



N° d'ordre :

UNIVERSITE DE M'SILA
FACULTE DES MATHÉMATIQUES ET DE L'INFORMATIQUE
Département d'Informatique

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de **MASTER**

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Technologies de l'Information et de la Communication

Par: OTMANI Kenza

SUJET

**Recherche d'information en langue arabe à base
d'ontologie**

Soutenu publiquement le : 16 / 06 / 2015 devant le jury composé de :

Mr. MOUHOUB

Mr. MAHDJOUBI Roussafi

Mr. CHALABI

Université de M'sila

Université de M'sila

Université de M'sila

Président

Rapporteur

Examineur

Promotion : 2014 /2015

Table des Matières

Introduction générale	I
Chapitre I : Recherche d'information	
1. Recherche d'information	4
1.1. Définitions	4
1.2. Les principaux acteurs de RI	5
1.3. Les approches de RI.....	5
2. Le système de recherche d'information (SRI)	6
2.1. Définitions	6
2.2. Notions de base dans un SRI.....	6
3. Architecture générale d'un SRI	7
3.1. Modèle de représentation.....	9
3.2. Modèle de recherche d'information (Appariement)	9
3.3. Reformulation	9
4. La fonctionnalité d'un SRI	10
5. L'indexation.....	11
5.1. Définition de l'indexation.....	11
5.2. Les types d'indexation	11
5.2.1. Indexation manuelle.....	12
5.2.2. Indexation automatique.....	12
5.2.3. Indexation semi-automatique.....	13
5.3. Critères d'une bonne indexation	14
5.3.1. La cohérence	14
5.3.2. L'adéquation entre les représentations	14
6. L'interrogation	15
7. Les modèles de recherche d'information.....	15
7.1. Le modèle booléen.....	16
7.2. Le modèle probabiliste.....	17
7.3. Le modèle vectoriel	18
7.3.1. Définition	18
7.3.2. Pondération des termes	19
7.3.3. L'expansion de requête	20
8. Evaluation d'un SRI.....	21
8.1. Collection de test	22

8.2. Mesures d'évaluation.....	23
Conclusion	24

Chapitre II : Les ontologies et la langue arabe

1. Les ontologies	25
1.1. Définitions	25
1.2. Rôle de l'ontologie.....	26
1.3. Composantes d'une ontologie.....	26
1.4. Types d'ontologies.....	28
1.5. Langages d'ontologies	29
1.5.1. RDF et RDF-S	29
1.5.2. OIL.....	29
1.5.3. DAML.....	29
1.5.4. OWL	30
2. Principe d'utilisation des ontologies par les SRI.....	30
2.1. Indexation à partir d'une ontologie.....	31
2.1.1. Indexation sémantique	32
2.1.2. Indexation conceptuelle	32
2.2. Appariement à partir d'une ontologie.....	33
2.3. Reformulation de requête à partir d'une ontologie.....	34
3. Langue utilisée pour la construction de l'ontologie : La langue arabe	34
3.1. Présentation de la langue arabe.....	34
3.2. La morphologie de la langue arabe.....	35
3.2.1. Les éléments essentiels de la morphologie arabe	35
3.2.2. Principe de composition des mots arabes	39
3.2.3. Les catégories du mot	40
3.3. Les méthodes d'extraction de la racine d'un mot arabe	41
3.3.1. Les méthodes basées sur la génération systématique	41
3.3.2. Les méthodes basées sur la transcription ou la traduction.....	42
3.3.3. Les méthodes basées sur l'analyse statistique	42
3.3.4. Les méthodes basées sur l'analyse morphologique	43
3.3.5. Les méthodes basées sur l'analyse statistique et morphologique	43
Conclusion	44

Chapitre III : Conception et réalisation du système

1. Le système proposé.....	45
2. L'architecture générale du système	46
2.1. Le module d'indexation	47
2.2. Le module de recherche	51
2.3. Le module d'ontologie.....	52
2.3.1. Choix d'une méthodologie de construction de l'ontologie.....	52
2.3.2. Etapes de conception de l'ontologie الجامعة.....	53
3. Les étapes de création de notre système de recherche.....	61
4. Environnement de développement.....	61
4.1. Les langages utilisés	61
4.1.1. Java	61
4.1.2. Serveur de base de données (PostgreSQL).....	61
4.2. Les outils utilisés	62
4.2.1. Eclipse.....	62
4.2.2. Protégé	62
4.2.3. Les APIs JENA.....	62
5. Les fenêtres de notre système	63
5.1. La fenêtre principale	63
5.2. La fenêtre d'ajout de documents.....	64
5.3. La fenêtre de l'affichage de la base de données	64
Conclusion	65
Conclusion générale.....	66
Bibliographie	68

INTRODUCTION GENERALE

Aujourd'hui, l'information joue un rôle primordial dans le quotidien des individus et dans l'essor des entreprises. Cependant, le développement de l'internet et la généralisation de l'informatique dans tous les domaines ont conduit à la production d'un volume d'information sans précédent. En effet, la quantité d'information disponible, se mesure en milliards de pages. Il est par conséquent, de plus en plus difficile de localiser précisément ce que l'on recherche dans cette masse d'information. La recherche d'information (RI) est le domaine par excellence qui s'intéresse à répondre à ce type d'attente. En effet, l'objectif principal de la RI est de fournir des modèles, des techniques et des outils pour stocker et organiser des masses d'information et localiser celles qui seraient pertinentes relativement à un besoin en information d'un utilisateur, souvent, exprimé à travers une requête. Ces outils sont appelés systèmes de recherche d'information (SRI). De manière générale, le fonctionnement d'un SRI consiste à construire une représentation des documents et de la requête et d'établir une comparaison entre ces deux représentations pour retourner les documents pertinents. Cette comparaison est réalisée au moyen d'un modèle de recherche. Afin d'obtenir un SRI performant, il est nécessaire de construire une bonne représentation du document et de la requête, et de développer un modèle de RI qui supporte ces représentations.

La plupart des SRI existants représentent les documents comme un ensemble de mots clés, ce que l'on appelle communément une représentation par sac de mots. Ces mots clés sont généralement pondérés en utilisant des schémas de pondération tels que $tf \times idf$ qui prend en compte les statistiques suivantes : la fréquence des termes dans le document (tf), sa fréquence dans la collection (idf) et la taille du document. Ce modèle suppose que les mots clés sont indépendants. Cette hypothèse d'indépendance entre termes facilite grandement les calculs. De ce fait, l'ordre des termes dans une phrase est donc ignoré. Ceci peut de toute évidence conduire à l'ambiguïté entre termes qui pourraient engendrer des résultats bruités.

On utilise les ontologies pour rendre les systèmes de recherche d'information plus efficaces. En effet, les ontologies offrent des ressources sous la forme de relations sémantiques, qui peuvent constituer une solution (parmi d'autres) pour résoudre le problème des variations sémantiques. Elles permettent d'étendre le champ de recherche d'une requête ; ce qui a pour conséquence, l'amélioration des résultats de la recherche.

Le fonctionnement des systèmes de recherche d'information actuels opèrent essentiellement sur les termes (en mesurant leur présence ou leur absence) sans faire appel à des relations sémantiques. Dans le cadre de la recherche d'information, pour les textes en langue arabe, la récupération de mots clés est jugée insuffisante, car les termes utilisés dans la requête peuvent présenter par rapport aux documents de la base, des différences sur plusieurs plans, par exemple :

- **Des variations morphologiques** comme « مدرسة » et « مدرستان », « خيل » et « خيول ».
- **Des variations lexicales** (on utilise pour le même sens des mots différents) comme dans le cas dans « حصان » et « خيل ».
- **Des variations sémantiques** pour un même mot comme dans le cas de « مرادف الصخر » : « الحجر » et « أنثى الخيل ».

Le problème qui se pose : Comment découvrir le document rare ou caché, celui qui contient de réelles informations répondant à la requête ? Comment sélectionner des documents, les informations pertinentes pour des objectifs précis ?

Pour localiser et délivrer des documents en langue arabe à un utilisateur, en fonction de son besoin en information, une manière pertinente est de chercher l'ensemble de contenus dans un document, et d'introduire plus de sémantique dans la recherche d'information sur les textes. On a donc intégré l'ontologie dans la phase de recherche, ce qui permettra de combiner un modèle de recherche d'information classique qui est le modèle vectoriel avec l'exploitation de l'ontologie. Ceci permettra aussi d'améliorer la précision de recherche par la compréhension de son objectif et la signification contextuelle des termes qui mènera rapidement et efficacement à l'information désirée. Le système reçoit alors une requête en langue arabe en entrée et produit en sortie une liste de documents en langue arabe également qui répondent à cette requête.

L'ontologie permet ainsi de présenter ses produits de manière cohérente et structurée facilitant la localisation des documents pertinents ; d'où on a introduit la notion d'ontologie dans les SRI. Elle permet d'étendre le champ de recherche d'une requête, ce qui a pour conséquence l'amélioration des résultats de la recherche.

Notre projet consiste à implémenter un système de recherche d'information en langue arabe à base d'ontologie en utilisant le modèle vectoriel qui dépasse les limites d'une recherche classique. Pour l'implémentation du système, on se servira de l'environnement de

développement constitué du langage de programmation (Java), du Serveur de bases de données (PostgreSQL), l'éditeur d'ontologie Protégé et les API jena.

Ce mémoire est organisé en trois chapitres :

Dans le premier chapitre, nous exposerons les points essentiels du sujet. Tout d'abord, nous parlerons de la recherche d'informations en général, en donnant sa définition, ses principaux acteurs ainsi que les différentes approches utilisées. On exposera par la suite, les systèmes de recherche d'information, leurs définitions ainsi que leur architecture générale. Ensuite, Nous définirons l'étape d'indexation et son processus ainsi que les différents modèles de la recherche d'information, en particulier, le modèle booléen, le modèle probabiliste et le modèle vectoriel. Le dernier point traité dans ce chapitre concernera l'évaluation des systèmes de recherche d'information.

Dans le chapitre suivant, nous décrirons la notion d'ontologie, sa définition, son rôle, ses composantes et langages, ensuite, nous décrirons le principe d'utilisation des ontologies dans la recherche d'information. Nous finirons dans ce chapitre par la description de la langue arabe, les éléments essentiels de sa morphologie ainsi que les différentes méthodes d'extraction de la racine.

Le chapitre trois servira à présenter les détails de la conception de notre système de recherche, sa réalisation complète impliquant la construction et l'utilisation de l'ontologie et les outils adéquats pour son développement ainsi que les résultats obtenus.

Enfin, nous clôturerons ce mémoire par une conclusion qui synthétisera le travail mené dans ce contexte.

CONCLUSION GENERALE

Le travail présenté dans cette thèse s'inscrit dans le contexte général de la recherche d'information. Il a pour but, la réalisation d'un système de recherche à base d'ontologie dont la collection de données et les requêtes d'utilisateur sont en langue arabe.

La langue arabe est l'une des langues les plus largement répandues dans le monde, mais malheureusement, peu d'études ont été faites sur la restitution par des moteurs de recherche de documents dans cette langue.

Aujourd'hui, dans les systèmes de recherche d'informations, la recherche ne se fait plus uniquement sur la base d'une correspondance entre des mots clés (métadonnées) stockés et les mots clés formulés par les utilisateurs. Le processus de recherche actuel est plus intelligent car capable de prendre en compte le sens des mots clés, plutôt que de les considérer comme de simples chaînes de caractères dépourvues de toute signification. C'est là que la notion d'ontologie intervient, en l'organisant sous forme d'un ensemble de concepts (mots-clés) par des relations sémantiques. C'est une façon technique de simuler de la connaissance et prendre l'initiative d'augmenter les résultats de recherche.

Nous pouvons dire que l'une des difficultés remarquables dans notre travail réside dans le traitement de la langue arabe qui possède une morphologie très riche avec des aspects très complexes, tels que la polysémie, les formes irrégulières de certains mots, ses propriétés dérivatives et concaténatives, ainsi que l'ambiguïté lexicale due à l'absence, dans la plupart des cas, de voyelles.

Le traitement morphologique de cette langue s'avère être particulièrement important pour la recherche d'information afin de déterminer une forme appropriée de mots utilisés comme index. Dans ce contexte, la lemmatisation des mots arabes a fait l'objet de plusieurs travaux sur la recherche d'information en langue arabe, desquels nous avons adopté, pour notre travail, la méthode de Shereen khoja [25].

Enfin, nous pouvons dire que nous avons réussi à implémenter un système de recherche d'information utile et efficace basé sur trois grands axes : l'indexation, la recherche et l'ontologie. L'ontologie que nous avons construite et intégrée dans le système proposé permet l'amélioration de la sémantique des documents et des requêtes et faciliter ainsi la recherche dans le SRI.

Ce travail, comme tout autre travail de mémoire, fait émerger plusieurs idées et perspectives susceptibles d'être réalisées pour améliorer le système réalisé. Parmi celles-ci, on peut proposer l'extension du système pour d'autres langues afin de le rendre multilingue, l'enrichissement et l'expansion de l'ontologie en ajoutant d'autres concepts et relations pour avoir plus d'informations sur le domaine, l'intégration d'autres techniques et méthodes de classification du résultat obtenu pour une plus grande précision.

BIBLIOGRAPHIE

- [01] Abbadeni, N., Ziou, D., ang, S. « Recherche d'image basé sur le contenu : Représentation de la texture par le modèle autorégressif ». Rapport technique, université de Sherbrooke, N° 216, 1998.
- [02] Abd el salam, Hajjar. « Extraction et gestion de l'information à partir des documants arabes ». thèse de doctorat en informatique, université de Paris VIII – SAINT DENIS ECOLE DOCTORALE Cognition, pp. 25-34, 2010.
- [03] Adamson, G., Boreham, J. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Journal of Information Storage and Retrieval*, vol 10, N° 7-8, pp. 253 -260, 1974.
- [04] Ahmed, F., and Nürnberger, A. N-Grams Conflation Approach for Arabic Text, In: *Proceedings of the International Workshop on improving Non English Web Searching iNEWS07*) In conjunction with The 30th Annual International (ACM SIGIR Conference). Amsterdam City, Netherlands, pp. 39-46, 2007.
- [05] Al Ameen, S., Al Ketbi, A., Al Kaabi, K., Al Shebli, N., Al Shamsi, N. and Al Nuaimi, S., Al Muhairi, Arabic Light Stemmer: A new Enhanced Approach, *The Second International Conference on Innovations in Information Technology (IIT'05)*, pp. 1-9, 2005.
- [06] Al Dakkak, O., Ghneim, N., Alshalaby, A., Sonbol, R. and Saïd Desouki, M. Arabic Language Resources in HIAST, *Medar 2nd International Conference on Arabic Language Resources and Tools*, 2009.
- [07] Aljlayl, M., and Frieder, O. On arabic search: Improving the retrieval effectiveness via a light stemming approach. In *Proceedings of ACM Eleventh Conference on Information and Knowledge Management*, Mclean, VA, pp. 340-347, 2002.

- [08] Aussenac-Gilles, N., Biébow, B., Szulman, N. « Revisiting Ontology Design: a method based on corpus analysis ». Proc of KAW'2000. Juan-Les-Pins (F). Oct 2000. Lecture Notes in Artificial Intelligence Vol 1937, Springer Verlag, pp. 172-188, 2000.
- [09] Baeza-Yates, R., Ribeiro-Neto B. A. Modern Information Retrieval. Pearson Education Ltd., Harlow, UK, 2nd edn, 2011.
- [10] Baneyx, A. « Construire une ontologie de la pneumologie, aspects théorique, modèles et expérimentations ». Thèse de doctorat, Université Pierre et Marie Curie, 2007.
- [11] Baziz, M. « Indexation conceptuelle guidée par ontologie pour la recherche d'information », Thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2005.
- [12] Boubekeur, F. « Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets », thèse de doctorat en informatique, Université Paul Sabatier, 2008.
- [13] Boucham, S. « Une approche basée Ontologies pour l'indexation automatique et la Recherche d'Information Multilingue (RIM) », thèse de magister en informatique, Université M'hamed Bougara de Boumerdes, 2009.
- [14] Cruzel, S. L. « Conception de systèmes de recherche d'informations : accès aux documents numériques scientifiques ». HDR, Université Claude Bernard Lyon 1, 2001.
- [15] Daoud, M. « Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche », thèse de doctorat en informatique, Université Paul Sabatier, 2009.
- [16] Darwish, K. Building a Shallow Arabic Morphological Analyzer in One Day. The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, USA, 2002.

- [17] De Roeck, A. N., and Al-Fares, W. A morphologically sensitive clustering algorithm for identifying Arabic roots. In Proceedings ACL-2000. Hong Kong, pp. 199 -206, 2000.
- [18] Douzidia, F. S. Résumé automatique de texte arabe, Mémoire de M.Sc en informatique, Université de Montréal, 2004.
- [19] Fox, C. Lexical analysis and stoplists, Frakes W B, Baeza - Yates R (eds) Prentice Hall. New jersey, pp. 102 -130, 1992.
- [20] Goweder, A., Poesio, M., De Roeck, A. and Reynolds, J. Identifying Broken Plurals in Unvowelised Arabic Text". Proceedings of EMNLP, pp. 246-253, 2003.
- [21] Grefenstette, G., Semmar, N., F. and Fluhr, C. Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications. ACL computational approaches to semitic languages workshop ANN ARBOR USA, June 29 2005.
- [22] Gruber, T. « A translation approaches to portable ontology specifications. Knowledge Acquisition », vol 5 N°2, pp. 199-220, 1993.
- [23] Guarino, N., Giaretti, P. « Ontologies and knowledge bases: Towards a terminological clarification ». In Towards Very Large Knowledge Bases. N. J. I. Mars, Ed., IOS Press, pp. 25-32, 1995.
- [24] Hernandez, N. « Ontologie de domaine pour la modélisation du contexte en recherche d'information », Thèse de doctorat en informatique, Université Paul Sabatier, 2006.
- [25] Khoja, S., Garside, R., and Knowles G. « A Tagset for the Morphosyntactic Tagging of Arabic ». Proceedings of the Corpus Linguistics. Lancaster University (UK), 2001.
- [26] Koo, S., Lim, S.Y., Lee, S.J. « Building an Ontology based on Hub Words for Informational Retrieval », In Proceedings of the IEEE/WIC International Conference on Web Intelligence, 2003.

- [27] Krovetz, R. « Homonymy and polysemy in information retrieval ». In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (A CL-97}, pp. 72-79, 1997.
- [28] LV, Y., Zhai, C. « Positional Relevance Model for Pseudo-Relevance Feedback ». Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 579-586, 2010.
- [29] Maron, M. E., and Kuhns, J. L. « On relevance, probabilistic indexing and information Retrieval ». J. ACM, vol 7 N° 3, pp. 216 –244, 1960.
- [30] Marsi, E., Bosch, A.v.d. and Soudi, A. « Memory-based morphological analysis generation and part-of-speech tagging of Arabic ». In Computational Approaches to Semitic Languages Workshop Proceedings, University of Michigan Ann Arbor, Michigan, USA, 2005.
- [31] Mustafa, M., AbdAlla, H. and Suleman, H. « Current Approaches in Arabic IR: A Survey ». In Proceedings The Annual International Conference on Asia-Pacific Digital Libraries (ICADL), Bali, Indonesia. 2008.
- [32] Mesfar, S. « Analyse morpho-syntaxique automatique et reconnaissance des entites nommees en arabe standard ». Thèse de l'université de Franche-Comte, 2008.
- [33] Nathalie, H. « ontologies de domaine pour la modélisation du contexte en recherche d'information ». Thèse pour obtenir le grade de docteur, décembre 2005.
- [34] Neches, R., Fikes, R.E., Finin T, T., Ruber, T.R. Patil, R., Senator, T., W.R. Wartout. « nabling technology for knowledge sharing ». AI Magazine, vol 12 N°3, pp. 16- 36, 1991.
- [35] Noy N. F., McGuinness D. L. Developpement d'une ontologie 101 : Guide pour la creation de votre premiere ontologie. Universite de Stanford, Stanford, CA 94305, 2001.
- [36] Ottens, K. « Un système multi-agent adaptatif pour la construction d'ontologies à partir de textes ». page 14, Octobre 2007.

- [37] Qiu, Y., Frei, H.P. Concept Based Query Expansion. Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160-169, 1993.
- [38] Ren, F., Lan, L., Nie, J-Y. SAAK Approach: How to Acquire Knowledge in an Actual Application System. International Conference on Artificial Intelligence and Soft Computing, Honolulu, pp. 136-140, 1999.
- [39] Rijsbergen, V. « Information Retrieval ». Butterworths & Co, Ltd, London, 1979.
- [40] Robertson, S., Sparck Jones, K. « Relevance weighting for search terms ». Journal of The American Society for Information Science, vol 27, N° 3, pp. 129-146, 1976.
- [41] Roussey, C. « Une méthode d'indexation sémantique adaptée aux corpus multilingues ». Thèse pour obtenir le grade de docteur, décembre 2001.
- [42] Saias, J., Quaresma, P. « A Methodology to Create Ontology-Based Information Retrieval Systems », In Proceedings of the EPIA Conference, pp. 424 - 434, 2003.
- [43] Salton, G., and McGill. Introduction to Modern Information Retrieval. Mc Graw-Hill, New York, 1983.
- [44] Salton, G., A comparison between manual and automatic indexing methods. Journal of American Documentation, vol 20, N°1, pp. 61-71, 1971.
- [45] Sanan, M., Etude Des Méthodes De La Recherche D'information Et De L'indexation Sur Les Documents Electroniques : Cas De La Langue Arabe, Doctorat Informatique, UNIVERSITE PARIS VIII - SAINT DENIS, ECOLE DOCTORALE Cognition, Langage et Interaction (CLI), 2008.
- [46] Sauvagnat, K. « Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés », Thèse en vue de l'obtention du Doctorat de l'Université Paul Sabatier, 2005.

- [47] Semmar, N., Elkateb-Gara, F., and Fluhr, C. Using à Stemmer in a Natural Language Processing system to treat Arabic for Cross-language Information Retrieval. International conference on Machine Intelligence, Tozeur, TUNISIE. 2005.
- [48] Singhal, A., Salton, G., Mitra, M., Buckley, C. Document length normalization. Information Processing and Management, vol 32, N° 5, pp. 619 – 633, 1996.
- [49] Smeaton, A. F. «Information retrieval and natural language processing». In proceedings of a conference jointly sponsored by ASLIB, University of York. page 2, March 1989.
- [50] Tambellini, C. « Un système de recherche d'information adapté aux données incertaines: adaptation du modèle de langue ». Thèse de doctorat en informatique, Université de Nice-Sophia Antipolis-UFR sciences, 2007.
- [51] Tribune des Droits Humains – Genève(TDHG). Censure de l'internet dans les pays arabes, 2006 - www.humanrights-geneva.info, 2006.
- [52] Vallet, D., Fernández, M., Castells, P. « An Ontology-Based Information Retrieval Model », In Proceedings of the 2nd European Semantic Web Conference, pp. 455-470, 2005.
- [53] Van Heijst, G., Schreiber, A., Wielinga, B. « Using explicit ontologies in KBS development ». Int. J. of Human-Computer Studies, pp. 183-292, 1997.
- [54] Welty, C., Guarino, N. Supporting Ontological Analysis of Taxonomic Relationships Data et Knowledge Engineering (39), pp. 51-74, 2001.
- [55] Wikipedia: Arabic_grammar (WAG), website: http://en.wikipedia.org/wiki/Arabic_grammar. 2010.
- [56] Woods, A.W. « Conceptual indexing: A better way to organize knowledge ». Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April 1997. www.sun.com/research/techrep/1997/abstract-61.html.

Résumé :

L'objectif de ce projet est de réaliser un système de recherche d'information en langue arabe à base d'ontologie pour améliorer la description sémantique des documents et des requêtes. Le domaine choisi pour cette dernière est celui de l'université. Le système ainsi proposé, intègre une indexation fondée sur le formalisme du modèle vectoriel et tient en compte d'une part, de la reformulation des requêtes et d'autre part, du classement des résultats obtenus.

Un traitement spécifique de la langue arabe est décrit, visant à décomposer les mots pour obtenir leurs différents constituants servant de mots-clés au système de recherche.

Mots-clés : recherche d'information, indexation automatique, ontologie, modèle vectoriel, langue arabe.

Abstract:

The objective of this project is to realize an information search system in Arabic based ontology to improve the semantic description of documents and queries. The area chosen for the latter is that of the university. The proposed system and integrates indexing based on the formalism of the vector model and takes into account on the one hand, the reformulation of queries and secondly, the ranking results.

Specific treatment of the Arabic language is described, for breaking words for their various constituents using keywords in the search system.

Key words: information retrieval, indexing automatic, ontology, vector model, arabic language.

ملخص:

الهدف من هذه المذكرة هو انجاز نظام بحث عن المعلومات باللغة العربية ذو قاعدة مبنية على الانطولوجيا من أجل تحسين الوصف الدلالي في النصوص و الاستفسارات. المجال المختار لهذه الانطولوجيا هي الجامعة. النظام المقترح, يدمج عملية الفهرسة التي تعتمد على شكلية النموذج الشعاعي ويأخذ في عين الاعتبار كلا من إعادة صياغة الاستفسارات و ترتيب النتائج التي تم الحصول عليها.

المعالجة الخاصة للغة العربية, تهدف الى تحليل الكلمات من اجل الحصول على مختلف مكوناتها لاستخدامها في الكلمات المفتاحية لنظام البحث.

الكلمات المفتاحية : البحث عن المعلومات , الفهرسة الآلية , الانطولوجيا , النموذج الشعاعي , اللغة العربية.