Extraction automatique d'information inattendue à partir de textes.

François Jacquenet, Christine Largeron

Université Jean Monnet de Saint-Etienne
EURISE
23 rue du docteur Paul Michelon
42023 Saint-Etienne Cedex 2
Francois.Jacquenet@univ-st-etienne.fr
Christine.Largeron@univ-st-etienne.fr

Résumé. Dans cet article, nous proposons d'utiliser des techniques de fouille de textes pour extraire des informations, automatiquement et à des fins stratégiques, à partir de bases de données scientifiques et techniques. Ce contexte de veille technologique introduit une difficulté inhabituelle par rapport aux domaines d'application classiques de la fouille de textes, puisqu'au lieu de rechercher de la connaissance fréquente cachée dans les données, il faut rechercher de la connaissance inattendue, qualifiée par les veilleurs de signal faible. Les mesures usuelles d'extraction de la connaissance à partir de textes doivent, de ce fait, être revues. Pour ce faire, nous avons développé le système *UnexpectedMiner* dans lequel de nouvelles mesures permettent d'estimer le caractère inattendu d'un document. Notre système est évalué sur une base de résumés d'articles scientifiques.

1 Introduction

Du fait de l'augmentation croissante des capacités de production et de stockage des données, des travaux ont porté tout d'abord sur le développement de méthodes et d'algorithmes permettant de les analyser et d'en extraire automatiquement des connaissances utiles. Mais rapidement, en plus du problème du volume des données c'est celui de leur diversité et de leur hétérogénéité qui a suscité l'intérêt des chercheurs. L'enjeu est en effet, de sortir du carcan tabulaire à n lignes (les individus) et p colonnes (les attributs associés à des types standards de données : booléen, nominal ou ordinal, réel) pour répondre aux besoins de nouvelles applications. Parmi les formes de données considérées on peut citer les données multi-relationnelles, issues de plusieurs tables d'une base de données relationnelles, qui ont conduit à la conception de Data Warehouses et de systèmes OLAP (*On-Line Analytical Procesing*). On peut aussi évoquer les données transactionnelles, où chaque enregistrement d'un fichier est formé d'un identifiant et d'un ensemble d'items (dont l'exemple type est le panier d'une ménagère) et, qui ont pu être traitées efficacement à l'aide d'algorithmes d'extraction de règles d'association. On peut mentionner également les données séquentielles décrivant l'évolution dans le temps d'un phénomène, et qui ont été étudiées par différentes communautés, par exemple en apprentissage automatique à l'aide d'automates ou en statistiques par des modèles stochastiques et régressifs. Sans parler de données encore plus complexes