Cube de textes et opérateur d'agrégation basé sur un modèle vectoriel adapté

Lamia Oukid*, Ounas Asfari**
Fadila Bentayeb**, Nadjia Benblidia*, Omar Boussaid**

*Université Saad Dahlab Blida (Laboratoire LRDSI)
B.P. 270, Route de Soumaa; 09000 Blida, Algérie.
o.lamia@hotmail.fr, benblidia@yahoo.com
**Université de Lyon (ERIC, Lyon 2)
5, avenue Pierre Mendès France 69676 Bron Cedex, France
{ounas.asfari, fadila.bentayeb, omar.boussaid}@univ-lyon2.fr

Résumé. Les technologies d'entreposage de données et d'analyse en ligne (*On-Line Analytical Processing* OLAP) ont largement fait leurs preuves pour l'analyse de données structurées, mais elles sont inadaptées pour l'analyse des données textuelles, faute d'outils et de méthodes adaptés. Nous proposons dans cet article, un modèle de cube textuel nommé *TCube*, qui comporte plusieurs dimensions sémantiques, pour une meilleure prise en charge de la sémantique des données textuelles. Les attributs de chaque dimension sémantique sont regroupés dans une hiérarchie de concepts, extraite à partir d'une ontologie de domaine utilisée comme une ressource externe. Notre cube de textes comprend une mesure d'analyse textuelle qui s'appuie à la fois sur un modèle vectoriel adapté à l'analyse OLAP et sur une technique de propagation de pertinence. Il est également associé à un nouvel opérateur d'agrégation appelé *ORank*(*OLAP-Rank*) permettant d'agréger les données textuelles dans un environnement OLAP. Les résultats préliminaires de notre étude expérimentale montrent l'intérêt de notre approche.

1 Introduction

De nos jours, les technologies d'entrepôts de données et d'analyse en ligne (OLAP) sont efficaces pour traiter des données numériques. Néanmoins, une grande partie des données circulant dans les entreprises (texte, images, vidéo, etc.) reste hors de portée des systèmes décisionnels. Ces dernières sont appelées *données complexes*. La plupart d'entre elles représentent des données textuelles (rapports, e-mails, etc.). Dans cet article, nous nous intéressons à l'analyse OLAP de ces données textuelles.

L'intégration des informations issues de données textuelles dans un processus d'analyse en ligne représente un défi pour les systèmes décisionnels. D'autre part, l'agrégation des données numériques s'effectue à l'aide de fonctions d'agrégat (somme, moyenne, min, max, etc.). Or, ces dernières ne sont pas adaptées à l'agrégation de données textuelles. La nature non structurée de ces dernières les rend difficile à analyser. Par conséquent, les questions suivantes se posent :