

Modèle conceptuel pour l'analyse multidimensionnelle de documents

Franck Ravat**, Olivier Teste*
Ronan Tournier*, Gilles Zurfluh**

*IRIT SIG/ED, UMR5505 – Université Toulouse 3, 118 rte. de Narbonne,
F31062 Toulouse CEDEX 9, France

** IRIT SIG/ED, UMR5505 – Université Toulouse 1, 2 rue du doyen G. Marty,
F31042 Toulouse CEDEX 9, France
{ravat, teste, tournier, zurfluh}@irit.fr

Résumé. OLAP et les entrepôts de données sont utilisés pour l'analyse de données transactionnelles. De nos jours, avec l'évolution d'Internet et le développement de formats d'échange de données semi-structurées comme par exemple XML, il est possible de considérer les documents comme source d'analyse. En conséquence, un environnement d'analyse multidimensionnel adapté à ce type de données est nécessaire. Dans cet article, nous introduisons un modèle conceptuel multidimensionnel adapté à l'analyse de données documentaires, reposant sur un unique concept : une dimension. Nous définissons aussi un ensemble d'opérateurs d'analyse multidimensionnelle adaptés.

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 et reposent sur un outil de centralisation des données: un entrepôt de données (Kimball, 1996).

1.1 Contexte et problématique

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.

L'analyse multidimensionnelle basée sur des BDM factuelles numériques est une tâche bien maîtrisée de nos jours (Sullivan, 2001). Ces BDM sont souvent construites sur des