Masseglia, D. Tanasa, Brigitte Trousse. *Web Usage Mining: Sequential Pattern Extraction with a Very Low Support*. APWeb 2004: 513-522. 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April, 2004

[10] Bettina Berendt, Andreas Hotho, Gerd Stumme. *Towards Semantic Web Mining*. Proceedings of the First International Semantic Web Conference on The Semantic Web. Pages 264 – 278, 2002.

[11] S. Parent, B. Mobasher, and S. Lytinen. *An Adaptive Agent for Web Exploration*

Based of Concept Hierarchies. Proceedings of the 9th International Conference on Human Computer Interaction New Orleans, August 2001.

[12] Honghua (Kathy) Dai, Bamshad Mobasher. *A Road map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining*. Proceedings of the International Conference on Internet Computing, IC '03, Las Vegas, Nevada, USA, June 23-26, 2003.