Bien cube, les données textuelles peuvent s'agréger !

Sandra Bringay*,**, Anne Laurent*,
Pascal Poncelet*, Mathieu Roche*, Maguelonne Teisseire*,***

*LIRMM – CNRS, 161 rue Ada, Montpellier, France
{bringay,laurent,poncelet,mroche,teisseire}@lirmm.fr
**Univ. Montpellier 3
***CEMAGREF – UMR TETIS, maguelonne.teisseire@cemagref.fr

Résumé. La masse des données aujourd'hui disponibles engendre des besoins croissants de méthodes décisionnelles adaptées aux données traitées. Ainsi, récemment de nouvelles approches fondées sur des cubes de textes sont apparues pour pouvoir analyser et extraire de la connaissance à partir de documents. L'originalité de ces cubes est d'étendre les approches traditionnelles des entrepôts et des technologies OLAP à des contenus textuels. Dans cet article, nous nous intéressons à deux nouvelles fonctions d'agrégation. La première propose une nouvelle mesure de TF-IDF adaptative permettant de tenir compte des hiérarchies associées aux dimensions. La seconde est une agrégation dynamique permettant de faire émerger des groupements correspondant à une situation réelle. Les expériences menées sur des données issues du serveur HAL d'une université confirment l'intérêt de nos propositions.

1 Introduction

Avec le développement de l'Internet, de plus en plus de documents textuels sont disponibles. Extraire de la connaissance ou analyser et interroger de tels volumes de données est un enjeu important et de nombreux travaux de recherche se sont intéressés à ces problématiques. Ainsi, par exemple, les travaux menés autour de la fouille de textes ont proposé de nouvelles approches pour classer automatiquement des documents (Sebastiani (2002)), rechercher les nouvelles tendances (Saga et al. (2009)) ou extraire de l'information dans des données textuelles (Chang et al. (2006)). Plus récemment, de nouvelles approches fondées sur des cubes de textes proposent d'utiliser les technologies OLAP pour analyser et extraire de la connaissance. L'un des avantages de ces approches est notamment de pouvoir utiliser des opérateurs comme $Roll$-Up ou $Drill$-$Down$ pour naviguer au travers des hiérarchies et ainsi agréger les données en fonction des requêtes utilisateurs.

De manière à illustrer les problématiques que nous étudions dans cet article, considérons, par exemple, les documents extraits de dépêches concernant le virus de la grippe $A(H_1N_1)$. En étudiant les différents articles, il est aisé de constater que plusieurs catégories de documents peuvent apparaître : articles sur le vaccin, articles sur de nouveaux cas déclarés, articles sur les recommandations ou même articles généraux. Dans un processus d'aide à la décision, si nous désirons retrouver les mots caractéristiques de chaque catégorie, nous pouvons utiliser