

République algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique



Université M'hamed BOUGARA de BOUMERDES
Faculté des sciences
Département Informatique

MEMOIRE

Présenté pour l'obtention du diplôme de Magister

Spécialité:

INFORMATIQUE

Option:

Spécification de logiciels et traitement de l'information

Par : BOUCHAM Souhila

Thème

**Une approche basée Ontologies pour l'indexation automatique
et la Recherche d'Information Multilingue (RIM)**

Soutenu devant le jury:

Pr. M. Mezghiche
Dr. O. Nouali
Pr. Alimzighi Zaia
Mme Aliane Hassina

Professeur, Université de Boumerdes
Maître de recherche, CERIST
Professeur, USTHB
Chargée de recherche / CERIST

Président
Examineur
Rapporteur
Invitée

Année universitaire 2008-2009

Résumé

Notre travail se situe dans le contexte de la *recherche d'information (RI)*, plus particulièrement la recherche d'information multilingue (RIM).

L'objectif de ce projet est de proposer une solution pour la recherche d'information multilingue afin d'explorer l'apport des approches Web Sémantique en particulier l'utilisation des ontologies pour améliorer la description sémantique des documents et des requêtes.

Nous proposons dans ce travail une approche pour l'indexation et la recherche d'information pour un corpus trilingue : arabe, français et anglais. Le système proposé est fondé sur un formalisme de représentation de connaissances, plus précisément les graphes sémantiques qui supportent une ontologie de domaine. Les documents et les requêtes sont aussi représentés dans ce formalisme.

L'ontologie du domaine constitue le noyau du système et est utilisée aussi bien pour l'indexation que pour la recherche. Le système d'indexation utilise une méthode d'extraction qui est basée sur le calcul de segments répétés en utilisant des filtres linguistiques. Le système de recherche consiste en une comparaison de graphes pour trouver les documents qui répondent à la requête étendue de l'utilisateur.

Mots-clés : *recherche d'information multilingue, Indexation automatique, Ontologie, Extraction de connaissances, expansion de la requête, Graphes sémantiques.*

Abstract

This work deals with information retrieval (IR), more particularly multilingual information retrieval (MIR).

The aim of our project is to propose a solution to MIR and explore the contribution of semantic web approaches, particularly the use of ontologies to improve the semantic description of documents and queries.

We propose an approach to indexing and retrieval in a trilingual corpus: arabic, french and english. The proposed system is founded on a knowledge representation formalism, namely semantic graphs which supports a domain ontology. Documents and queries are also represented in this formalism. The domain ontology constitutes the kernel of the system and is used both for indexing and retrieval.

The indexing system uses an extraction method based on repeated segments calculation. When identified, repeated segments are submitted to a filtering procedure using linguistic filters.

The retrieval system consists of a graph comparison to find relevant documents for the extended user query.

Keywords: multilingual information retrieval, automatic indexing, ontology, Extracting information, query expansion, semantic graphs.

Table des matières

Introduction générale

Partie I

Chapitre 1 : De la Recherche d'Information (RI) à la Recherche d'Information

Multilingue (RIM)

1. La recherche d'information

- 1.1. Introduction
- 1.2. définition d'un SRI
- 1.3. Architecture générale des SRI
- 1.4. Modèle de représentation
- 1.4.1. Les entités d'indexation
- 1.4.2. Les langages d'indexation
- 1.4.3. Les approches d'indexation
- 1.4.3.1. Indexation manuelle
- 1.4.3.2. Indexation automatique
- 1.5. Type d'indexation ou de représentation : Modèles de RI
- 1.5.1. Indexation à plat du modèle booléen
- 1.5.2. Indexation pondérée des modèles vectoriel et probabiliste
- 1.5.3. Indexation structurée et le modèle logique
- 1.5.4. Indexation sémantique : indexation basée sur les connaissances
- 1.6. Méthodes d'évaluation
- 1.7. Critères d'une bonne indexation
- 1.7.1. La cohérence
- 1.7.2. L'adéquation entre les représentations
- 1.8. Conclusion : Vers l'utilisation des techniques de TALN

2. Traitement Automatique des Langues et recherche d'information

- 2.1. Introduction
- 2.2. Définition
- 2.3. Grands domaines du Traitement Automatique des Langues
- 2.3.1. La morphologie
- 2.3.2. La syntaxe
- 2.3.3. La sémantique
- 2.3.4. La pragmatique
- 2.4. Quelques pièges du langage naturel
- 2.5. Techniques de TAL pour la recherche d'information
- 2.5.1. Palier morphologique
- 2.5.1.1. Segmentation en unités linguistiques
- 2.5.1.2. Racinisation
- 2.5.2. Palier syntaxique
- 2.5.2.1. Etiquetage ou désambiguïsation syntaxique
- 2.5.2.2. Analyse « peu profonde » ou « surfacique »
- 2.5.2.3. Indexation sur les syntagmes et variation
- 2.5.2.4. Reconnaissance des entités nommées
- 2.5.3. Paliers sémantique et pragmatique

2.5.3.1.	Etiquetage sémantique	
2.5.3.2.	Résolution d'anaphores	
2.5.4.	Techniques transversales	
2.5.4.1.	Statistiques textuelles	
2.5.4.2.	Traduction automatique et RI interlangue	
2.6.	Conclusion	
3.	Extraction des connaissances à partir des textes	
3.1.	Introduction	
3.2.	Unités lexicales et conceptuelles	
3.2.1.	Mots clés	
3.2.2.	Termes	
3.2.3.	Unités de sens : Concepts ou catégories conceptuelles	
3.3.	Relations sémantiques	
3.3.1.	Relations d'inclusion et d'identité	
3.3.1.1.	Synonymie	
3.3.1.2.	Hyponymie	
3.3.1.3.	Méronymie	
3.3.2.	Relations d'exclusion et d'opposition	
3.3.2.1.	Co-hyponyme	
3.3.2.2.	Complémentation	
3.3.2.3.	Antonyme	
3.4.	Les approches d'extraction de termes	
3.4.1.	Les Méthodes à base de patrons	
3.4.1.1.	Patrons morpho-syntaxiques	
3.4.1.2.	La méthode de Jacques Vergne	
3.4.1.3.	Système ANA	
3.4.1.4.	Patrons morphologiques	
3.4.2.	Mesures d'association (mesures statistiques).....	
3.4.2.1.	Fréquence de co-occurrence	
3.4.2.2.	Le test du χ^2	
3.4.2.3.	Le coefficient de Jaccard	
3.4.2.4.	L'information mutuelle	
3.4.2.5.	Coefficient de Dice	
3.4.2.6.	Limites des mesures d'association	
3.4.2.	Évaluation des résultats de l'extraction terminologique	
3.5.	Les approches d'extraction de relations sémantiques	
3.5.1.	Vecteurs et graphes de co-occurrences	
3.5.2.	Classification	
3.5.3.	Patrons lexico-syntaxiques	
3.5.4.	Utilisation de la structure interne des termes	
3.5.4.1.	Utilisation de la structure lexicale des termes polylexicaux	
3.5.4.2.	Utilisation de la structure morphologique des termes simples	
3.5.5.	Evaluation des résultats d'acquisition de relations sémantiques	
3.6.	Syntagmes et la recherche d'information	
3.6.1.	Notion de syntagme	
3.6.2.	Utilisation des syntagmes en recherche d'information	
3.7.	Conclusion	

4. Recherche d'Information Multilingue (RIM)
4.1. Introduction
4.2. Contexte de la recherche d'information multilingue
4.2.1. Requête multilingue
4.2.2. Base multilingue de documents
4.2.3. Document multilingue
4.3. Problèmes de la recherche d'information multilingue
4.4. Indexation multilingue: les différentes approches de la traduction automatique
4.4.1. Approche basée sur la traduction de la requête
4.4.2. Approche basée sur la traduction des documents
4.4.3. Approche basée sur le langage pivot
4.5. Ressources linguistiques pour le traitement d'information multilingue
4.5.1. système de traduction automatique
4.5.2. Les bases lexicales
4.5.2.1. les dictionnaires de transfert
4.5.2.2. utilisation des bases de connaissances: ontologies et thésaurus
4.5.3. Les corpus
4.5.3.1. les corpus parallèles
4.5.3.2. les corpus comparables
4.6. Exemple d'un SRIM : SyDoM : Système Documentaire Multilingue
4.6.1. Le module de gestion du thésaurus sémantique
4.6.1.1. Le niveau conceptuel (support)
4.6.1.2. Le niveau terminologique
4.6.2. Le module d'indexation
4.6.2.1. Les annotations
4.6.2.2. L'index du document
4.6.3. Le module de recherche
4.7. Conclusion
Chapitre 2. Les Ontologies
2.1. Introduction
2.2. Bases théoriques
2.2.1. Qu'est ce qu'une ontologie ?
2.2.2. Au-delà des définitions
2.2.3. Les objectifs de l'ontologie
2.2.4. Composants des ontologies
2.2.5. Types d'ontologies
2.2.6. Les différents modes de représentation des ontologies
2.2.6.1. Réseaux sémantiques
2.2.6.2. Les graphes conceptuels
2.2.6.3. Les frames
2.2.6.4. Les logiques de description
2.3. Langages de spécification d'ontologie pour le Web sémantique
2.3.1. SHOE : (Simple HTML Ontology Extension)
2.3.2. Ontobroker
2.3.3. Ontoseek
2.3.4. Webkb
2.3.5. CONCERTO
2.3.6. RDF (Resource Description Framework)
2.3.7. RDFSchéma

2.3.8. OWL (Ontology Web Language)	
2.4. Conclusion	

Chapitre 3. Utilisation des ontologies pour la recherche d'information et l'extraction de connaissances

3.1. Introduction	
3.2. principe d'utilisation des ontologies par un SRI	
3.3. Indexation sémantique: Indexation à partir d'ontologies	
3.1. Identification des concepts et des instances existant dans l'ontologie	
3.1.1. Extraction des termes du document	
3.1.2. Recherche des labels correspondant à des concepts de l'ontologie	
3.1.3. Désambiguïsation des labels	
3.1.4. Extraction de nouvelles instances	
3.2. Pondération des concepts et instances	
3.2.1. Pondération statistique	
3.2.2. Pondération à partir de similarité conceptuelle	
3.3. Appariement à partir d'ontologies	
3.4. Reformulation de requête à partir des termes de l'ontologie	
3.4. Apports de l'ontologie dans le domaine de la RI	
3.5. Les ontologies les plus connues	
5.1. Ontologies de représentation des connaissances	
5.2. Ontologies de haut niveau	
5.3. Ontologies linguistiques	
5.4. Ontologies d'ingénierie	
3.6. Conclusion	

Partie II

Chapitre 4: Vers une approche basée Ontologie pour l'indexation automatique et La recherche d'information multilingue

4.1. Introduction	
4.2. L'extension du modèle des GC pour la RI	
4.3. Formalisme des graphes sémantiques	
4.4. Observations sur les langages documentaires	
4.5. Un modèle de SRIM basé sur l'ontologie de domaine	
4.5.1. Vue globale de l'approche	
4.5.2. L'ontologie de domaine : Thésaurus sémantique	
4.5.2.1. La conceptualisation du domaine ou Support	
4.5.2.2. hiérarchie des types de concepts	
4.5.2.3. hiérarchie des types de relations	
4.5.2.4. les relations entre types	
4.5.2.5. Définition formelle de l'ontologie (thésaurus sémantique).....	
4.6. Construction manuelle du Thésaurus sémantique	
4.7. Indexation, extraction et génération des graphes sémantiques	
4.7.1. Extraction des termes à partir des textes	
4.7.1.1. Contexte	
4.7.1.2. Analyse de surface pour l'extraction des syntagmes nominaux ...	
4.7.1.3. Fonctionnement de notre extracteur de termes	

4.7.1.4. Expansion de la liste des candidats-termes	
4.7.2. Démarche de l'extraction des relations sémantiques.....	
4.7.2.1. les relation syntagmatique	
4.7.2.2. relations paradigmaticues	
4.7.3. Génération de graphe sémantique	
4.8. La recherche, étendre la requête via une ontologie	
4.9. Architecture du SRIM basée sur un thésaurus sémantique	
4.10. Conclusion	
Conclusion générale	
Annexe	

Bibliographie