Emploi

"À ceux qui brassent de grandes quantités de documents numériques, le text mining offre des techniques pour indexer des documents autrement qu'à partir de mots clés et extraire du sens à partir des résultats."

Jérôme Azé, docteur en informatique, est maître de conférences et chercheur au Laboratoire de recherche en informatique (LRI), CNRS UMR 8623 - Université Paris-Sud. Sa thèse, soutenue en 2003, porte sur l'extraction de connaissances à partir de données numériques et textuelles. Il y propose une méthode permettant d'éviter de fixer le support minimal indispensable pour filtrer correctement les connaissances. Il organise, avec Mathieu Roche, l'atelier DEfi Fouille de Textes (DEFT) qui se tiendra lors la Semaine du document numérique du 18 au 22 septembre 2006, à Fribourg (Suisse).

À quoi sert le text mining ?

Le text mining permet d'extraire automatiquement des informations pertinentes et d'établir dans un corpus de textes des corrélations entre différentes informations en fonction de thèmes prédéfinis. Il permet aussi d'extraire des "régularités" sous forme de mots clés ou de termes appartenant à un domaine, et de découvrir les informations sans a priori. Si, par exemple, l'expression "carte bleue" apparaît fréquemment, on pourra savoir si le texte concerne le domaine bancaire ou celui du jeu.

Comment, dans ce cas, le système fait-il la distinction entre la carte bancaire et la carte de jeu ?

Sur la simple base de la fréquence des mots et de leur co-occurrence dans un ensemble de textes.

Les termes apparaissent automatiquement et leur sens découle de la fréquence de leur apparition, des mots auxquels ils sont associés, d'un étiquetage grammatical et de manière plus générale du contexte dans lesquels ils ont été observés.

Mais c'est un être humain qui définit ensuite, à partir des résultats obtenus, le domaine auquel l'expression "carte bleue" se réfère, et qui lui associe une étiquette sémantique.

Le text mining appartient encore au monde de la recherche...

Oui, mais les recherches dans ce domaine connaissent un véritable essor depuis ces dix dernières années alors que l'on parle de datamining depuis plus de trente ans. Les text miners sont très souvent, comme moi, des informaticiens qui travaillent en étroite collaboration avec des linguistes. Mais les objectifs des deux communautés

ne sont pas forcément les mêmes. Si les text miners veulent extraire des "régularités" dans un ensemble de textes pour progresser dans la compréhension globale d'un problème, les linguistes entendent extraire des informations leur permettant de définir des modèles linguistiques.

Plusieurs équipes travaillent dans ce domaine en France et ce secteur est appelé à se développer. Aujourd'hui, il est essentiellement orienté vers les domaines de la biologie ou de la médecine, soit vers des domaines où les publications sont nombreuses, où il y a beaucoup de connaissances à extraire et où l'impact financier est important. Comprendre, à partir des articles analysés, la corrélation qui peut exister, par exemple, entre les différents gènes présentés, aide à rédiger des articles mais aussi à orienter la recherche ou faciliter le dépôt de brevets.

Vous semblez privilégier la médecine et la biologie, alors que les applications en datamining se font surtout dans le domaine du commerce et de la gestion.

Le datamining et le text mining ne ciblent pas les mêmes domaines car les ressources ne sont pas les mêmes. Les données en très grand nombre ont effectivement des applications commerciales. Les résultats d'enquêtes, l'analyse des tickets de caisse permettent de gérer de manière optimale un stock, un magasin, d'organiser un rayon, de proposer les bonnes promotions, de gérer les têtes de gondoles. Le datamining permet aussi de cibler avec plus d'efficacité la publicité, notamment dans la téléphonie mobile où elle a permis de stabiliser les churners (ceux qui changent souvent de forfaits) en leur proposant une offre pertinente et ainsi éviter une publicité très coûteuse, car touchant un segment de population très limité.

Pour faire du text mining, il faut disposer d'un corpus important de textes sous une forme électronique. En médecine et en biologie, on a très souvent uniquement accès aux résumés du texte, soit à une ressource très standardisée dont le langage est clair, ce qui permet d'extraire facilement de l'information. Il serait plus difficile, voire impossible aujourd'hui, d'obtenir les même résultats à partir de sources plus compliquées comme des articles de presse qui font appel à des connaissances plus variées.

Mais c'est sans doute l'objet de vos recherches...

Oui. Nous aimerions bien atteindre ce niveau mais nous en sommes encore assez loin pour l'instant.

Est-ce que cela signifie que les outils de text mining sont réservés aux experts ?

La recherche est majoritairement orientée vers des secteurs spécialisés tels que les secteurs du médicament et des brevets. Mais elle concerne d'autres domaines : le résumé automatique, la veille économique, l'analyse automatique de courrier électronique (détection de spams, redirection de mails, réponse automatique) ou encore l'extraction d'informations pour construire automatiquement des formulaires (applications d'intelligence économique).

Aujourd'hui, la communauté du text mining est capable de proposer à ceux qui brassent de grandes quantités de documents numériques, des techniques leur permettant d'indexer des documents autrement qu'à partir de mots clés et d'extraire du sens à partir des résultats.

Quels résultats allez-vous présenter lors de la Semaine du document numérique à Fribourg ?

Avec Mathieu Roche (maître de conférences en informatique à l'université Montpellier 2), j'organise depuis l'année dernière le Défi francophone de fouille de textes (DEFT). Nous animerons donc le deuxième atelier de ce défi lors de la Semaine du document numérique.

L'objectif de ce deuxième défi consiste à développer des méthodes permettant de détecter automatiquement des ruptures thématiques dans un corpus de discours politiques d'origines différentes, dans un corpus de lois européennes ainsi que dans un ouvrage de référence en langue française sur l'apprentissage dont le nombre de pages était important. On parvient aujourd'hui à trouver automatiquement le passage à un nouvel article, ce qui est

plutôt difficile à obtenir.

L'indexation automatique dans un grand corpus de données présente un grand intérêt pour les documentalistes. Pourrons-nous disposer prochainement de tels outils ?

Pour l'instant, mes travaux appartiennent au domaine de la recherche : la collaboration avec le monde industriel reste difficile.

Mais on trouve de nombreux exemples de recherche appliquée aux États-Unis, où la recherche répond depuis les années 80 à divers impératifs, tels que sécurité, questions-réponses, bioinformatique, recherche d'informations dans des teraoctets de données. Dans le domaine de la sécurité, les campagnes MUC (Message Understanding Conferences) sont une référence. Elles ont été lancées par le département américain de la Défense, qui voulait analyser de manière automatique les dépêches parues dans les journaux sur les attentats. Ce type d'analyse permet aussi de concevoir les outils de détection automatique de spam, soit une application grand public essentiellement fondée sur la fouille de texte.

Rédigé par ADBS.
Publication le 22 juin 2006 - Mise à jour le 14 octobre 2008

URL : https://www.adbs.fr/groupes/adbs-site-internet/portrait-de-pro-jerome-aze-287276