UNIVERSITÉ LUMIÈRE LYON 2

ÉCOLE DOCTORALE INFORMATIQUE ET MATHÉMATIQUES

T H È S E

pour obtenir le titre de

Docteur en Informatique

de l'Université de Lyon - Lumière Lyon2

Présentée et soutenue par

Rachid AKNOUCHE

Entrepôt de textes : de l'intégration à la modélisation multidimensionnelle de données textuelles

---

Sous la direction de

M. Omar BOUSSAID et Mme Fadila BENTAYEB

préparée au sein du Laboratoire ERIC-Université Lumière Lyon2

soutenue publiquement le 26 avril 2014

Jury :

Rapporteurs :	Esteban ZIMANYI	-	Professeur, Université Libre Bruxelles
	Mohand BOUGHANEM	-	Professeur, Université Toulouse 3
Examinateurs :	Bernard ESPINASSE	-	Professeur, Aix-Marseille Université
	Jérôme GENSEL	-	Professeur, Université de Grenoble
Directeurs :	Omar BOUSSAID	-	Professeur, Université Lyon 2
	Fadila BENTAYEB	-	MCF-HDR, Université Lyon 2

Résumé : Le travail présenté dans ce mémoire vise à proposer des solutions aux problèmes d'entreposage des données textuelles. L'intérêt porté à ce type de données est motivé par le fait qu'elles ne peuvent être intégrées et entreposées par l'application de simples techniques employées dans les systèmes décisionnels actuels. Pour aborder cette problématique, nous avons proposé une démarche pour la construction d'entrepôts de textes. Elle couvre les principales phases d'un processus classique d'entreposage des données et utilise de nouvelles méthodes adaptées aux données textuelles. Dans ces travaux de thèse, nous nous sommes focalisés sur les deux premières phases qui sont l'intégration des données textuelles et leur modélisation multidimensionnelle.

Pour mettre en place une solution d'intégration de ce type de données, nous avons eu recours aux techniques de recherche d'information (RI) et du traitement automatique du langage naturel (TALN). Pour cela, nous avons conçu un processus d'ETL (*Extract-Transform-Load*) adapté aux données textuelles. Il s'agit d'un framework d'intégration, nommé ETL-Text, qui permet de déployer différentes tâches d'extraction, de filtrage et de transformation des données textuelles originelles sous une forme leur permettant d'être entreposées. Certaines de ces tâches sont réalisées dans une approche, baptisée RICSH (Recherche d'information contextuelle par segmentation thématique de documents), de prétraitement et de recherche de données textuelles.

D'autre part, l'organisation des données textuelles à des fins d'analyse est effectuée selon TWM (*Text Warehouse Modelling*), un nouveau modèle multidimensionnel adapté à ce type de données. Celui-ci étend le modèle en constellation classique pour prendre en charge la représentation des textes dans un environnement multidimensionnel. Dans TWM, il est défini une dimension sémantique conçue pour structurer les thèmes des documents et pour hiérarchiser les concepts sémantiques. Pour cela, TWM est adossé à une source sémantique externe, Wikipédia, en l'occurrence, pour traiter la partie sémantique du modèle. De plus, nous avons développé WikiCat, un outil pour alimenter la dimension sémantique de TWM avec des descripteurs sémantiques issus de Wikipédia. Ces deux dernières contributions complètent le framework ETL-Text pour constituer le dispositif d'entreposage des données textuelles.

Pour valider nos différentes contributions, nous avons réalisé, en plus des travaux d'implémentation, une étude expérimentale pour chacune de nos propositions. Face au phénomène des données massives, nous avons développé dans le cadre d'une étude de cas des algorithmes de parallélisation des traitements en utilisant le paradigme MapReduce que nous avons testés dans l'environnement Hadoop.

Mots clés : Intégration des données textuelles - Entrepôts de textes - ETL-Text- Modélisation multidimensionnelle des données textuelle- RICSH- TWM- Recherche d'information- MapReduce - Enrichissement de documents- Wikipédia

Abstract : The work, presented in this thesis, aims to propose solutions to the problems of textual data warehousing. The interest in the textual data is motivated by the fact that they cannot be integrated and warehoused by using the traditional applications and the current techniques of decision-making systems. In order to overcome this problem, we proposed a text warehouses approach which covers the main phases of a data warehousing process adapted to textual data. We focused specifically on the integration of textual data and their multidimensional modeling.

For the textual data integration, we used information retrieval (IR) techniques and automatic natural language processing (NLP). Thus, we proposed an integration framework, called ETL-Text which is an ETL (Extract- Transform- Load) process suitable for textual data. The ETL-Text performs the extracting, filtering and transforming tasks of the original textual data in a form allowing them to be warehoused. Some of these tasks are performed in our RICSH approach (Contextual information retrieval by topics segmentation of documents) for pretreatment and textual data search.

On the other hand, the organization of textual data for the analysis is carried out by our proposed TWM (Text Warehouse Modelling). It is a new multidimensional model suitable for textual data. It extends the classical constellation model to support the representation of textual data in a multidimensional environment. TWM includes a semantic dimension defined for structuring documents and topics by organizing the semantic concepts into a hierarchy. Also, we depend on a Wikipedia, as an external semantic source, to achieve the semantic part of the model. Furthermore, we developed WikiCat, which is a tool permit to feed the TWM semantic dimension with semantics descriptors from Wikipedia. These last two contributions complement the ETL-Text framework to establish the text warehouse device.

To validate the different contributions, we performed, besides the implementation works, an experimental study for each model. For the emergence of large data, we developed, as part of a case study, a parallel processing algorithms using the MapReduce paradigm tested in the Apache Hadoop environment.

Keywords : Integration of textual data- Text Warehouses - ETL-Text - RICSH - Text Warehouse Model - TWM - Information Retrieval - MapReduce - Enrichment of textual documents - Wikipedia

Table des matières