

# THÈSE DE DOCTORAT

Présentée par

**Fadoua Ataa Allah**

**Titre :** Information Retrieval: Applications to  
English and Arabic Documents

**Discipline :** Sciences de l'ingénieur

**Spécialité :** Informatique et Télécommunications

**U.F.R :** Informatique et Télécommunications

**Période d'accréditation :** 2005-2008

**Directeur de l'UFR :** Prof. Driss ABOUTAJDINE

**Directeur de thèse :** Driss ABOUTAJDINE

**Soutenance :**

**Date :** 23 Mai 2008

**Heure :** 15h00

**Lieu :** Amphi Bel Mahi

**Devant le jury**

**Président :**

D. Aboutajdine                      PES à la Faculté des Sciences de Rabat.

**Examineurs :**

M. ABBAD                              PES à la Faculté des Sciences de Rabat.

W. I. Grosky                            PES à l'université de Michigan-Dearborn, USA.

N. Mouaddib                            PES à l'école Polytechnique de l'université de  
Nantes, France.

A. Soudi                                  PES à l'école Nationale de l'Industrie Minérale,  
Rabat.

M. Diab                                  PA à l'université de Columbia, USA.

A. El Qadi                                PA à l'école Supérieure de Technologie de  
Meknès.

---

## **Abstract:**

Arabic information retrieval has become a focus of research and commercial development due to the vital necessity of such tools for people in the electronic age. The number of Arabic-speaking Internet users is assumed to achieve 43 millions during this year; however, on the other side, few full search engines are available to Arabic-speaking users. This dissertation focuses on three naturally related areas of research: information retrieval, document clustering, and dimensionality reduction.

In information retrieval, we propose an Arabic information retrieval system, based on light stemming in the pre-processing phase, and on the Okapi BM-25 weighting scheme and the latent semantic analysis model in the processing phase. This system has been suggested after performing and analyzing many experiments dealing with Arabic natural language processing and different weighting schemes found in literature. Moreover, it has been compared with another proposed system based on noun phrase indexation.

In clustering, we propose to use the diffusion map space based on the cosine kernel and the singular value decomposition (that we denote by the cosine diffusion map space) for clustering documents. We illustrate experimentally, using the k-means clustering algorithm, the robustness of document indexation in this space compared to the Salton's space. We discuss the problems of the reduced dimension determination and the choice of clusters' number, and we provide some solutions for these issues. We provide some statistical results and discuss how the k-means algorithm performs better in the latent semantic analysis model space than in the cosine diffusion map space in the case of two clusters, but not in the case of multi-clusters. We propose a new approach for online clustering, based on the cosine diffusion map and the updating singular value decomposition method.

Concerning dimensionality reduction, we use singular value decomposition technique for feature transformation, while we propose to supplement this reduction by a generic term extracting algorithm for features selection in the context of information retrieval.

---

**Keywords (5):** Information Retrieval, Natural Language Processing, Document Clustering, Indexation, Diffusion Map.

---

---

## **Résumé:**

La recherche d'information en langue Arabe est devenue de plus en plus importante. Néanmoins, peu de moteurs de recherche spécialisés en cette langue existent, d'où la nécessité de mener des recherches dans ce contexte. A cette fin, nous nous sommes intéressés à plusieurs disciplines : le traitement automatique des langues naturelles, la classification et l'indexation.

Nous avons réalisé un système de recherche d'informations dédié à la langue Arabe, fondé sur la pseudo-racinisation, la pondération Okapi BM-25 et le modèle d'Analyse de la Sémantique Latente. Ce système a été évalué sur un corpus, que nous avons construit, traitant la thématique d'environnement.

Par ailleurs, dans le cadre de l'optimisation du temps de réponse de notre système, nous avons proposé une nouvelle technique de réduction de la taille de la base d'index, en utilisant l'extraction des termes génériques. Ainsi que nous avons conçu une approche de diffusion map basé sur le noyau du cosinus et la décomposition en valeurs singulières pour le regroupement des documents en toutes langues, en particulier l'anglais. Cette approche a résolu le problème du choix du nombre des groupes et celui de la détermination de la dimension de l'espace.

---

**Mots-clefs (5):** Recherche d'Information, Traitement Automatique des Langues Naturelles, Regroupement des Documents, Indexation, Diffusion Map.

---