Text Mining

Jérôme Champavère Didier Devaurs Kaouther Drira
Nawal Guermouche Mohammed Toukourou Meriem Zidouni

15 novembre 2005

L'analyse et le traitement d'informations textuelles sont devenus un enjeu majeur avec l'explosion du Web : environ 90% de l'information accessible l'est sous forme textuelle (bibliothèques électroniques, pages HTML, forums de discussion, etc.). Cependant les tâches d'exploration et de récupération de l'information dans ces réservoirs de connaissances deviennent extrêmement ardues.

Face à ce problème, la fouille de texte, ou *text mining*, vise à faciliter l'extraction des connaissances "cachées" dans les documents. Ce domaine de recherche tente de mettre à profit la surabondance d'informations textuelles en utilisant des techniques d'informatique linguistique, de *data mining*, d'apprentissage automatique et de statistiques.

Qu'est-ce que le *text mining* ?

Le *text mining* est définit comme le « *processus non trivial d'extraction d'informations implicites, précédemment inconnues, et potentiellement utiles, à partir de données textuelles non structurées dans de grandes collections de textes* » [1]. Le *text mining* comprend la succession des étapes suivantes, permettant de passer de documents textuels à un ensemble de données formalisé.

Recherche de documents. L'étude à effectuer porte toujours sur un ensemble de textes ayant une caractéristique commune. Cette caractéristique peut concerner le sujet, l'auteur, l'année de publication, etc. La première étape consiste donc à effectuer une simple recherche au sein des ressources disponibles, en général à partir du Web et de bases de données bibliographiques ou textuelles, pour trouver les documents ayant cette caractéristique. Le résultat de la recherche constitue le contexte de fouille.

Structuration des données. L'ensemble de documents à fouiller n'est pas structuré (du moins au sens informatique, car le contenu des documents possède bien sûr une structure sémantique). Mais, de nombreuses méthodes, essentiellement issues du domaine du traitement automatique des langages, permettent de pallier à ce problème. La plus couramment utilisée consiste à rechercher, et si nécessaire à filtrer, les mots-clés ou les phrases-clés contenus dans les documents, et éventuellement les relations existant entre ces divers éléments clés. Mais, il est également possible de faire des résumés des documents, d'effectuer une catégorisation ou du *clustering* (la différence étant que pour la catégorisation, les différentes catégories possibles sont connues à l'avance, et pas pour le *clustering*).

Exploration des données structurées. Les outils utilisés à ce niveau proviennent pour la plupart du monde de la fouille de données classique. Cette étape repose sur une forte interaction entre l'utilisateur et le système. La machine apporte la puissance de calcul et les capacités de mémorisation. L'utilisateur est le seul à pouvoir dominer l'aspect sémantique des résultats. Néanmoins, le système agit comme un assistant, dans le sens où il est capable de faire des suggestions à l'utilisateur et de prendre des initiatives, tout en justifiant ses choix.

Un système respectant ces critères permettrait de quantifier un texte ou les parties d'un texte pour en extraire les structures signifiantes les plus fortes, établir des liens entre les termes et les documents, et établir des règles de classification automatique de documents.

Mise en œuvre des concepts de fouille de textes

Le projet LINDI (*Linking Information for Novel Discovery and Insight*) permet d'illustrer chacune des étapes du processus de fouille de textes décrites précédemment [2].

Les objectifs du projet LINDI, rebaptisé aujourd'hui BioText, étaient de déterminer comment les chercheurs pourraient utiliser de grandes collections de textes pour la découverte de nouvelles informations importantes, et de mettre en place un système pouvant les y aider. Le contexte de mise en place de ce projet était la biologie moléculaire, et plus précisément la découverte automatique des fonctions des gènes nouvellement séquencés. Les chercheurs travaillant sur le génome étudient la façon dont les gènes s'expriment. En partant de gènes dont on connaît déjà l'expression et la fonction à laquelle ils participent, et en comparant leur expression à celle d'un gène nouvellement séquencé, il est possible d'en déduire des hypothèses concernant la fonction à laquelle participe ce nouveau gène. Par exemple, considérons deux gènes A et X, A étant un gène connu, et X un gène "nouveau" dont on aimerait connaître la fonction. Si l'on sait que le gène A intervient dans le développement du cancer de la prostate, et si l'on sait que l'expression du gène X est relativement similaire à celle du gène A, alors il y a de fortes chances que le gène X intervienne aussi dans le cancer de la prostate.

Pour pouvoir émettre des hypothèses concernant la fonction d'un gène, il suffirait donc d'explorer la littérature biomédicale, à la recherche de renseignements concernant les différents gènes, afin de pouvoir faire des recoupements d'informations. Mais le problème est que cette tâche est en pratique impossible à réaliser à cause de l'énorme quantité de documents publiés dans ce domaine. C'est donc à ce niveau qu'intervient la fouille de textes.

Les différentes étapes de la fouilles de textes apparaissent à travers le fonctionnement du système LINDI.

Recherche de documents. Dans le cadre de ce projet, les documents à traiter concernent tous la recherche biomédicale, et plus précisément l'étude du génome. Concrètement, on recherche sur le Web des articles mentionnant certains gènes, par exemple A, B et C, dont l'expression et la fonction à laquelle ils participent sont connues. Notons que cette recherche est évidemment restreinte à la langue anglaise.

Structuration des données. Afin de pouvoir effectuer des traitements sur les données contenues dans les documents obtenus lors de la recherche, celles-ci doivent être structurées. Pour ce faire, le système LINDI utilise la recherche de mots-clés. Plus précisément, cette tâche est effectuée au sein de chaque ensemble de documents traitant d'un même gène, en comptant le nombre d'occurrences de chaque mot. Parmi ces mots fréquents, il faut enlever les mots non signifiants (article, préposition, etc.) et les mots qui ne présentent pas d'intérêt pour l'utilisateur.

Exploration des données structurées. Les ensembles de mots-clés résultant de la phase précédente représentent une source exploitable par la machine, qui va alors calculer leur intersection. Les mots-clés trouvés *via* cette opération sont ordonnés par le système (par exemple en fonction du nombre d'occurrences), et sont ensuite présentés à l'utilisateur par le biais d'une interface graphique. Celui-ci choisit dans la liste proposée ceux qu'il trouve pertinents ; il peut également les réordonner. Il obtient alors un ensemble de mots-clés, éventuellement réduit, caractérisant l'expression des gènes de départ. Une nouvelle recherche est ensuite effectuée dans la littérature biomédicale afin de récupérer les documents mentionnant les trois gènes A, B et C, et au moins un des mots-clés les mieux classés (selon l'ordre défini par l'utilisateur). A partir de

cet ensemble de documents restreint, l'utilisateur pourra alors parfaire sa connaissance de l'expression des gènes A, B et C, et essayer de la mettre en relation avec ce qu'il sait de l'expression d'un gène X, nouvellement séquencé, et dont il veut connaître la fonction.

Un autre exemple qui permet de mieux se rendre compte du principe d'interaction qui peut être mis en place entre l'homme et la machine est celui du système AIDE (*Assistant Intelligent for Data Exploration*) [3]. Il s'agit d'un système de planification à initiative mixte : il poursuit ses buts de manière autonome tout en permettant à l'utilisateur d'intervenir dans ses décisions, et l'aide à s'orienter dans le paysage du processus d'exploration. Non seulement le système propose des voies d'exploration, mais en plus il justifie et ordonne ses propositions. Ainsi l'utilisateur contrôle tout le processus de recherche et d'analyse sans avoir nécessairement à diriger les initiatives prises par le système.

Un dernier exemple académique, illustrant, lui, les techniques de visualisation, est DocMiner (*Document Maps in INformation Elicitation and Retrieval*). DocMiner est un outil qui montre des cartes où les textes sont regroupés par régions en fonction de leur similarité. Il permet à l'utilisateur d'en analyser visuellement le contenu et d'interagir avec la carte de documents en changeant d'échelle.

Quelques domaines de recherche en fouille de textes

Le principe du projet LINDI repose sur la recherce de mots-clés fréquents dans un grand ensemble de documents. Dans ce genre de système, on considère qu'une information est potentiellement intéressante si elle apparaît fréquemment dans le texte. Cependant, il existe d'autres approches pour considérer qu'une information peut être pertinente. Par exemple, des recherches sont engagées dans la découverte d'informations inattendues, c'est-à-dire se situant en dehors d'une référence ou d'un modèle fixés par ailleurs. Une autre voie explorée par les chercheurs est la découverte de connaissances pouvant émerger d'un processus de classification de documents.

Classification automatique de documents. La plupart des outils d'accès à l'information ne donnent aucune explication sur les méthodes utilisées pour la classification des documents trouvés suite à une recherche. C'est pour remédier à cela qu'une branche de recherche en fouille de textes s'intéresse à apporter un peu de lumière sur la catégorisation d'un ensemble de documents selon leurs principaux thèmes. Les idées développées dans cette branche peuvent être illustrées à l'aide du système TileBars [4].

Il s'agit d'un outil qui a pour objectif de montrer à l'utilisateur sous forme de graphiques les relations existant entre les termes contenus dans sa requête et les documents trouvés. L'interface de TileBars requiert que l'utilisateur définisse sa requête par le biais d'une liste de thèmes. Par exemple, “médecine”, “médical” et “hôpital” peuvent tous être choisis pour la recherche de documents médicaux. Si tous les mots de la requête apparaissent de façon insistante dans un certain passage d'un document, alors ce document est d'une importance élevée. L'utilisateur voudra donc pouvoir se diriger directement au niveau de ce passage au lieu d'avoir à les lire tous. L'interface du système permet à l'utilisateur de jauger l'intérêt de chaque document et de ses passages constitutifs. TileBars permet donc de montrer simultanément et de façon compacte la longueur relative de chaque document trouvé suite à une recherche, la fréquence des mots de recherche au sein des documents et la distribution de ces mots.

Découverte de connaissance à partir d'informations inattendues. La nouvelle économie et avec elle la gestion croissante d'informations et de connaissances dans la vie des organisations sont des facteurs définissant un nouvel horizon pour la veille et l'intelligence économique. Ce nouveau contexte est favorable à un accroissement de la demande de fouille de textes et à une réorientation de la recherche. La veille introduit en effet une difficulté inhabituelle par rapport aux domaines d'application classiques des techniques de fouille de textes, puisque, au lieu de

rechercher de l'information fréquente cachée dans les données, il faut rechercher de l'information inattendue. L'idée d'une telle recherche provient d'une analogie que l'on peut faire avec la théorie de l'information. Selon cette théorie, une donnée apporte une quantité d'information d'autant plus grande qu'elle est inattendue.

Les algorithmes d'extraction de motifs séquentiels fréquents, employés habituellement en fouille de données, sont inappropriés pour effectuer de la veille. Des chercheurs en fouille de textes se concentrent actuellement dans la mise en place du système Unexpected Miner qui tente de répondre à cette problématique [5].

Ce système vise à extraire, de corpus documentaires, des documents pertinents pour le veilleur en ce sens qu'ils traitent de sujets inattendus et inconnus de celui-ci auparavant. Ce dernier définit un ensemble restreint de documents de référence. Le système consulte ensuite de nouveaux documents dans divers corpus et sélectionne ceux qui sont suffisamment similaires aux documents de référence. C'est dans cet ensemble de documents qu'il va chercher des informations ayant un caractère inattendu, c'est-à-dire absentes des documents de référence. La mesure de similarité constitue un pôle actif de la recherche dans ce domaine car il n'existe pas encore de méthode réellement efficace.

Conclusion

La fouille de textes est un domaine dont l'objectif est d'utiliser de grandes collections de documents pour découvrir de nouvelles informations ("découvrir" au sens de "expliciter"). La puissance des outils de fouille de textes repose essentiellement sur les interactions mises en place entre l'homme et la machine. En effet, ces interactions permettent de mettre à profit les compétences (complémentaires) des deux parties, à savoir la puissance de calcul et de mémoire pour la machine, et les capacités de raisonnement et d'interprétation pour l'homme. La mise en pratique de ces concepts a donc permis de résoudre la problématique liée au traitement de grandes quantités d'informations textuelles dans de nombreux domaines applicatifs. Cette réussite s'est traduite par un très fort engouement commercial et l'apparition de nombreux logiciels portant l'étiquette "*text mining*". Espérons que cet élan bénéficie également aux initiatives plus orientées vers la recherche.

Références

[1] Ibekwe-Sanjuan (F.) et Sanjuan (E.), « Ingénierie linguistique et fouille de textes », dans *Veille stratégique, scientifique et technologique (VSST'2004)*, 2004.

[2] Hearst (M.). « Untangling text data mining », 1999.

[3] Amant (R. S.) et Cohen (P. R.), « Interaction with a mixed-initiative system for exploratory data analysis », dans *Intelligent User Interfaces*, p. 15–22, 1997.

[4] `http://www.sims.berkeley.edu/~hearst/tb-overview.html`.

[5] Jacquenet (F.), Largeron (C.) et Chapaux (S.), « Veille technologique assistée par la fouille de textes », dans *Actes de la conférence Extraction et Gestion des Connaissances (EGC'2004)*, p. 429–440. Editions Cépaduès, Janvier 2004.