Query Optimization Using Genetic Algorithms in the Vector Space Model

Eman Al Mashagba[1], Feras Al Mashagba[2] and Mohammad Othman Nassar[3]

[1] Computer Information Systems, Irbid Private University, Irbid, 22110, Jordan

[2] Computer Information Systems, Amman Arab University, Amman, 11942, Jordan

[3] Computer Information Systems, Amman Arab University, Amman, 11942, Jordan

Abstract

In information retrieval research; Genetic Algorithms (GA) can be used to find global solutions in many difficult problems. This study used different similarity measures (Dice, Inner Product) in the VSM, for each similarity measure we compared ten different GA approaches based on different fitness functions, different mutations and different crossover strategies to find the best strategy and fitness function that can be used when the data collection is the Arabic language. Our results shows that the different GA approaches have differences in their results, the best IR system found is the one that uses the Inner Product similarity as a fitness with one-point crossover operator.

Keywords: *information retrieval, vector space model, query optimization, genetic algorithms.*

1. Introduction

Information retrieval (IR) can be defined as the study of how to determine and retrieve from a corpus of stored information the portions which are responsive to particular information needs [1]. IR is also concerned with text representation, text storage, text organization, and the retrieval of stored information items that are similar in some sense to information requests received from users. The major information retrieval model includes: the vector space model, Boolean model, Fuzzy sets model and the probabilistic retrieval model. These models are used to find the similarity between the query and the documents in order to retrieve the documents that represent the query. Vector space model usually use Cosine, DICE, Jaccard, or Inner Product as a similarity measures. The similarity then used to evaluate the effectiveness of IR system using two measures: Precision which is a ratio that compares the number of relevant documents found to the total number of returned documents [2], and Recall which is the system's ability to retrieve all related documents of a query [2]. The IR models have the local optimal solution problem that can be solved using GA.

A GA is a heuristic search algorithm based on the natural selection and genetics ideas [3]. The GA approach has gained importance and popularity because of its ability in finding a global solution in many difficult problems such as NP-hard problems. In this paper; and for each similarity measure (Dice, and Inner Product) in the vector space model we will implement and compare ten different genetic algorithms settings with different crossover techniques, mutation techniques, and fitness functions to optimize the user query. As a test collection; we are going to use an Arabic data collection which was presented for the first time by [16]; this collection contains 242 documents and 59 queries, the correct answer (relevant documents) are known in advanced for each query.

More than twenty one Arab countries uses the Arabic as the official language, and more than one billion Muslims around the world use it as their religious language. The difficulty of Arabic language compared to Indo-European languages comes from syntactic, morphologic, and semantic differences [13]. Arabic is more sparsed than English language which negatively affect the retrieval quality in Arabic language, Sparseness means that for the same text length, English words are repeated more often than Arabic [14, 15]. In written Arabic, many forms of writing for the same letter are exist. Letters in Arabic can have punctuation associated with them, this may change the meaning of two identical words. Finally; Arabic roots are more complex than English roots.

The special properties for the Arabic language compared to other languages, and the absence of similar studies in the literature is our driver to conduct this study based on Arabic data collection.

2. Previous Studies

There are several studies that used GA in information retrieval systems to optimize the user query based on English data collections such as [4, 5, 7, 6, 7, 5, 9, 10, 11, 12, 18].

In their experiments for the VSM [8,4,6], the authors presents many methods: the connectionist Hopfield network; the symbolic ID3/ID5R, evolution- based genetic algorithms, symbolic ID3 Algorithm, evolution-based genetic algorithms, Simulated Annealing, neural networks, genetic programming. The previous mentioned techniques are promising in their ability to analyze user queries and information needs, they can also suggest alternatives for the user search queries.

In [9, 11, 7, 5,12] different mutation probabilities, new crossover operation, new fitness functions for the GA have been tested to improve the IR results in the VSM. Mercy and Naomie [10]; and based on Vector Space Model and Probability Model they propose a data fusion approach that combine the retrieval status values, they used GA to find the most suitable linear combination of weights assigned to the scores of different retrieval system to get the best optimal retrieval performance.

Improving the performance of Arabic information system using GA is rare in the literature. In [17] an Arabic information retrieval system based on vector space model and GA was created to enhance the performance. The researchers used an adaptive matching function, which obtained from a weighted combination of four similarity measures (Dot, Cosine, Jaccard and Dice).

Arabic data collections gained a very little focus in the literature, and since the information retrieval (IR) is one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency, it is important to conduct a comprehensive comparison between different genetic algorithms settings for each similarity measure in the VSM to find the most useful setting when used with the Arabic data collections.

3. Vector Space Model (VSM)

The vector space model (VSM) is an IR model that represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents [27].

In the vector space model the query and the document vectors are usually compared using different similarity measures [27] such as Cosine, DICE, Jaccard, and Inner Product, those similarity measures are shown in Table 1. $W_{i,j}$ in table 1 are the weights of the i^{th} term in document j, and in the query respectively.

4. Genetic Algorithms (GA)

The basic concept of GA is designed to simulate processes in natural systems necessary for evolution. As such they represent an intelligent exploitation of a random search within a defined search space to solve the given problem.

GAs uses the idea of the survival of the fittest individuals within a given population. A population of strings (solutions to a specified problem) are created and maintained by the GA. The GA then iteratively creates new populations from the old by ranking the strings, then choose the fittest for interbreeding to create new strings that are hopefully will be closer to the optimum solution for the problem.

The GA algorithm flowchart is illustrated in Figure 1. Genetic algorithm operations can be used to generate new and better generations. As shown in Figure 1 the genetic algorithm operations include:

A. The selection of the fittest individuals using the fitness function; this is called Reproduction.

B. The exchange of genes between two individual chromosome; this is called Crossover. There are many crossover strategies like n-point crossover [11], restricted crossover [7], uniform crossover [30], fusion operator [7] and dissociated crossover [7]. For the details about the mentioned crossover strategies you can see the related references.

C. The process of randomly altering the genes in a particular chromosome is called Mutation. In mutation there are two types of mutation:

   1) Point mutation: in this type of mutation single gene is changed.
   2) Chromosomal mutation: in this type of mutation a number of genes is changed completely.

Table 1: Different Similarity Measures.

Similarity Measure	Evaluation for Binary Term Vector	Evaluation for Weighted Term Vector
Cosine	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert^{1/2} \bullet \lvert q \rvert^{1/2}}$	$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$
Dice	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2}$
Jaccard	$sim(d,q) = \frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert - \lvert d \cap q \rvert}$	$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j} \times w_{i,q}}$
Inner Product	$\lvert d_i \cap q_k \rvert$	Sim= $\sum_{k=1}^{t} (d_{ik} \bullet q_k)$

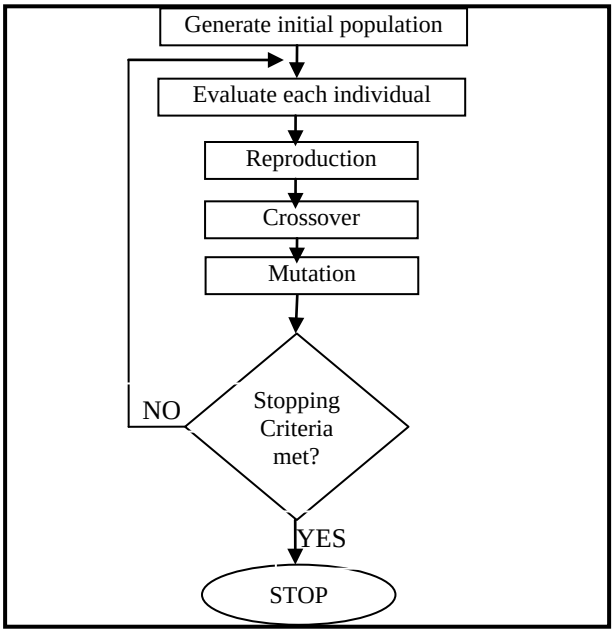

Fig. 1: Flowchart for Typical genetic Algorithm.

5. Experiment

In this study we used the same IR system that was built and implemented by Hanandeh [6] to deal with the 242 Arabic abstracts collected from the Saudi Arabian National Conference [16]. In this study we extracted the significant terms from the relevant and irrelevant documents then we assign weights to those extracted terms. A query vector is created using the binary weights of the terms, and then each query vector considered as a chromosome, then the GA is applied and considered with the final goal of getting an optimal or near optimal query vector, finally the result of the GA approach is compared with the result of the traditional IR system without using a GA to find differences.

This study was conducted using the following approach:

1) The chromosomes are represented as following:

a) Binary representation is used for the chromosomes, then a random function is used to convert the chromosomes to a real representation.
b) We will use the same number of genes as the query and the feedback documents.
c) The size of the chromosomes will be chosen to be equal for the number of terms in the following set (feedback documents set + the query set).
d) The query vector will be represented as a binary vector.
e) Terms will be modified using random function.
f) All of our proposed GA approaches will receive an initial population chromosomes equal to 15 , these 15 chromosomes are to the top 15 documents retrieved from traditional IR with respect to that query.

2) We will use Dice, and Inner Product similarity measures as fitness functions in this study.

3) the selection process for the chromosomes depends on the fitness function. The higher values of the fitness function the higher probability to be selected in the next generation.

4) In this study, two GA operators are used to produce offspring chromosomes, they are:

   1. Crossover: it is used to mix two chromosomes together to form new one. In this paper crossover occurs with probability of Pc (Pc=0.8). In this study; different crossover strategies were used for the VSM :
      a) One-point crossover operator.
      b) Restricted crossover operator.
      c) Uniform crossover operator.
      d) Fusion operator.
      e) Dissociated crossover.

   2. Mutation involves the modification of the gene values of a solution with a probability Pm. In this paper we used a mutation probability equals to (Pm=0.7), also we adapted two different mutation strategies, they are:
      a) Point mutation.
      b) Chromosomal mutation.

Finally we used the previous information to create a 10 GA strategies for each similarity measure (Dice, and Inner Product) in the VSM, Those strategies are as following:

1) GA1: GA that use one-point crossover operator and point mutation.
2) GA2: GA that use one-point crossover operator and chromosomal mutation.
3) GA3: GA that use restricted crossover operator and point mutation.
4) GA4: GA that use restricted crossover operator and chromosomal mutation.
5) GA5: GA that use uniform crossover operator and point mutation.
6) GA6: GA that use uniform crossover operator and chromosomal mutation.
7) GA7: GA that use fusion operator and point mutation.
8) GA8: GA that use fusion operator and chromosomal mutation.
9) GA9: GA that use dissociated crossover and point mutation.
10) GA10: GA that use dissociated crossover and chromosomal mutation.

6. Results for the GA strategies Using Dice Similarity

The results for the GA strategies using Dice similarity are shown in Table 2 and Table 3. From those tables we notice that GA1, GA2, GA4, GA5, GA7, GA8, GA9 and GA10 give a high improvement than traditional IR system with 2.726679%, 4.256249%, 3.051032%, 5.940507%, 5.98964%, 6.095792%, 10.83388% and 9.757293% respectively while GA3 and GA6 give a low improvement than traditional IR system with -1.19504% and -4.68231% respectively. Which means that GA9 that use dissociated crossover and point mutation gives the highest improvement over the traditional approach with 10.83388%.

7. Results for the GA strategies Using Inner Product Similarity

The results for the GA strategies using Inner product similarity are shown in Table 4 and Table 5. From those tables we notice that GA1, GA2, GA3, GA4, GA5,GA9 and GA10 give a high improvement than traditional IR system with 11.9444%, 3.355853%, 3.271745%, 3.203264%, 2.912908%, 5.074422% and 6.307254% respectively while GA6, GA8 and GA9 give a low improvement than traditional IR system with -2.71346%, -2.32334% and -3.60072% respectively. This means that GA1 that use GA that use one-point crossover operator and point mutation gives the highest improvement over the traditional approach with 11.9444%.

Table 2: Average Recall and Precision Values for 59 Query by Applying GA's on Dice Similarity.

Recall	Dice	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	0.131	0.133	0.143	0.129	0.132	0.134	0.136	0.146	0.134	0.141	0.139
0.2	0.172	0.173	0.177	0.165	0.183	0.191	0.187	0.187	0.182	0.197	0.193
0,3	0.262	0.266	0.268	0.254	0.277	0.288	0.177	0.274	0.285	0.298	0.297
0.4	0.214	0.233	0.225	0.212	0.221	0.242	0.232	0.229	0.254	0.277	0.278
0.5	0.357	0.367	0.387	0.363	0.366	0.376	0.223	0.393	0.399	0.402	0.393
0.6	0.379	0.388	0.389	0.385	0.389	0.384	0.391	0.387	0.389	0.408	0.397
0.7	0.383	0.389	0.401	0.375	0.386	0.387	0.386	0.399	0.387	0.396	0.401
0.8	0.388	0.395	0.399	0.387	0.398	0.403	0.394	0.405	0.402	0.412	0.414
0.9	0.431	0.446	0.432	0.422	0.443	0.455	0.437	0.437	0.432	0.441	0.431
Average	**0.3018**	**0.31**	**0.3134**	**0.2991**	**0.3105**	**0.3177**	**0.2847**	**0.3174**	**0.3182**	**0.3302**	**0.327**

Table 3: GA's Improvement in Dice Similarity (GA's Improvement %).

Recall	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	1.526718	9.160305	-1.52672	0.763359	2.290076	3.816794	11.45038	2.290076	7.633588	6.10687
0.2	0.581395	2.906977	-4.06977	6.395349	11.04651	8.72093	8.72093	5.813953	14.53488	12.2093
0,3	1.526718	2.290076	-3.05344	5.725191	9.923664	-32.4427	4.580153	8.778626	13.74046	13.35878
0.4	8.878505	5.140187	-0.93458	3.271028	13.08411	8.411215	7.009346	18.69159	29.43925	29.90654
0.5	2.80112	8.403361	1.680672	2.521008	5.322129	-37.535	10.08403	11.76471	12.60504	10.08403
0.6	2.37467	2.638522	1.583113	2.638522	1.319261	3.166227	2.110818	2.638522	7.651715	4.74934
0.7	1.56658	4.699739	-2.08877	0.78329	1.044386	0.78329	4.177546	1.044386	3.394256	4.699739
0.8	1.804124	2.835052	-0.25773	2.57732	3.865979	1.546392	4.381443	3.608247	6.185567	6.701031
0.9	3.480278	0.232019	-2.08817	2.784223	5.568445	1.392111	1.392111	0.232019	2.320186	0
Average	**2.726679**	**4.256249**	**-1.19504**	**3.051032**	**5.940507**	**-4.68231**	**5.98964**	**6.095792**	**10.83388**	**9.757293**

Table 4: Average Recall and Precision Values for 59 Query by Applying GA's on Inner Product Similarity.

Recall	Dice	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	0.132	0.146	0.134	0.146	0.134	0.135	0.139	0.134	0.129	0.139	0.135
0.2	0.178	0.208	0.182	0.187	0.191	0.185	0.186	0.167	0.169	0.192	0.192
0,3	0.265	0.301	0.285	0.274	0.288	0.288	0.177	0.256	0.272	0.268	0.287
0.4	0.221	0.283	0.254	0.229	0.242	0.242	0.227	0.223	0.211	0.231	0.255
0.5	0.376	0.405	0.399	0.393	0.376	0.376	0.366	0.365	0.344	0.399	0.399
0.6	0.381	0.409	0.389	0.387	0.384	0.384	0.391	0.377	0.371	0.408	0.397
0.7	0.391	0.413	0.387	0.399	0.387	0.387	0.386	0.389	0.386	0.411	0.404
0.8	0.394	0.437	0.402	0.405	0.403	0.403	0.394	0.386	0.393	0.425	0.422
0.9	0.456	0.487	0.432	0.437	0.455	0.455	0.445	0.423	0.408	0.459	0.466
Average	**0.3104**	**0.3432**	**0.3182**	**0.3174**	**0.3177**	**0.3172**	**0.3012**	**0.3022**	**0.2981**	**0.3257**	**0.328556**

Table 5: GA's Improvement in Inner Product Similarity (GA's Improvement %).

Recall	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	10.60606	1.515152	10.60606	1.515152	2.272727	5.30303	1.515152	-2.27273	5.30303	2.272727
0.2	16.85393	2.247191	5.05618	7.303371	3.932584	4.494382	-6.17978	-5.05618	7.865169	7.865169
0,3	13.58491	7.54717	3.396226	8.679245	8.679245	-33.2075	-3.39623	2.641509	1.132075	8.301887
0.4	28.0543	14.93213	3.61991	9.502262	9.502262	2.714932	0.904977	-4.52489	4.524887	15.38462
0.5	7.712766	6.117021	4.521277	0	0	-2.65957	-2.92553	-8.51064	6.117021	6.117021
0.6	7.349081	2.099738	1.574803	0.787402	0.787402	2.624672	-1.04987	-2.62467	7.086614	4.199475
0.7	5.626598	-1.02302	2.046036	-1.02302	-1.02302	-1.27877	-0.51151	-1.27877	5.11509	3.324808
0.8	10.91371	2.030457	2.791878	2.284264	2.284264	0	-2.03046	-0.25381	7.86802	7.106599
0.9	6.798246	-5.26316	-4.16667	-0.2193	-0.2193	-2.41228	-7.23684	-10.5263	0.657895	2.192982
Average	**11.9444**	**3.355853**	**3.271745**	**3.203264**	**2.912908**	**-2.71346**	**-2.32334**	**-3.60072**	**5.074422**	**6.307254**

8. Comparison between the Best GA's Strategies

Table 6 shows the comparison between Dice (GA9) and Inner Product (GA1). From this table we notice that the Inner Product (GA1) is better than Dice (GA9) in all recall levels. Which means that Inner Product(GA1) that use one-point crossover operator and point mutation and use Inner Product similarity as a fitness function represent the best IR system in VSM to be used with the Arabic data collection.

9. Conclusions

For each similarity measure (DICE and Inner Product) in the VSM we compared ten different GA approaches, and by calculating the improvement of each approach over the traditional IR system, we noticed that most approaches (GA1, GA2, GA4, GA5, GA8, GA9 and GA10) give improvements compared to the traditional IR system, also we noticed that in the inner product the one-point crossover operator and point mutation gives the highest improvement over the traditional approach.

Table 6: Comparison Between the Best GA Strategies (Each Similarity Measures).

Recall	Dice(GA9)	Inner Product(GA1)
0.1	0.141	0.146
0.2	0.197	0.208
0,3	0.298	0.301
0.4	0.277	0.283
0.5	0.402	0.405
0.6	0.408	0.409
0.7	0.396	0.413
0.8	0.412	0.437
0.9	0.441	0.487
Average	0.330222	0.343222

References

[1] Tengku M.T., Sembok, C.J., and van Rijsbergen, "A simple logical-linguistic document retrieval system", Information Processing & Management, Volume 26, Issue 1, pp. 111-134, 1990.

[2] J. Carlberger, H. Dalianis, M. Hassel, O. Knutsson, "Improving Precision in Information Retrieval for Swedish using Stemming", In the Proceedings of NoDaLiDa-01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001, Uppsala, Sweden.

[3] Goldberg, D. E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.

[4] Hsinchun C., "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms", Journal of the American Society for Information Science. Volume 46 Issue 3, April 1995.

[5] D. Vrajitoru, "Crossover improvement for the genetic algorithm in information retrieval", Information Processing& Management, 34(4), pp. 405–415, 1998.

[6] Hananda E, "Evaluation of Different Information Retrieval models and Different indexing methods on Arabic Documents", Phd Thesis, ARAB Academy, 2008.

[7] Vicente P., Cristina P., "Order-Based Fitness Functions for Genetic Algorithms Applied to Relevance Feedback", Journal Of The American Society For Information Science And Technology, 54(2):152–160, 2003.

[8] Andrew T., "an Artificial Intelligence Approach to Information Retrieval", Information Processing and Management, 40(4):619-632, 2004.

[9] Rocio C., Carlos Lorenzetti, Ana M., Nelida B., "Genetic Algorithms for Topical Web Search: A Study of Different Mutation Rates", ACM Trans. Inter. Tech., 4(4):378–419, 2005.

[10] Mercy T., Naomie S., "A Framework for Genetic-Based Fusion of Similarity Measures In Chemical Compound Retrieval", International Symposium on Bio-Inspired Computing, Puteri Pan Pacific Hotel Johor Bahru, 5 - 7 September 2005.

[11] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, Osman A. Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems", proceedings of world academy of since, engineering and technology, volume 17, ISSN 1307-6884, 2006.

[12] Abdelmgeid A., "Applying Genetic Algorithm in Query Improvement Problem", International Journal "Information Technologies and Knowledge, Vol.1, p 309-316. 2007.

[13] Khoja, S., "APT:Arabic part-of-speech tagger", proceedings of the student workshop at second meeting of north American chapter of Association for Copmputational Linguistics (NAACL2001), Pittsburgh, Pennsylvania, pp. 20-26, 2001.

[14] yahaya, A., "on the Complexity of the initial stage of Arabic text processing", First Great Lakes Computer Science Conference, Kalamazoo, Michigan, USA, October, 1989.

[15] Goweder, A., De Roeck, A., "Assessment of a Significant Arabic Corpus", Arabic Natural Language Processing Workshop (ACL2001), Toulouse, France. Downloaded from: (http://www.elsnet.org/acl2001 arabic.html).
[16] I. Hmedi, and G. Kanaan and M. Evens, "design and implementation of automatic indexing for information retrieval with Arabic documents", Journal of American society for information science, Volume 48 Issue 10, pp. 867-881, 1997.
[17] Bassam Al-Shargabi, Islam Amro, and Ghassan Kanaan, "Exploit Genetic Algorithm to Enhance Arabic Information Retrieval", 3rd International Conference on Arabic Language Processing (CITALA'09), Rabat, Morocco, pp. 37-41, 2009.
[18] Fatemeh Dashti, and Solmaz Abdollahi Zad," Optimizing the data search results in web using Genetic Algorithm", international journal of advanced engineering and technologies, Vol 1, Issue No. 1, 016 – 022, ISSN: 2230-781, 2010.

First Author Dr. Eman Fares Al-mashakba is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Irbed University, Irbed, Jordan. She holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, Security, E-learning and image processing.

Second Author Dr. Feras Fares Al-mashakba is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Amman Arab University, Amman, Jordan. He holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, artificial intelligence, M-commerce.

Third Author Dr. Mohammad Othman Nassar is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Amman Arab University, Amman, Jordan. He worked as Assistant Professor at the Computer Information Systems department in the Arab Academy for Banking & Financial Sciences University. He holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, supply chain management, reengineering, outsourcing, and security. Dr. Nassar has published more than 12 articles in these fields in various journals and international conferences. He is included in the Panel of referees of "International Journal of Modeling and Optimization" and in the "International Journal of Computer Theory and Engineering", he was reviewer in the 2011 3rd International Conference on Machine Learning and Computing, also he is currently reviewer in A collection of open access journals called (academic journals).