Analyse des données textuelles : Une approche d'extraction de contenu sémantique et un opérateur d'agrégation Top_KRankedTopics

Sarah Attaf*, Nadjia Benblidia*
Omar Boussaid **

*Laboratoire LRDSI Département d'informatique
Faculté des sciences université Saad Dahlab Blida,
route de Soumaa BP 270 Blida(09000)
sarah.attaf@gmail.com
benblidia@yahoo.com
** Laboratoire ERIC University of Lyon 2,
5 AV. P. Mends-France 69676 Bron Cedex Lyon, France
omar.boussaid@univ-lyon2.fr

Résumé. La prise en compte de la sémantique des données textuelles lors d'une analyse OLAP est une tâche complexe, qui n'est pas prise en charge par les systèmes décisionnels classiques. Pour répondre à cette problématique, nous proposons dans cet article une nouvelle approche pour l'extraction des descripteurs sémantique des données textuelles afin de les utiliser dans l'analyse. L'approche proposée est basée sur l'utilisation de la méthode Latent Dirichelet allocation (LDA) et la taxonomie Open Directory Project (ODP) comme une source de connaissance externe pour identifier les sujets pertinents dans un document textuel. Notre approche vise à construire pour chaque document textuel une hiérarchie sémantique à base des concepts du ODP. Pour prendre en compte cette hiérarchie sémantique lors d'une analyse OLAP, nous proposons une fonction de pondération ainsi qu'un opérateur d'agrégation qui sélectionne les k premiers sujets et retourne pour chaque sujet une liste de documents pondérés.

1 Introduction

Le document électronique représente aujourd'hui un vecteur et un support d'information que les organisations ne doivent pas négliger. En effet, il est entendu que plus de 80 % des données nécessaires au fonctionnement d'une organisation sont encapsulées dans des documents, et non uniquement dans les bases de données opérationnelles. Ces données textuelles restent hors de portée des systèmes décisionnels, ce qui induit qu'une grande partie de l'information demeure inaccessible. Pour répondre à cette problématique et afin de pouvoir prendre profit des informations contenues dans ces documents, il est devenu plus que nécessaire d'intégrer ces données textuelles dans des systèmes d'information décisionnels permettant leur analyse.