TLabel: Nouvel opérateur d'agrégation par catégorisation dans les cubes de textes

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

*Université de Blida 1 (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 Mondès France 69676 Bron Cedex, France.
{omar.boussaid, fadila.bentayeb}@univ-lyon2.fr

Résumé. L'analyse en ligne (OLAP) dans les cubes de textes nécessite la définition de nouveaux types d'opérateurs d'analyse appropriés aux données textuelles. En effet, les opérateurs d'agrégation classiques ont montré leur efficacité pour l'analyse en ligne des données numériques, mais ils sont inadaptés pour l'analyse des données textuelles. Dans cet article, nous proposons un nouvel opérateur d'agrégation par catégorisation nommé *TLabel* (*Text Label*) permettant d'agréger les données textuelles en plusieurs classes de documents. A chaque classe sera associée une étiquette (*Label*) qui représente le contenu sémantique des données textuelles de la classe grâce à une adaptation des techniques de fouille de textes à l'OLAP. Nous avons effectué une étude expérimentale sur notre opérateur *TLabel*. Les résultats préliminaires montrent l'intérêt de notre approche pour l'analyse en ligne des données textuelles.

1 Introduction

Les technologies d'entreposage de données et d'analyse en ligne (OLAP) ont largement fait leurs preuves pour l'analyse en ligne des données numériques. Néanmoins, une grande partie des données circulant dans les entreprises sont présentées sous forme de données textuelles (rapports, e-mails, etc.). Ces dernières restent peu exploitées par les systèmes décisionnels actuels. Permettre la prise en compte de ce type de données par les systèmes OLAP revient à définir de nouvelles techniques permettant d'intégrer la sémantique des données textuelles dans le processus d'analyse en ligne. Un des principaux challenges dans ce contexte est l'agrégation de données textuelles. En effet, l'agrégation des données numériques s'effectue à l'aide de fonctions d'agrégation classiques (somme, moyenne, min, max, etc.) qui ne sont pas adaptées aux données textuelles. La nature peu ou pas structurée de ces dernières les rend difficiles à analyser. Pour analyser en ligne des données textuelles, il est donc nécessaire de faire évoluer les cubes de données classiques vers des cubes de textes tout en proposant de nouveaux opérateurs permettant l'agrégation de ces données.