Prise en compte de la structure des documents pour la découverte d'informations inattendues*

Using the structure of documents to discover unexpected information

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

Résumé

La découverte automatique d'informations inattendues ou nouvelles dans des documents textuels est une tâche difficile mais particulièrement intéressante compte tenu de ses applications potentielles. Il peut s'agir par exemple de repérer parmi les questions adressées à des listes de diffusion celles qui n'ont pas encore été répertoriées dans les foires aux questions (FAQ), ou bien de trouver parmi des nouvelles publiées dans la presse celles qui traitent d'un nouveau sujet ou encore, dans le cadre de la veille technologique, on peut identifier des signaux faibles dans des bases d'articles scientifiques et techniques, de brevets, etc. Dans ces différentes applications, le but est de repérer des informations inattendues en ce sens qu'elles étaient inconnues auparavant de l'utilisateur. Faisant suite à un premier travail visant à concevoir et implanter des mesures d'inattendu dans un système baptisé UnexpectedMiner, nous avons cherché à améliorer les performances de celui-ci en prenant en compte la structure des documents analysés. Chaque partie des documents est ainsi pondérée par des coefficients dont les valeurs sont déterminées par un algorithme d'optimisation. Ces coefficients sont alors intégrés dans les mesures d'inattendu utilisées par UnexpectedMiner pour déterminer si un document présente un caractère inattendu ou pas. Les performances de notre nouveau système ont été évaluées sur un corpus d'articles scientifiques et mettent en évidence les améliorations induites par la prise en compte de la structure des documents.

Mots Clefs

Fouille de textes, Information inattendue, Recherche d'information, Structure du document.

Keywords

Text mining, Unexpected information, Information retrieval, Structure of documents.

*Ce travail a été partiellement soutenu par le projet BINGO de l'ACI Masses de Données 2004-2007, financé par le ministère de la recherche

1 Article de vulgarisation scientifique

Alors même que les enjeux liés à la maîtrise de l'information venaient d'être identifiés, l'augmentation croissante des capacités de production et de stockage des données rendaient l'accès à cette information plus difficile. C'est probablement ce qui a conduit au développement de la fouille de données, depuis le milieu des années 1990. La fouille de données a en effet été définie comme "l'extraction non triviale, à partir de données, d'une information potentiellement utile, implicite et inconnue auparavant" (Fayyad et al., 1996). Lorsque les données considérées se présentent sous la forme de textes, qu'ils soient structurés ou non, on parle alors de fouille de textes. Par analogie avec la fouille de données, la fouille de textes, introduite en 1995 par Feldman (Feldman et al. 1995), est définie par Sebastiani (Sebastiani, 2002) comme l'ensemble des tâches qui, par analyse de grandes quantités de textes et la détection de modèles fréquents, essaie d'extraire de l'information probablement utile. Ainsi, en faisant appel à des algorithmes d'extraction de motifs séquentiels fréquents (Agrawal et al. 1995), les premiers travaux réalisés ont cherché à extraire des informations qui apparaissent fréquemment dans les textes. A l'opposé, d'autres recherches se sont focalisées sur la découverte de ce qui est appelé, selon les auteurs, des informations inattendues ou nouvelles, des événements rares ou encore des sujets émergents.

La découverte automatique de telles informations, qui en général n'apparaissent pas avec une fréquence élevée dans les documents textuels, est une tâche difficile mais particulièrement intéressante compte tenu de ses applications potentielles. Il peut s'agir par exemple de repérer parmi les questions adressées à des listes de diffusion celles qui n'ont pas encore été répertoriées dans les foires aux questions (FAQ), ou bien de trouver, parmi des nouvelles publiées dans la presse celles qui traitent d'un nouveau sujet, ou encore, dans le cadre de la veille technologique, on peut identifier des signaux faibles dans des bases d'articles scientifiques et techniques, de brevets, *etc.*

Parmi les premiers travaux portant sur ce sujet, on peut citer le programme *TDT* (*Topic Detection and Tracking*) lancé par la *DARPA* dès 1996 dans le but d'identifier de nouveaux événements dans un flux de nouvelles journalistiques (Allan et al., 1998, Wayne, 1998). Plus récemment, un challenge a été organisé sur le thème de la détection de nouveauté (*Novelty detection*) dans le cadre de la conférence *TREC* (Soboroff et al., 2003). Toutefois, la liste des documents fournis, qu'il s'agisse des phrases pour *TREC* ou des nouvelles journalistiques pour *TDT*, est triée par ordre chronologique, ce qui n'est pas le cas des documents considérés dans un grand nombre d'applications. De plus, les corpus traités sont souvent composés de textes courts tels que des phrases ou des requêtes. Enfin, la plupart des travaux consacrés à la recherche d'informations nouvelles dans des textes, considèrent uniquement le contenu des documents et rares sont ceux qui exploitent également leur structure. Or les études récentes menées en recherche d'information (Fourel, 1998, Piwowarski, 2003) ont montré l'intérêt de prendre en compte ces deux types d'information. Dans le cadre de la recherche de nouveauté, la prise en compte de la structure paraît également justifiée car on peut supposer que toutes les parties d'un document n'ont pas le même poids et qu'il en est de même pour les termes apparaissant dans ces parties. De plus, les documents considérés sont souvent fortement structurés.

C'est pourquoi, suite à un précédent travail (Jacquenet et al. ,2004) visant à concevoir et implanter des mesures de recherche d'information inattendue dans un système baptisé UnexpectedMiner, nous avons cherché à améliorer les performances de celui-ci en prenant en compte la structure des documents analysés. Ce système vise à extraire, de corpus documentaires, des documents pertinents pour l'utilisateur en ce sens qu'ils contiennent des informations inconnues auparavant de celui-ci et se rapportant au sujet qui l'intéresse. Il est bien adapté pour la recherche de signaux faibles dans le cadre de la veille scientifique et technique car il traite des textes intégraux. Il est articulé autour de trois modules et il requiert deux ensembles de documents. Le premier est composé d'une dizaine de documents de référence, fourni par l'utilisateur et qui permettent de cibler le sujet qui l'intéresse. Le second correspond à l'ensemble des nouveaux documents, issus de différents corpus, susceptibles de contenir des informations inattendues et intéressantes pour l'utilisateur. Les deux ensembles de documents vont subir un prétraitement à l'issue duquel chaque document est représenté sous une forme matricielle. Le but du second module est d'extraire de la base des nouveaux documents, ceux qui sont les plus similaires aux documents de référence, à l'aide de la distance du *cosinus*. Enfin, dans cet ensemble de nouveaux documents jugés similaires, le troisième module, vise à rechercher ceux qui contiennent des informations inattendues par rapport à celles figurant dans les documents de référence à l'aide de mesures que nous avons proposées (Jacquenet, 2005). Les performances de ce système ont été évaluées sur un corpus d'articles scientifiques et mettent en évidence les améliorations induites par la prise en compte de la structure des documents.