International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-2, Issue-11)

Research Article | **November 2013**

A New Developed Model for Arabic Information Retrieval System based on Knowledge Base System

Mohamed A. Abdelhadi [*], **Tiruveedula Gopi Krishna**
Department of Computer Science
Faculty of Arts & Science, Hoon
Sirt University, Libya

Ghassan Kanann
Department of Comp. Science & Information System,
Univ.of Banking & Financial Sciences
Faculty of Technology & Infm. Jordan, Amman

Abstract—

Optimization in Arabic Information Retrieval is new area to study and implement. We have chosen this area because of its importance and growth need. Knowledge Base is the most important key for optimization. We have selected some efficient methods used to optimize the retrieval techniques by means of Knowledge Base methods. One of the best choices was to use the benefit of expert system technique to optimize the system queries and its retrieval. Our KBS-Model has two new advantages in advance by using (HML-classification method) and (Hill-Climbing-Search-Technique) with respect to other models implemented to deal with Arabic Information Retrieval Systems.

Keywords— Arabic Language, Information Retrieval, Knowledge Base System, Expert System, Optimization.

I. INTRODUCTION

Optimization in general is defined to be as certain methods used to develop and scale up any System productivity in term of reliability and efficiency. Knowledge Base is the most important key for optimization. Many Expert Systems are not difficult to realize. An Expert System usually has the ability to explain why data is needed and what and/or how conclusions were reached [1], [4]. A system can normally be highly interactive (directly asking the user questions) or embedded where all input comes from another program. The range of problems that can be handled by Expert Systems became vast. Expert systems can be developed with Expert System Shells. An Expert System Shell is a Software Programming Environment which enables the construction of Expert or Knowledge Based Systems. Most Expert Systems Software can be developed for any problem that involves a selection from among a definable group of choices where the decision is based on logical steps [2], [5]. Any area where a person or group has special expertise needed by others is a possible area for an Expert System. Expert Systems can help Automate anything from complex Regulations to aiding Users in selecting from among a group of Products, or diagnosing equipment problems. Expert Systems have become a popular Method for representing large bodies of Knowledge for a given field of Expertise and Solving Problems by use of this Knowledge. An Expert System often consists of three parts, namely: a Knowledge Base, an Inference Engine, and a User Interface, a Dialogue is conducted by the User Interface between the User and the System. The User provides Information about the Problem to be solved and the System then attempts to provide insights derived (or inferred) from the Knowledge Base. These insights are provided by the Inference Engine after examining the Knowledge Base [39], [40], [45].

II. RELATED WORKS

Raid Al-Shalabi and et.al [2006]; have studied and compared the performance of a search engine before and after expanding the query through Interactive Word Sense Disambiguation (WSD). Their objectives were to focus mostly on studying how effective and efficient expanding queries through disambiguating the word senses via user feedback on an Arabic IR system. Alaa M. El-Halees [2007]; has studied and test out the Arabic Text Classification Using Maximum Entropy. His Research paper focuses on classifying Arabic text documents. In his approach, he first preprocessed data using natural language processing techniques such as tokenizing, stemming and part of speech. Then, he used maximum entropy method to classify Arabic documents. Abdelhadi Soudi, Antal van den Bosch, and Guenter Neumann [2007]; has introduced and studied by publishing a book which explained the need to covering the area of an increased interest in Arabic natural language processing, and in particular computational morphology. Ezra Daya [2002]; has in his thesis as declared that the study of advanced computational linguistics techniques to an investigation of certain linguistic features of Arabic dialects spoken in Northern Palestine. His study was about the use of computational morphological analyser for the dialects, based on finite-state technology, and uses it in order to process a corpus of transcribed texts. Hend S. Al-Khalifa and Areej S. Al-Wabil [2007]; have in their research investigated the Semantic Web technologies which are used since 90's to processing Latin family scripts, thus, an apparent lack of Arabic script support in these technologies kept the research in the Semantic Web for the Arabic language not tackled. Karim Bouzouba and Adil Kabbaj [2007]; were interested in natural language processing; they had studied the morphological system, syntactic system, semantic system. In their research, they have introduced a platform as a tool for developing intelligent systems which can be used as an integrated platform to deal with all aspects of the Natural language Processing. Venus Samawi1, Akram Mustafa, and Abeer Ahmad [2013], they have presented in their research a system which has provided with the required domain

dictionary to be used by the Arabic morphological system. They have studied the benefit of merging of morphological system with knowledge acquisition.

III. BLOCK DIAGRAM

A. *A developed Model for Arabic IRs based on KBS*

The new proposed Model consisting of three components as shown in the block diagram figure1. Our KBS-Model has two new advantages in advance by using (HML-Classification Method), [34], [35], [37] and (Hill-Climbing-Search-Technique). We have built our KBS-Model in three phases to ensure that each phase has passed all design and analysis requirements. Our KBS-Model has two new advantages in advance by using (HML-Classification Method) and (Hill-Climbing-Search-Technique). We have built our KBS-Model in our proposed model to ensure that each phase has passed all design and analysis requirements. We have tested each phase during the research work and in each phase we have evaluated the System Performance with respect to the System-Retrieval –Optimization. The evaluation of Arabic-IRs based on KBS-Model; we have done many Tests on different Arabic queries related to the most important topics in our Test-Corpus. The results will be explained in the section of Results analysis.

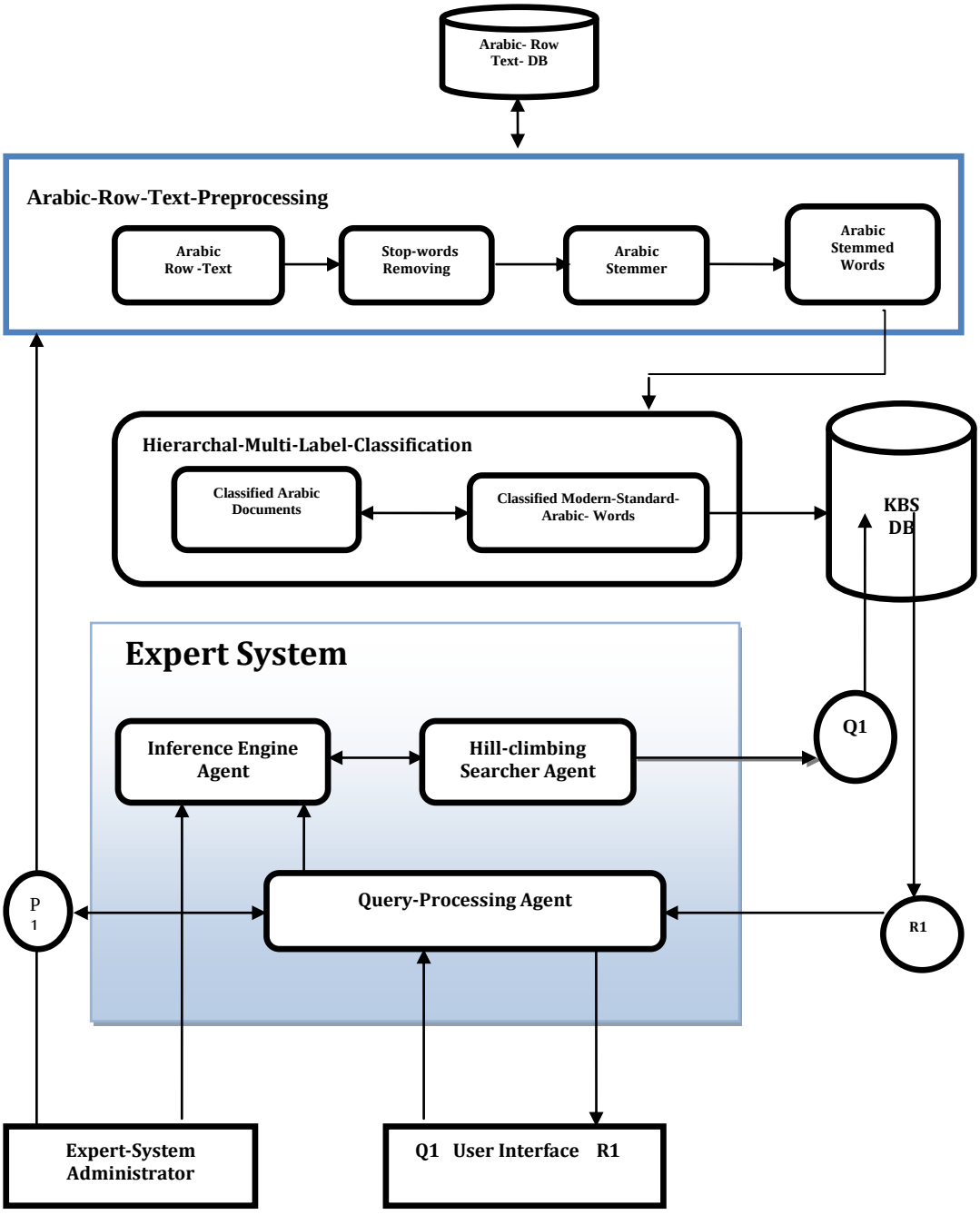

Fig1: Arabic-Irs Model Based On KBS Block Diagram

B. Arabic-IRs based on KBS Preprocessing (Text Document Preprocessing)

We have built our Data Base Model to explain how the Arabic–IRs preprocessing phase has been analyzed. In this phase we have filtered Arabic Document for the following:

a. Stop Words.
b. Raw text.

Any Document will be inserted into the system can be processed to make the three main steps:

1- Remove all Special characters, numbers and non-Arabic letters (Stemming).
2- Remove all Stop Words.
3- Save all raw Words into an array initially.

To filter the Arabic Row-Text we have designed special enhancement on Regular Expression Parser for extracting Arabic words and also to remove the Arabic stop-words from the collected Arabic Row-Text. This Parser is based on the technique of Regular Expression which has the facility of using word pattern extraction as it has been stated in [Beesley, Kenneth R. 1996].

We have used the benefit of that technique in our Arabic Retrieval System to remove the Arabic stop-word and also to stem the Arabic words, finally we have found that there were possibilities to improve our system by using the New Parser in extracting Arabic words Roots too. After having especial studies on the Arabic Roots and how can we extract the Root from it; we found that there is a specific 3 or 4 letters and the other letters are boundaries Letters, so we can make a RegEx pattern for each Arabic-Words- Roots then we can comparing the Roots within the Text to remove the boundaries Letters.

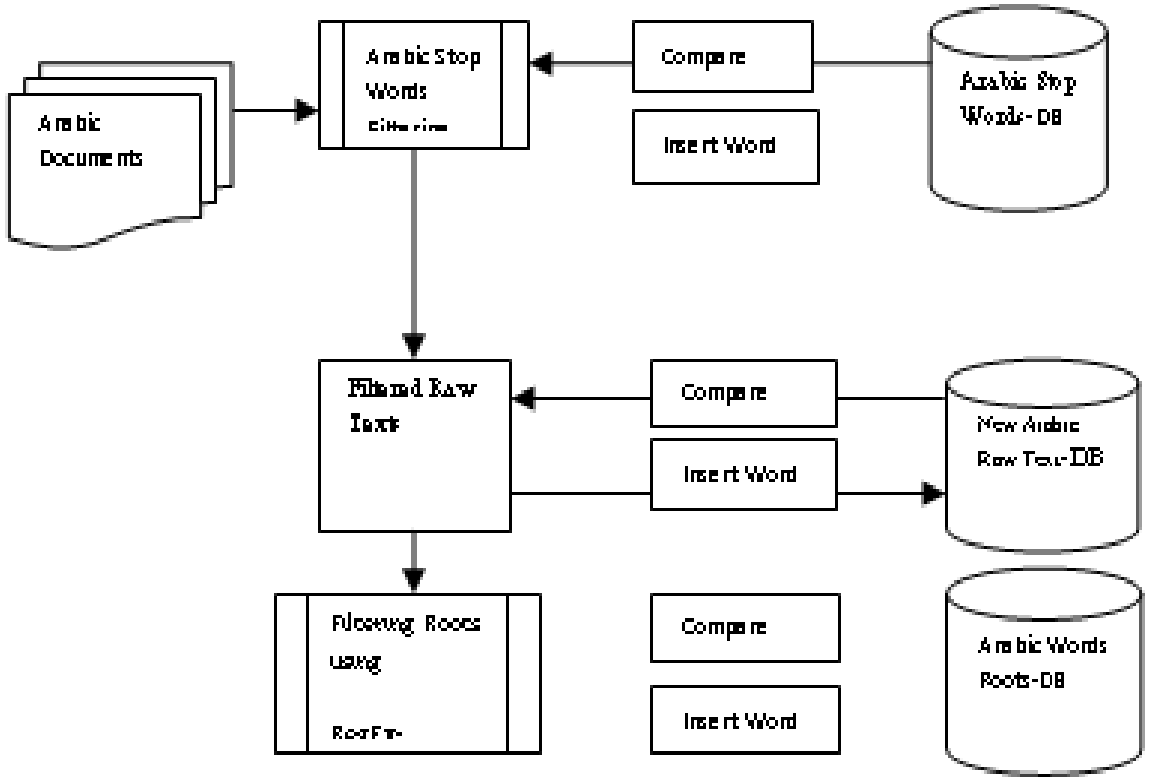

Fig2. Arabic- IRS –Pre-processing block-diagram

C Hierarchical-Multi-Label-Classification

Text categorization is defined as a process of classifying text document into some pre-defined categories [34], [35]. It is usually formulated as binary classification which has a binary classifier; and can be designed for each category of interest. Multi-class or multi-label decisions are the classes which based on how well the input text document matches with the set of multiple binary classifiers. multi-class classification approach are used in many pattern Recognition problems, such as Automatic Speech Recognition, Speaker Identification, Face Recognition and Optical Character Recognition. Classification refers to the task of learning from a set of classified instances a model that can predict the class of previously unseen instances and also used in Text classification. Hierarchical-Multi-Label-classification (HMLC) differs from normal classification in two ways, (1): a single example may belong to multiple- classes simultaneously; and (2): the classes are organized in a hierarchy: an example that belongs to some class automatically belongs to all its super classes (the hierarchy constraint); [37], [38], [44].

1) Definitions

1- Single-label classification: – set of examples associated with a single label (l)-from a set of disjoint labels L, |L|>1.(2)- If |L|=2, then it is called a binary classification problem.(3)- While if |L|>2, then it is called a multi-class classification problem.(4)- In multi-label classification, the examples are associated with a set of labels {Y belong or equal to L}.(5)- For an input X the corresponding output is a vector Y= {y1, y2, yL} ^T.

2) Single-label (Multi-class) Classification

Examples: D = {X1, Xn} Labels: L = {l1, lm} .Each example is associated with one label: (X, l belongs to L).

3) Multi-label Classification

Examples: D = {X1, Xn} Labels: L = {l1, lm}. Each example is associated with a subset of labels: (x, S belongs or equal to L.).

D. Evaluation on Classification

TP is the number of True Positives (correctly predicted positive examples) and FP is the number of False Positives (positive predictions that are incorrect) and FN is the number of False Negatives (positive examples that are incorrectly predicted negative) [6],[7]. Precision and Recall are traditionally defined for a binary Classification task with Positive and Negative classes. Precision is the proportion of Positive predictions that are correct, and Recall is the proportion of Positive examples that are correctly predicted Positive. 1- Precision = TP/ TP + FP. 2- Recall = TP / TP + FN.
F-Measure =2 x (Precision x Recall) / (Precision + Recall),[8].

E. Evaluation for System Performance by Hierarchal Multi-Label-Classification

As shown in the figure2 below, we have used in our Arabic-IRs model the technique of the Hierarchal -Multi-label-Classification to assess the performance of the system with respect to the classification method used in our Arabic-IRs. We have used the performance measurement methods, which has the ability to measure the Recall\precision by calculating the percentage of all possible solutions needed for the predicted outcome values by which the whole categorized words evaluated.

TABLE I
CLASSIFICATION PERFORMANCE EVALUATION

no	Category	Total Docs In Arabic=104	Total Words	Total Retrieved Docs in CAT	Total Retrieved Docs	True Positive	False Positive	False Negative	Recall	Precision	F-Measure
1	Economic	19	654	18	19	18	19	1	0.94	0.49	0.6
2	Arts	54	844	42	54	42	54	12	0.78	0.44	0.6
3	History	3	4457	2	3	2	3	1	0.67	0.4	0.5
4	Politics	24	12536	18	24	18	24	6	0.75	0.4	0.55
5	Sports	4	930	4	4	4	4	0	1	0.5	0.67

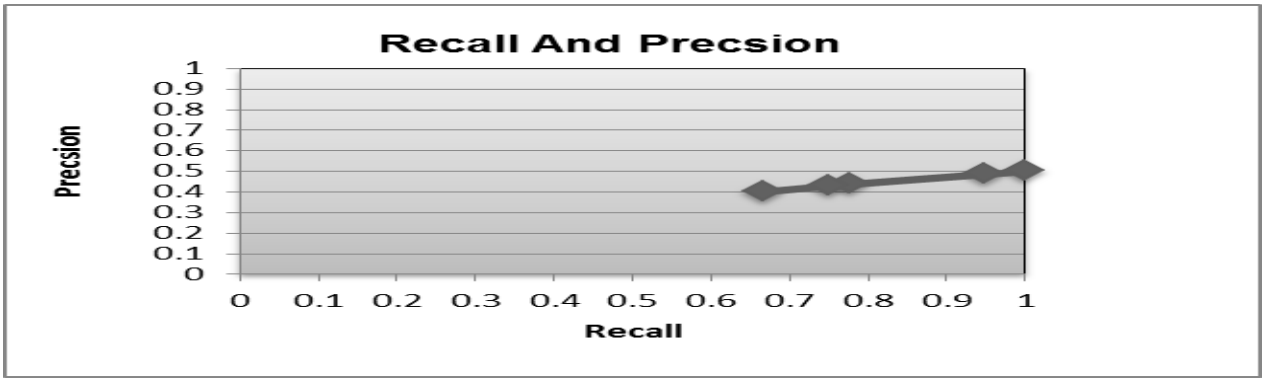

Fig3. Recall and Precision for Arabic-IRs based on Classification

IV. PERFORMANCE EVALUATION

Arabic-IRs Model based on KBS Performance Evaluation

The most important results analysis for our new proposed Arabic-IRs model was to evaluate the total of 163 Arabic documents selected to be as Test-Corps; which were divided into different categories to measure performance in both sides as document retrieval and query relevancy. The level of Retrieval by new proposed Arabic-IRs were exact and more a curate. Figure2 has showed us that the level of Recall and Precision, which has surprised us in its accuracy by reaching to level of (0.98:1.0); the level of optimality in both sides (Recall=0.98 and Precision=1.0) in some cases.

TABLE II
EXPERT SYSTEM PERFORMANCE EVALUATION

Queries in Arabic / Total Documents=163	Relevant	Retrieved	Relevant ‖ Retrieved	Recall	Precision	F-Measure
Politics and history of Sciences	55	38	34	0.62	0.89	0.73
Islamic Policy	09	40	40	0.68	1.00	0.81
Alcohol and Lever Disease	80	59	59	0.74	1.00	0.85
Physics Scientist	45	49	36	0.80	0.73	0.77
Economic of Islamic World	97	80	80	0.82	1.00	0.90
Engineering of Computer	07	08	06	0.86	0.75	0.80
Islamic Scientists	61	71	56	0.92	0.79	0.85
Islam and Hajj	28	26	26	0.93	1.00	0.96
Heart Disease	56	55	55	0.98	1.00	0.99
			Average:=	0.82	0.91	0.86

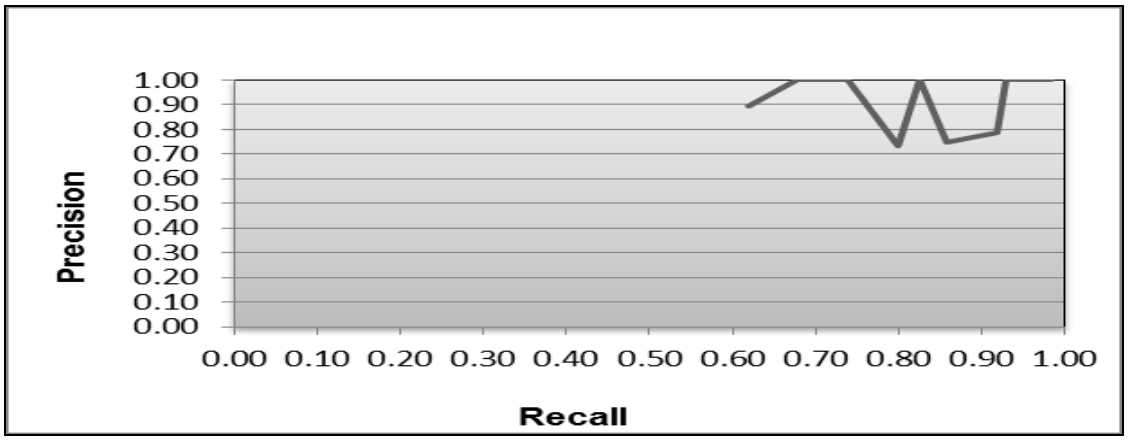

Fig4. Recall and Precision for Arabic-IRs based on KBS Evaluation

V. CONCLUSIONS

We have evaluated our new developed Arabic-IRs-Model based on KBS by analyze the results expected by using an excellent Hierarchal-Multi-Label-Classification which has improved our new model %100 in Arabic-Stemmed-Words classification; this has led to improve our new model by its results of Figure2 which has showed us the level of Recall and Precision. In Arabic-stemmed-words classification were evaluated by ran some Arabic queries on the system

retrieval optimization. Finally the most important results analysis for our new Arabic-IR model was to evaluate the Performance in both sides as document retrieval and query relevancy. We have not handled the Multi-Media-Files, because of some obstacles that we have met in our research work; especially the Arabic-Multi-Media-Tools which are not available to be used to extract such kind of Media files. Our presented work was an Arabic-IR based on KBS; as we have done, it was for Desktop purposes programmed but it would be in our Future work as Web-based-Oriented.

ACKNOWLEDGMENT

I would like to thank all of my co-others who helped me in the practical sessions to fulfil this research work.

REFERENCES

[1] Hussein O., Monzer Q., and Hazim F.,"Framework Model for Shell Expert System, "International Journal of Computer Science and Network Security, vol. 9, no. 11, pp. 56-68, 2009.
[2] Joseph G. and Gary R., Expert Systems: Principles and Programming, Thomson Course Technology, 2004.
[3] Keith D., the Essence of Expert System, Pearson Education Limited, 2000.
[4] Kiong W., Abd-latif R., Mohd Z., and Azwan A., "Expert System in Real World Applications,"Available at: http://www.generation5.org/content/2005/Expert_System.asp, last visited 2005.
[5] Pamela G. and Xenogeny G., "A Map-Based Expert-Friendly Shell," Bouncier D., (ED.), in Proceedings of Legal Knowledge and Information Systems, Amsterdam, pp. 51-60, 2003.
[6] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison-Wesley, 1999.
[7] S. E. Robertson, S. Walker, and M. Beaulieu, Okapi at TREC–7: automatic ad hoc, filtering, VLC and filtering tracks. In Proceedings of 7th Text Retrieval Conference (TREC-7), pages 253–264. NIST special publication, 1999.
[8] Khoja, S. and Garside, R. "Stemming Arabic Text", Computing Department Lancaster University, Lancaster, 1999.
[9] *A. Sajjanhar, J. Hou, and Y. Zhang." Algorithm for web services matching". In Advanced Web Technologies and Applications, pages 665–670. Springer Verlag 2004*.
[10] R. Sindhgatta." Using an information retrieval system to retrieve source code samples". In Proceedings of 28th IEEE/ACM International Conference on Software Engineering, pages 905–908, Shanghai, China, 2006. IEEE CS Press.
[11] A. Singhal, C. Buckley, and M. Mitra. "Pivoted document length normalization". In Proceedings of 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 21–29, Zurich, Switzerland, 1996. ACM Press..
[12] A. Singhal, J. C. D. Hindle, D. Lewis, and F. Pereira. At &t at TREC-7. In Proceedings of 7th Text Retrieval Conference (TREC-7), pages 239–252. NIST Special Publication, 1999.
[13] A. Singhal, "Modern information retrieval: A brief overview", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4), 35–43, 2001.
[14] Abdelhadi Soudi, Antal van den Bosch, and Gunter Neumann, "Arabic Computational Morphology: Knowledge-Based and Empirical Methods", 2007. Ezra Daya," Computational Investigation of Palestinian Arabic Dialects", 2002.
[15] Jacques Savoy, Yves Rasolofo," Report on the TREC-11 Experiment: Arabic, Named Page and Topic Distillation Searches", Jacques Savoy, Yves Rasolofo, 2002.
[16] Kareem Darwish and Douglas W. Oard, CLIR Experiments at Maryland for TREC-2002: "Evidence combination for Arabic-English retrieval", 2002. Aitao Chen and Fredric Gey "Building an Arabic Stemmer for Information Retrieval", 2002.
[17] Paul McNa, JHU/APL at TREC 2002:" Experiments in Filtering and Arabic Retrieval", 2002.
[18] Karim Bouzouba;" an integrated development platform for Arabic language Processing", 2007.
[19] Riyad Al-Shalabi, Ghassan Kanaan, Mustafa Yaseen; ,Bashar Al-Sarayreh4 and Nada A. Al-Naji "Arabic Query Expansion Using Interactive Word Sense Disambiguation", 2006.
[20] Ibrahiem M.M. El and Ja'far Atwan; "Designing and Building an Automatic Information Retrieval System for Handling the Arabic Data", 2005.
[21] Alaa M. El-Halees; "Arabic Text Classification Using Maximum Entropy", 2007.
[22] Abdelali, A. Localization in modern standard Arabic, Journal of the American Society for Information Science and Technology, 55, 1, 23- 28.2004.
[23] Suleiman, Y," The Arabic Language and National Identity". Washington, D.C. Georgetown University Press.2003.
[24] Beesley, Kenneth R. , "Arabic Finite-State Morphological Analysis and Generation". In COLING 1996: The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 89-94, 1996.
[25] Beesley, Kenneth R.," Arabic Morphological Analysis on the Internet". In The 6th International Conference and Exhibition on Multilingual Computing, Cambridge, UK.,1998.
[26] Beesley, Kenneth R. , "Finite-State Morphological Analysis and Generation of Arabic at Xerox Research": Status and Plans in 2001. In The ACL 2001Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France,2001.

[27] Buckwalter, Tim.," Buckwalter Arabic Morphological Analyzer Version 1.0. In Linguistic Data Consortium. ", Catalog number LDC2002L49, and ISBN 1-58563-257-0.2002.
[28] Buckwalter, Tim.," Issues in Arabic Orthography and Morphology Analysis". In The Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, pp. 31-34.
[29] Ditters, Everhard. ,"A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic". In Workshop on Arabic Processing: Status and Prospects at ACL/EACL, Toulouse, France,2001.
[30] Stephen D. Bay." Combining nearest neighbor classifiers through multiple feature subsets". In Proceedings of the 17th International Conference on Machine Learning, pages 37–45, Madison, WI, 1998. Erin J. Bredensteiner and Kristin P. Bennett. Multicategory classification by support vector machines. Computational Optimization and Applications, 12:53–79, January 1999.
[31] Yangchi Chen, M. Crawford, and J. Ghosh." Integrating support vector machines in a hierarchical output space decomposition framework". In Proceedings of Geoscience and Remote Sensing Symposium, volume 2, pages 949–952, 2004.
[32] Zafer Barutcuoglu, Robert E. Schapire, and Olga G. Troyanskaya. "Hierarchical multi-label prediction of gene function. Bioinformatics", 22(7):830–836, 2006.
[33] H. Blockeel, M. Bruynooghe, S. Dˇzeroski, J. Ramon, and J. Struyf, "Hierarchical multi-classification". In Proceedings of the ACM SIGKDD 2002 Workshop on Multi-Relational Data Mining (MRDM 2002), pages 21–35, 2002.
[34] Blockeel, L. De Raedt, and J. Ramon," Top-down induction of clustering trees". In Proceedings of the 15th International Conference on Machine Learning, pages 55–63, 1998.
[35] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "Learning hierarchical multi-category text classification models". In L. De Raedt and S. Wrobel, editors, Proceedings of the 22nd International Conference on Machine Learning, pages 744 – 751. ACM Press, 2005.
[36] Klaus Brinker, Johannes Furnkranz, and Eyke Hullermeier," A united model for multi-label classification and ranking"; In Proceedings of the 17th Euro-pean Conference on Artificial Intelligence (ECAI '06), pages 489-493, Riva del Garda, Italy, Aug/Sept 2006.
[37] Pohl. J, "Intelligent Software Systems in Historical Context'; in Jain L. and G. Wren (eds.); 'Decision Support Systems in Agent-Based Intelligent Environments', Knowledge-Based Intelligent Engineering Systems Series, Advanced Knowledge International (AKI), Sydney, Australia, 2005.
[38] Pohl. J, "Knowledge Management Enterprise Services (KMES): Concepts and Implementation Principles"; InterSymp-2007, Proceedings Focus Symposium on Representation of Context in Software, Baden-Baden July 31, Germany, 2007.
[39] Juho Rousu, Craig Saunders, Sándor Szedmák, John Shawe-Taylor: Kernel-Based Learning of Hierarchical Multilabel Classification Models. Journal of Machine Learning Research 7: 1601-1626 (2006) .
Cesa-Bianchi, N., Gentile, C., Tironi, A., & Zaniboni, L," Incremental algorithms for hierarchical classification". Neural Information Processing Systems, 2004.
[41] Cesa-Bianchi, C. Gentile, and L. Zaniboni Incremental Algorithms for Hierarchical Classification. Journal of Machine Learning Research, 7:31—54,2006