

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 :

<i>Rapporteurs :</i>	Esteban ZIMANYI	-	Professeur, Université Libre Bruxelles
	Mohand BOUGHANEM	-	Professeur, Université Toulouse 3
<i>Examineurs :</i>	Bernard ESPINASSE	-	Professeur, Aix-Marseille Université
	Jérôme GENSEL	-	Professeur, Université de Grenoble
<i>Directeurs :</i>	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

1	Introduction	1
1.1	Contexte général	1
1.2	Problématique	3
1.3	Contributions de nos travaux de recherche	4
1.4	Organisation du mémoire	8
Partie I	État de l'art	11
2	Concepts de base	13
2.1	Entrepôt de données	13
2.1.1	Définition	13
2.1.2	Architecture d'entreposage des données	14
2.2	ETL : Extraction, Transformation et Chargement	17
2.2.1	Définition	17
2.2.2	ETL et données non structurées	18
2.3	Modélisation multidimensionnelle des données	20
2.3.1	Définition	20
2.3.2	Standards de modélisation	20
3	Techniques de recherche d'information	23
3.1	Introduction	23
3.2	Analyse et indexation des documents et des requêtes	24
3.2.1	Indexation des documents	24
3.2.2	Pondération des termes	24
3.2.3	Filtrage des mots fonctionnels	28
3.2.4	Racinisation (Lemmatisation Stemmatisation)	28
3.3	Modèles de RI	29
3.3.1	Modèles ensemblistes	29
3.3.2	Modèles algébriques	31
3.3.3	Modèles probabilistes	33
3.4	Contexte dans la recherche d'information	34
3.4.1	Contexte utilisateur	34
3.4.2	Contexte requête	35
3.5	Modèle de recherche d'information contextuelle basé sur la modélisation de langue	36
3.5.1	Modèle de langue en recherche d'information	36

4	Modèles d'entrepôts de documents	39
4.1	Modélisation multidimensionnelle des données complexes	39
4.1.1	Modèles multidimensionnels pour la construction des entrepôts de documents	40
4.2	Synthèse des travaux	50
4.2.1	Choix des paramètres de comparaison	50
4.2.2	Bilan et discussion	53
 Partie II Intégration des données textuelles dans un entrepôt de textes		 55
5	RICSH : Recherche d'information contextuelle par segmentation thématique de documents	57
5.1	Introduction	57
5.2	Exemple d'une démarche de RI classique	59
5.3	Approche RICSH : Recherche d'information contextuelle par segmentation thématique de documents	60
5.3.1	Phase de prétraitement	61
5.3.2	Phase de représentation et de classification des documents	65
5.4	Prise en compte du contexte pour améliorer la recherche d'information	68
5.4.1	Modèle de langue statistique	70
5.4.2	Discussion	70
5.4.3	Modèle général de recherche d'information	71
5.4.4	Construction du modèle de contexte requête	72
5.4.5	Construction du modèle de contexte utilisateur	73
5.5	Implémentation de l'approche RICSH	75
5.6	Conclusion	76
6	ETL-Text : Processus d'intégration des données textuelles dans un entrepôt de textes	77
6.1	Introduction	77
6.2	Modélisation d'un processus ETL	78
6.3	Un processus ETL-Text adapté aux entrepôts de textes	80
6.3.1	Phase d'extraction et d'épuration des données	82
6.3.2	Phase de transformation et de représentation	84
6.4	Implémentation et expérimentation	88
6.5	Conclusion	88
 Partie III Modélisation multidimensionnelle appropriée aux données textuelles		 89
7	TWM : Modèle Multidimensionnel de l'entrepôt de textes	91
7.1	Introduction	91

7.2	Modèles multidimensionnels pour la construction des entrepôts de documents	94
7.3	TWM : Modèle d'entrepôt de textes	95
7.3.1	Dimensions dans TWM	96
7.3.2	Faits dans TWM	99
7.3.3	Mesures de faits	99
7.3.4	Requête de RI dans TWM	101
7.3.5	Exemple d'illustration du modèle TWM	102
7.4	Conclusion	104
8	WikiCat : Un outil d'enrichissement de documents par des descripteurs sémantiques issus de Wikipédia	107
8.1	Introduction	107
8.2	Wikipédia	109
8.3	Les domaines d'application de Wikipédia	111
8.3.1	Wikipédia dans le domaine du Traitement Automatique du Langage Naturel	111
8.3.2	Wikipédia dans le domaine de l'extraction d'information	111
8.3.3	Wikipédia dans le domaine de la recherche d'information	112
8.4	Enrichissement de documents par des descripteurs sémantiques issus de Wikipédia	112
8.4.1	Mise en correspondance entre les documents et les articles Wikipédia	113
8.4.2	Mise en correspondance entre les articles Wikipédia et les catégories Wikipédia	113
8.4.3	Mise en correspondance entre les documents et les catégories Wikipédia	114
8.4.4	Chargement et fractionnement du fichier dump Wikipédia	114
8.4.5	Indexation des articles Wikipédia	115
8.4.6	Identification des catégories Wikipédia	116
8.5	Implémentation et évaluation de l'approche	117
8.6	Conclusion	117
Partie IV	Évaluation et validation	119
9	Implémentation	121
9.1	Introduction	121
9.2	Architecture générale de la plate-forme <i>Text Warehousing</i>	121
9.2.1	Module d'extraction (ME)	123
9.2.2	Module de transformation (MT)	123
9.2.3	Module de chargement (MC)	124
9.3	Implémentation de l'approche RICSH	124
9.3.1	Modules de prétraitement du corpus	125

9.3.2	Génération des termes-candidats - Module MTC	126
9.3.3	Segmentation thématiques	127
9.4	Implémentation de WikiCat	128
9.4.1	Chargement et fractionnement du fichier dump Wikipédia	128
9.4.2	Indexation des articles Wikipédia	128
9.4.3	Interrogation de l'index Wikipédia	129
9.4.4	Identification des catégories Wikipédia	130
9.4.5	L'accès aux résultats avancés de l'analyse	131
9.4.6	Interface de paramétrage	132
9.5	Traitement parallèle d'une requête dans RICSH	133
9.5.1	Le paradigme MapReduce	133
9.5.2	Hadoop	134
9.5.3	RICSH avec MapReduce	134
9.5.4	Conclusion	138
10	Expérimentations	141
10.1	Introduction	141
10.2	Évaluation des performances des systèmes de recherche d'information	142
10.3	Évaluation de l'approche RICSH	143
10.3.1	Corpus 20 Newsgroups	143
10.3.2	Description de la base de référence pour l'évaluation	143
10.3.3	Expérimentation de 20 Newsgroups avec Lemur	144
10.3.4	Résultats	145
10.4	ETL-Text : Intégration des données textuelles dans l'entrepôt	146
10.5	Évaluation de WikiCat	148
10.6	Évaluation de l'approche de parallélisation de RICSH	150
10.6.1	Temps d'exécution avec variation du nombre de nœuds	150
10.6.2	Temps d'exécution d'une requête de 2 mots sur un seul nœud	152
10.7	Conclusion	154
Partie V	Conclusion et perspectives	155
11	Conclusion et perspectives	157
11.1	Bilan et contributions	157
11.2	Perspectives	160
11.2.1	Analyse en ligne des données structurées et non structurées Text-Olap	160
11.2.2	Passage à l'échelle	160
A	Liste des publications	161
	Bibliographie	163