Top_Keyword : agrégation de mots-clefs dans un environnement d'analyse en ligne (OLAP)

Franck Ravat, Olivier Teste
Ronan Tournier, Gilles Zurfluh

IRIT SIG/ED, UMR5505, 118 rte. de Narbonne,
F31062 Toulouse CEDEX 9, France
{ravat, teste, tournier, zurfluh}@irit.fr
http://www.irit.fr

Résumé. Depuis plus d'une décennie, les travaux de recherche sur OLAP et les bases de données multidimensionnelles ont produit des méthodes, des outils et des moyens d'analyse de données numériques. L'accroissement de la disponibilité des documents numériques entraîne un besoin pour l'ajout de documents XML principalement constitués de données textuelles au sein de bases de données multidimensionnelles et d'un environnement adapté à leur analyse. En réponse à ce besoin, cet article présente une nouvelle fonction d'agrégation permettant l'agrégation de données textuelles au sein d'un environnement OLAP, au même titre que les fonctions d'agrégation arithmétique traditionnelles le permettent pour des données numériques. La fonction TOP_KEYWORD (ou TOP_KW) résume un ensemble de documents par leurs termes les plus significatifs, en employant une fonction de pondération issue de la recherche d'information : *tf.idf*.

1 Introduction

Les systèmes d'analyse en ligne OLAP (On-Line Analytical Processing) permettent aux analystes d'améliorer le processus de prise de décision. Ces systèmes facilitent la consultation et l'analyse de données économiques, statistiques ou scientifiques agrégées et historisées via une structuration adaptée au sein de bases de données multidimensionnelles (Colliat, 1996). Les systèmes d'aide à la décision, emploient des bases de données multidimensionnelles (BDM), qui permettent aux décideurs d'avoir une vision des performances d'une entreprise. Pour modéliser les BDM, des structures multidimensionnelles ont été définies permettant la représentation de sujets d'analyse, appelés *faits* et d'axes d'analyse, appelés *dimensions* (Kimball, 1996). Les faits sont des regroupements d'indicateurs d'analyse appelés *mesures*. Les dimensions sont composées d'attributs, agencés de manière hiérarchique, qui modélisent les différents niveaux de détails (granularité) des axes d'analyse.

Lors d'une analyse OLAP multidimensionnelle, les données représentant un sujet sont analysées en fonction de différents niveaux de détails ou niveaux de granularité. Le processus d'analyse agrège les données en fonction des niveaux de granularité sélectionnés via une