AENSI Journals

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

An Enhanced Arabic Information Retrieval Using Genetic Algorithms: An Experimental Study and Results

[1]Bassam AL-Shargabi, [2]Omar Sabri, [3]Shadi Aljawarneh

[1]*Software Engineering Dept, Faculty of Information Technology, Al Isra University, P.O.Box 33,Amman 11622, Jordan.*
[2]*Management Information System Dept , Faculty of Administration Sciences Al Isra University P.O.Box 33, Amman 11622, Jordan.*
[3]*Software Engineering Dept, Faculty of Information Technology, Al Isra University P.O.Box 33,Amman 11622, Jordan.*

ARTICLE INFO

Article history:
Received 14 October 2013
Received in revised form 19 November 2013
Accepted 20 November 2013
Available online 15 December 2013

Keywords:
Genetic Algorithm, Arabic Information Retrieval, Precision, Vector Space Model

ABSTRACT

Many key challenges influence on the use of Arabic information retrieval systems, one of these is the performance of the Arabic information retrieval systems in terms of precision and recall. In this paper, we present the Genetic Algorithms to improve performance of Arabic information retrieval system based on vector space model. The main idea in this paper is the usage of an adaptive matching function, which obtained from a weighted combination of four similarity measures (Dot, Cosine, Jaccard, and Dice). The Genetic Algorithms used to optimize these matching functions, through obtaining the best achievable combination of these weights. The proposed genetic process is tested on Arabic documents collection and then results has shown a considerable improvement on the precision as performance measure.

To Cite This Article: Bassam AL-Shargabi, Omar Sabri, Shadi Aljawarneh, An Enhanced Arabic Information Retrieval Using Genetic Algorithms: An Experimental Study and Results. ***Aust. J. Basic & Appl. Sci.,*** *7(13):242-248, 2013*

INTRODUCTION

The main idea behind using Information Retrieval System (IRS) is to facilitate the retrieving of relevant documents based on users query. An information retrieval system consists of a software that help to user to find information as per their needs. The process of retrieving documents in an information retrieval system that is based on extracting the keywords from the text document of documents collection and assign weights for each keyword, a matching functions is used such Dot, Cosine, Jaccard and Dice in vector space model , to calculate matching score. Accordingly, The precision and recall measures are used to evaluate the effectiveness of the information retrieval system in meeting users information requirements.

Genetic Agorithms (GAs) is not new to information retrieval. GAs used for representing a posting as a chromosome and using genetic algorithms to choose a good indexes, and using GAs for user feedback to select weights for search terms in a query as introduced by Pathak *et al.* (2000) the GAs was suggested based on matching function adoption.

It is becoming obvious the that GAs is being used profusely in the information retrieval systems in general, the use of GAs depends on which filed of retrieval are being exploited or at which component of the information retrieval operations the GAs serves. The use of GAs used in field of image retrieval as presented by Kato (1998). Shokouhi *et al.*(2005) and Priya (2013) utilized the GAs in the WWW for improving crawling process or for improving the performance (i.e. Precision, Recall) as introduced by Abe *et al.*(199) or for improving the performance of the ranking functions as suggested by Fan *et al.*(2004) , Singh *et al.* (2012), and Ghwanmeh (2012). In the field of textual information retrieval, the GAs has contributed in one of the best automated key phrase generators for multi-objectives as presented by Jia-Long *et al.*(2004) or for domain specific key phrases by Qin *et al.*(2005) [10]. With focusing on the classic operations of the information retrieval systems that employs the classic models; GAs have been utilized as well as in the query optimization as introduced by Kushchu (2005), Nassar *et al.* (2011), and Anuradha *et al.* (2013), although, Pathak *et al.* (2000) used GAs based matching functions adaptation .

Unfortunately; the significant penalty of GAs usage is that the time consumption of resources exhausting since time and space variables are luxuries that the information retrieval systems do no not have.

In this paper, the genetic algorithms is adopted to optimize the performance information retrieval based on the vector space model, we used an adaptive matching function formed from the a weighted combination of the classic matching functions of the vector space model (dot, cosine, Jaccard and dice). The proposed genetic process are used to find the best achievable values of weights combines the classical matching function of the

Corresponding Author: Bassam AL-Shargabi, Software Engineering Dept, Faculty of Information Technology, Al Isra University, P.O.Box 33,Amman 11622, Jordan.

vector space model. The proposed genetic process are tested on Arabic documents collection and the experimental results proved to improve the overall performance of the system with range of 10% in general, the results are achieved with a lower number of populations found in as debated by Pathak *et al.* (2000),which means a lower delay time and a better response time.

The rest of the paper organized as follows, in Section two the vector space model and the matching function optimization attempts are presented. The proposed matching function in this paper and its combination with the genetic algorithms are introduced in Section 3. The approach that is followed in this paper to improve the precision is presented in Section 4. The proposed genetic process is introduced in Section 5. The implantation of proposed genetic process is presented in Section 6. Experimental result and discussion are presented in Section 7. Finally , conclusion is drawn in Section 8.

Vector Space Model and Matching Function Optimization Attempts:

The vector space model treats documents as vectors, a document is expressed in term of its inverted file that holds a stemmed version of the document terms with each term frequency tf multiplied by the weight of that term of the whole document collection. The weight is obtained from $Log_{10}(N / n_i)$, Where N is the total number of documents in collection, and n_i is the term frequency in document i . The next step is applying the well known similarity measures: Dot, Cosine, Jaccard and Dice (matching functions), followed by ranking the result against the score of the chosen measure. The large numbers of matching functions have been tried in literature as mentioned by Pathak *et al.* (2000), and no one say that a single matching function is the best. There are so many factors affects any information retrieval system , such as the size of the Document collection, the documents topic, and the nature of the user community, all these factors affects the performance of the matching function Pathak *et al.* (2000).

The precision and recall have been used to evaluate the effectiveness of the information retrieval system in meeting users information requirements. Recall defined as the fraction of relevant retrieved documents to the total number of relevant documents available in the document collection. Precision defined as the fraction of relevant retrieved documents to the total number of retrieved documents. Relevance feedback is typically used in IRS to improve document descriptions asin Pathak *et al.* (2000) , or queries as presented by William *et al.*(1987) with expectation that the overall performance of the IRS will be improved after such a feedback.

Several attempts were conducted to optimize the matching functions, Salton *et al.*(1990) argued the use of numerical methods to optimize the parameters of a matching function, but they choose to optimize only the parameters involved in standard inner product measure. This adaptation lead to the use of one of the following matching functions: inner product, cosine. In this paper, the matching function adaptation is not restricted to a particular form of the matching function.

William *et al.*(1987) assumed that the IR model have a criteria (like ordering of documents) that are differentiable in nature. This assumption leads them to use numerical methods, but numerical methods may not always be useful for a single matching function. In this paper, we used a weighted combination of the four matching functions of vector space model, which lead to achieve a higher precision value as seen in next section.

Genetic Modification of Matching Functions:

This section describes the proposed genetic modification of the matching functions, and the proposed experimental design to test genetic modification of matching functions . In this paper we use the vector space model as introduced by Bartell *et al.*(1992), as the basic model in IRS. In vector space model, the documents and queries are located in a multidimensional vector space as shown bellow in Figure 1. Retrieval is accomplished by searching for documents that are close to the query vector. Typically a single matching function is used to match document vector with the query vector.

In this paper, we treat the overall matching function as a weighted sum of the scores returned by different matching functions as shown bellow in Equation 1:

$$(d_i, q_i) = \sum (\omega_{ti} * Mf_i(d_i, q_i)) \quad (1)$$

Where i ranges from 1 to the total number of matching functions used; MF1(Dot), MF2 (Cosine), MF3 (Jaccard), and MF4 (Dic). The scores produced by individual matching functions; and the weights wt1, wt2, etc. are the weights associated with these scores. The (dj,q) signifies that this matching function is utilized to calculate scores for the document dj, (j varying from 1 to the total number of documents) for the given query 'q'. The weights wt1, wt2, etc. range from 0.0 to 1.0. A higher weight signifies that the associated matching function is more important than the one with lower weight. Thus a matching function with a weight of 0.6 is doubly as important as that with a weight of 0.3. GA's are robust in searching a multidimensional space to find optimal

solution, and this motivated us to use the GA to search for the optimal solution in order to improve precision of the retrieved documents. the fitness function as shown bellow in equation 2, is a performance measure or reward function, which evaluates how each solution is good or not. GA's naturally require a single valued measure to evaluate fitness of an individual in the population. a single point measure which combines precision and recall measures as shown in Equation 2 as suggested by Pathak *et al*. (2000)

$$E = 1 - \frac{1}{\left[\frac{\alpha}{P} + \frac{(1-\alpha)}{R}\right]} \quad (2)$$

Where $\propto$ is a parameter to express the degree of users preference for precision (P) or recall (R).The higher value of $\propto$ characterizes that the user have less preference for recall, while a lower value of $\propto$ characterizes that the user have less preference for precision.

In this paper we use (1-E) as our fitness function so that higher values of our fitness function correspond with better performance for Arabic information retrieval system.

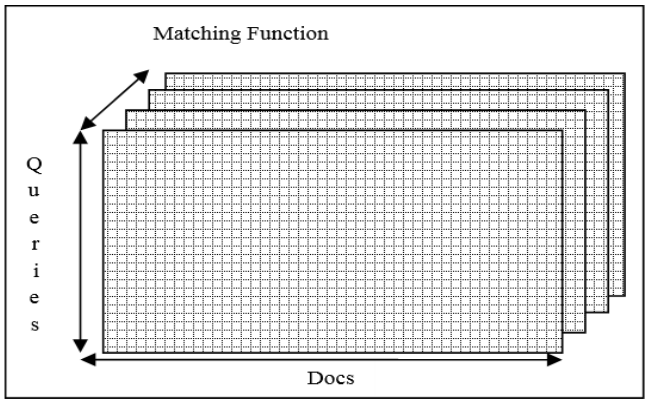

Fig. 1: Multidimensional Space represents each query with its relevant documents according to a certain matching function

Improving Precision:

In IRS, the precision measure is used to assess the accuracy of retrieved documents. To validate the proposed genetic process in this paper, we have used five inputs precision compared to one output precision. The five input precisions are the precisions of the four matching functions of vector space model (Dot, Cosine, Jaccard and Dice). The fifth precision of obtained from the retrieved documents in Equation 1. The randomly generated weights for the 5 matching functions are transferred to form the genotype of the genetic algorithm, the genetic algorithm then is used to find the best achievable weights, and then tested through the fitness function in Equation 2, during the genetic epoxies, precision and recall are being computed according the value of that function, we determined whether if this is a good combination or not, genetic flow guarantees that the accepted values are rewarded and the unaccepted values are panelized.

Once the genetic operation reaches its end the value of the highest fitness chromosome then is returned, pretension that this is the best achievable combination of weights. Then precision value is computed for this combination and fixed and comparing the value of this precision to the other five precisions as indicator to the usefulness of the proposed genetic process.

The Genetic Process:

The proposed genetic process that we followed to implement GA to enhance the Arabic information retrieval as shown in Figure 2. The genetic process implemented as follows:

Matching function variants:

Generating and assigning a random weight for each individual matching function (in the range 0.0 to 1.0). The overall matching function is a weighted combination of the individual function scores (individual matching function scores are normalized to be in the range of 0 to 1). Weights are encoded using the actual real numbers between 0.0 and 1.0. The initial population consisted of 30 (population size) such randomly chosen individuals.

The population consists of 30 chromosomes, each chromosome consists of 4 genes, each gene allele contains the binary representation of the weight values, we used 8 bit to express each allele, this has made each chromosome of 32 bit value under the disposal of genetic evolver.

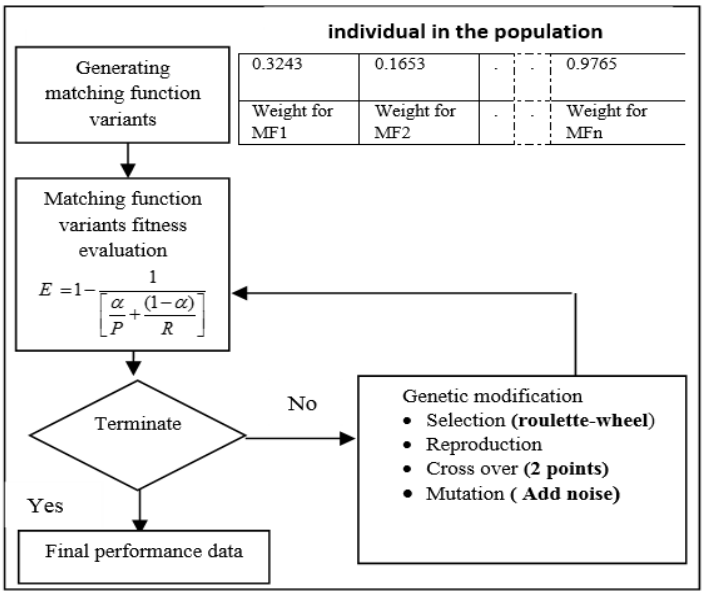

Fig. 2: Genetic Process.

***Matching function variants fitness evaluation*:**

For each individual in the population an overall matching score is calculated for each document and the documents in the document collection are arranged in a decreasing order. Based on the parameter for document cut-off value (DCV) the top DCV number of documents are retrieved. Based on the relevance judgments for this set of documents, precision and recall are calculated. These values are used to calculate fitness of the individual.

***Genetic Modification*:**

Genetic operators are applied to the individuals in the previous generation to generate the next generation of individuals, which it involves four stages as follows:

1.Selection and reproduction: All individuals in the previous generation were made available for reproduction in the next generation. The roulette –wheel reproduction process was used to select individuals for reproduction, because we are interested in having a sorted population against fitness.

2. *Crossover*: A two-point crossover was followed (exchanging information between two randomly selected points on the individual string). A parameter 'cross-over rate' determined the number of individuals that actually mate.

3. *Mutation*: Mutation was accomplished by introducing Gaussian noise.

4. *Process termination*: The process of genetic modification was terminated after a presetnumber of generations (80) as compared generations (60) used by Al-Shargabi *et al*.(2009) [18] .

Implementation:

An Arabic document collection are used , which consists of 200 documents and 50 queries. First, the vectors space variants were constructed with a java program, all of the related operations were carried out there. The inverted file generation and the similarity measures were calculated for each query against the relevant documents, for each query the relevant documents were retrieved against every similarity measure. The weighted function in Equation (1) is being calculated using the values of the four similarity measures. Therefore, the precision is calculated before the starting of the genetic process. The genetic process was implanted by Java Genetic Algorithms Package (JGAP), the variants generation is being explained before. At the end of genetic evolution, the chromosome with the highest fitness is accredited, the chromosome then is decomposed to its formulating genes, each gene allele is converted back to from an 8 bits binary string into a real number between 0 and 1. The precision calculated accordingly for this chromosome using Equation (1), forming the final result of the approach.

RESULTS AND DISCUSSION

As seen in Table 1, the first column represents queries sample , the second column represents PMF1 which is the precision of the first matching function, the first matching function was the dot measure. The PMF2 is the precision second matching function which is the Cosine measure, the PMF3 is the precision of the third matching function which is the Jaccard measure, PMF3 is the precision of the fourth matching function which is the Dice measure. The columns form the second to the fourth represents the precision of the query sample using that matching function.

Table 1: Holds a sample of 20 queries from the used 50 queries.

Matching functions / Queries	PMF1	PMF2	PMF3	PMF4	WF	WFGM
جامعه القاهرة	0.656	0.521	0.5233	0.6498	0.8345	0.9333
الجامعة الأردنية	0.6573	0.6103	0.5404	0.6065	0.7473	0.8372
الجمعية العلمية الملكية	0.6168	0.7072	0.7268	0.5413	0.8644	0.9222
الأردن	0.5467	0.6572	0.6356	0.6664	0.7677	0.8765
كلية الهندسة	0.5773	0.5704	0.6471	0.5302	0.737	0.827
الحرف العربي	0.5345	0.6232	0.7278	0.7267	0.8072	0.8987
قاموس للنصوص العربية	0.5526	0.5734	0.7703	0.5455	0.8415	0.9305
الكلمات العربية	0.6996	0.7919	0.5327	0.5508	0.8819	0.9719
اللغة العربية	0.5348	0.6094	0.5449	0.5898	0.6994	0.7894
المملكة العربية السعودية	0.5426	0.5654	0.7605	0.5464	0.8505	0.9405
معالجة اللغات الطبيعية	0.6261	0.7376	0.5356	0.7456	0.8259	0.9101
انظمة الحاسبات الالية	0.6017	0.6852	0.6137	0.5316	0.7751	0.8651

The precision of the weighted matching function calculated using Equation (1) is found on the fifth columns weighted function (WF). The WF precision value represents the precision of the weighted function before using the genetic optimization. As seen from Table 1, the precision of WF achieves a better precision from any individual matching function by its own, this is due to the cooperation between the values of the fitness function, each fitness function of the classical approaches participates in increasing the precision, but our proposed genetic process as explained earlier and as illustrated before in figure 3 achieved higher precision, which is represented by WFGM in Table 1.

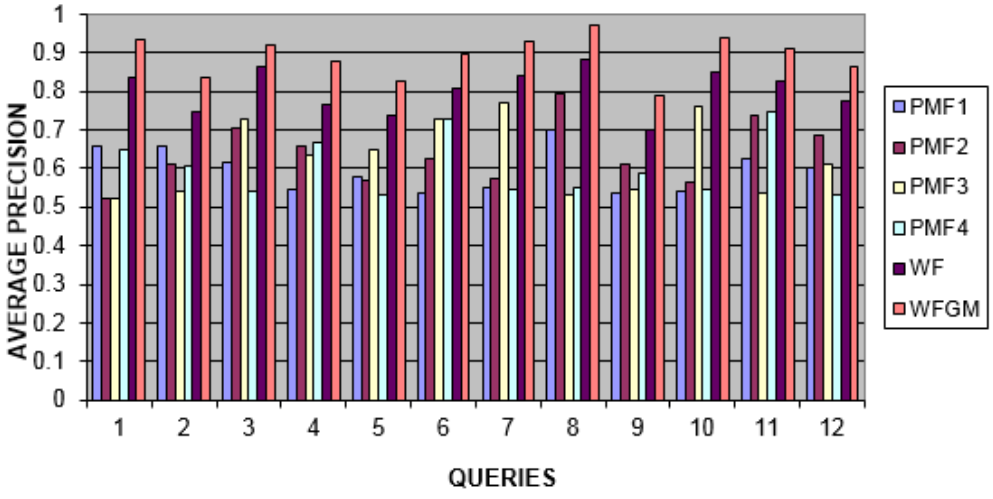

Fig. 3: Matching Function with Genetic Algorithm Adaptation.

The WFGM represents the modified weighted function with the genetic operation, the precision after the genetic evolutions are developed and the best achievable combination of weights is being converged as

illustrated in Figure 4 and 3. As noted the overall performance enhancement in the retrieval operation was achieved by the proposed genetic operation.

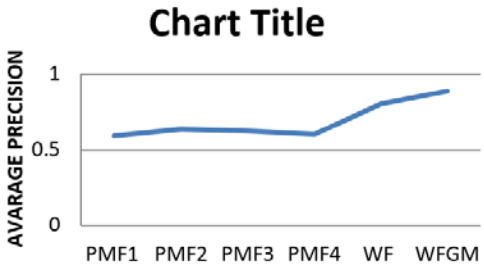

Fig. 4: Average precision enhancement by Genetic algorithm.

Conclusion:

The accuracy of the retrieved documents in information retrieval systems is a significant factor, in this paper we used the genetic algorithms to optimize the weighted function that's combines the four similarity measures of the vector space model (Dot, Cosine, Jaccard and Dice). The proposed genetic process was tested using 200 Arabic documents in document collection along with 50 queries, the results shown a better performance in terms of precision for the document collection thorough the optimization of the weighted function.

ACKNOWLEDGMENTS

I would like to thank everyone who contributed to the completion of this work.

REFERENCES

Abe, K., T. Taketa, H. Nunokawa, 1999. An Efficient Information Retrieval Method In WWW Using Genetic Algorithms. In proceeding of the International Conference On Parallel Processing Workshops (ICPPW'99).

Al-Shargabi, B., I. Amro, G. Kanaan, 2009. Exploit Genetic Algorithm to Enhance Arabic Information Retrieval. In proceeding of the 3rd International Conference on Arabic Language Processing (CITALA'09), 37-41.

Anuradha, T., K. Pallavi, C. Prajakta, K. Manisha, 2013. Introducing GA Based Information Retrieval System For Effectively Retrieving News Article. International Journal of Engineering Research & Technology (IJERT), 2(4): 1088-1090.

Bartell, B.T., G.W. Cottrell, R.K. Belew, 1992. Latent Semantic Indexing Is An Optimal Special Case Of Multidimensional Scaling. In Belkin, N., Editor, Proc. 15th Annual Intl. ACM SIGIR Conf.

Fan, W., M.D. Gordon, P. Pathak, W. Xi, E.A. Fox, 2004. Ranking Function Optimization For Effective Web Search By Genetic Programming: An Empirical Study. In Proceedings Of The 37th Annual Hawaii International Conference On System Sciences (HICSS'04) - Track 4.

Ghwanmeh, S., 2012. Enhanced Search Scheme Precision and Performance using a GA Approach with Application to Arabic Content. Journal of Advanced Computing, 1: 1-8.

Jia-Long, W., A.L. Agogino, 2004. Automating Keyphrase Extraction With Multi-Objective Genetic Algorithms Found In: Proceedings Of The 37th Annual Hawaii International Conference On System Sciences (HICSS'04)- Track4.

Kato, S., 1998. An Image Retrieval Method Based On A Genetic Algorithm. In proceeding of the 13th International Conference On Information Networking (ICOIN'98).

Kushchu, I., 2005. Web-Based Evolutionary And Adaptive Information Retrieval, Graduate Sch. Of Int. Manage., Int. Univ. Of Japan, Niigata, Japan; This Paper Appears In: Evolutionary Computation, IEEE Transactions On Publication Date: April.

Nassar, M.O., F. Al-Mashagba, E. Al-Mashagba, 2011. Improving the User Query for the Boolean Model Using Genetic Algorithms. International Journal of Computer Science, 8(1): 1694-0814.

Pathak, P., M. Gordon, W. Fan, 2000. Effective Information Retrieval Using Genetic Algorithms Based Matching Functions Adaptation. In Proceeding of the 33rd Hawaii International Conference On System Sciences.

Priya, I.B., L.P. Patil, 2013. Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization. International Journal of Future Computer and Communication, 2(6): 595-599.

Qin, J., H. Chen, 2005. Using Genetic Algorithm In Building Domain-Specific Collections: An Experiment In The Nanotechnology Domain. In Proceedings Of The 38th Annual Hawaii International Conference On System Sciences (HICSS'05) - Track 4.

Salton, G., C. Buckley, 1990. Improving Retrieval Performance By Relevance Feedback", Journal Of The American Society For Information Science, 41(4): 288-297.

Shokouhi, M., P. Chuba, Z. Raeesy, 2005. Enhancing Focused Crawling With Genetic Algorithms. In Proceeding of International Conference On Information Technology,Coding And Computing (ITCC'05) - Volume II.

Singh, J., J. Godara, 2012. Information Retrieval using Page Relevancy. International journal of Computer Science & Communication, 3(2): 29-31.

William, P., G. Jones, W. Furnas, 1987. Pictures Of Relevance: A Geometric Analysis Of Similarity Measures", Journal of The American Society of Information Science, 38(6): 420- 442.