

Exploit Genetic Algorithm to Enhance Arabic Information Retrieval

Bassam Al-Shargabi*, Islam Amro** and Ghassan Kanaan***

*Arab academy for Banking and Financial Science, Department of computer Information System. Jordan-Amman, Bassam20_152@yahoo.com

**Arab academy for Banking and Financial Science, Department of computer Information System. Jordan-Amman, iamro@qou.edu

***Arab academy for Banking and Financial Science, Department of computer Information System. Jordan-Amman, ghkanaan@aabfs.org

Abstract—In this paper, we explain the use of Genetic Algorithms to improve performance of Arabic information retrieval system, which based on vector space model. The performance was enhanced through 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 was used to optimize this matching functions. Optimization has been made through getting the best achievable combination of these weights. We used the precision as performance measure. The result of optimization process has shown a considerable improvement on the performance. The document collection was used is an Arabic documents collection.

Keywords—Cosine similarity, Jaccard similarity, Dice similarity, Fitness function, Genetic Algorithm, Information Retrieval.

I. INTRODUCTION

Genetic Algorithm (GA) is a probabilistic optimizing algorithm simulating the mechanism of natural selection of living organisms and is often used to solve problems having expensive solutions. In GA, the search space is composed of candidate solutions to the problem; each represented by a string is termed as a chromosome. Each chromosome has an objective function value, called fitness. A set of chromosomes together with their associated fitness is called the population. This population, at a given iteration of the genetic algorithm, is called a generation. Holland, De Jong and Goldberg were pioneered of GA in the context of continuous non-linear optimization [1, 2, 3].

Genetic algorithms (GAs) are not new to information retrieval. GA used for representing a posting as a chromosome and using genetic algorithms to select good indexes, and using GAs with user feedback to choose weights for search terms in a query, and Pathak suggested using GA based on matching function adoption.[4]

It is becoming obvious that GAs is being used intensively in the information retrieval systems in general, that depends on which field of retrieval GAs are being exploited or at which component of the information retrieval operations the GAs serves, in the field of image retrieval GAs based retrieval are one of

the best known approaches[16]. The WWW has also utilized the GAs for improving crawling process [17] or for improving the performance (i.e. Precision, Recall) [18] or for improving the performance of the ranking functions [19]. And for the textual information retrieval system, the GAs has contributed in one of the best automated key phrase generators for multi-objectives [22] or for domain specific key phrases [23].

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 [20] and matching functions adaptation [21]. Unfortunately; the significant penalty of the usage of GAs that is the time consumption and resources exhausting since time and space variables are luxuries the information retrieval systems do not have.

Sections two discusses the vector space model and matching function optimization Attempts, and the openness in the matching functions for discussion, there is no definite matching function for the vector space model, this is an introduction to our approach through understanding the needs for further accurate matching functions for the vector space model .

The next section explains the matching function suggested and its combination with the genetic algorithms this is where the approach details are being introduced theoretically and how the structure of the tool was built is configured. The following section explains the performance measure for the approach we are suggesting and how to tell that the approach has achieved a better desirable result. In the following section we explain the implantations used for the tool and how is it built and configured then we go to the results and discussions sections.

In this research, we exploited the genetic algorithms to optimize the vector space model of information retrieval, 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 genetic algorithms were used to find the best achievable values of weights combines these classic function, we were able to improve the overall performance of the system with range of 10% in general. This result was achieved with a lower number of populations found in [21] which means a lower delay time and a better response time, we have also used an Arabic document collection.

II. VECTOR SPACE MODEL AND MATCHING FUNCTION OPTIMIZATION ATTEMPTS.

As known; the vector space model of the information retrieval system treats documents as vectors, a document is expressed in term of its inverted file the holds a stemmed version of the document terms with each term frequency f_i multiplied by the weight of the term in the whole document collection, this weight is obtained from $\text{Log}_{10}(N / n_i)$ Where N is the total number of documents in collection and n_i is the term frequency in document i , then we apply any of the similarity measures known (dot, cosine, jaccard and dice) and we rank the result against the score of the chosen measure. also a large number of matching function have been tried in literature [5], and no one say single matching function is the best, there many factors on the retrieval environment such as the size of the Document collection., the document topic, and the nature of the user community affects which matching function will perform better [5]. Harman [1986] showed that by switching between different normalized inner product measures as matching functions it is possible to get a 12% improvement in average precision.

There are different criteria like precision and recall have been used to evaluate the effectiveness of the system in meeting users' information requirements. Recall is the fraction of relevant retrieved documents to the total number of relevant documents available in the document collection. Precision is defined as the fraction of relevant retrieved documents to the total number of retrieved documents. Relevance feedback is typically used by the system to improve document descriptions [4], or queries [13] with expectation that the overall performance of the system will improve after such a feedback.

There were several attempts to optimize the matching functions, Bartell, Cottrell, and Belew in [24] have used numerical methods to optimize the parameters of a matching function. But they have chosen to optimize only the parameters involved in a standard inner product measure. their adaptation leads to the use of one of the following matching functions: inner product, cosine, but our research looks at adaptation of various different forms of matching functions and is not restricted to a particular form of the matching function. Bartell et al. [24] have assumed that the IR model have 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. This leads to the use of the weighted combination of the known matching function and optimize this combination in order to achieve a higher precision value as seen in next section.

III. GENETIC MODIFICATION OF MATCHING FUNCTIONS

This section describes how to use GAs to modify the matching functions used and the experimental design to test our algorithm. We used the vector space model [15]

as the basic model in this research. Vector space model, documents and queries are located in a multi-dimensional vector space as seen in figure. Retrieval is accomplished by searching for documents that are close to the query vector. Typically a single such matching function is used to match document vector with the query vector.

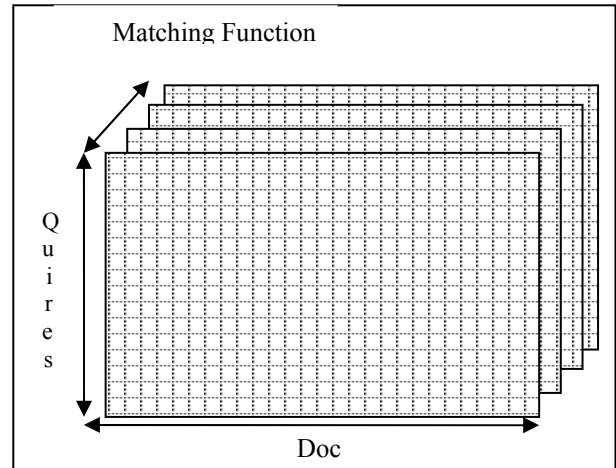


Figure 1. Multidimensional Space represents each query with its relevant documents according to a certain matching function.

In our research we treat an overall matching function as a weighted sum of the scores returned by different matching functions. Thus, overall matching function

$$(d_i, q_i) = \sum (\omega_i * Mf_i(d_i, q_i)) \dots\dots\dots(1)$$

Where i ranges from 1 to the total number of matching functions used; MF1, MF2 etc. are the scores produced by individual matching functions; and 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 that which is associated 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 in this research to search for the optimal solution. Fitness function is a performance measure or reward function, which evaluates how each solution, is good. GA's typically require a single valued measure to evaluate fitness of an individual in the population. a single point measure which combines precision and recall measures. It is:

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

Where α is a parameter to express the degree of users preference for precision (P) or recall (R). Which higher value of α characterizes a user with less preference for

recall, while a lower value of α characterizes one with a less preference for precision.

We decided to use $(1-E)$ as our fitness function so that higher values of our fitness function correspond with better performance.

IV. IMPROVING PRECISION

Precision improvement of retrieved documents is a reliable measure for verifying the acceptance of this approach, we have used five input precision compared to one output precision, the five input precisions are the precisions of the four classic vector space model (Dot, Cosine, Jaccard and Dice), the fifth precision of obtained from the retrieved documents in equation (1), the randomly generated weights then are transferred to form the genotype of the genetic algorithm, the genetic algorithm then is used to find the best achievable weights, this is tested through the fitness function of equation (2), during the genetic epxies, precision and recall is being calculated according the value of that function we determine if this is a good combination or not, genetic flow guarantees that the accepted values are rewarded and the unaccepted values are panelized.

After the genetic operation reaches its end the value of the highest fitness chromosome then is returned, claiming that this is the best achievable combination of weights. Then precision value is calculated for this combination and fixed, comparing the value of this precision to the other five precisions will indicated the usefulness of this approach.

V. THE GENETIC PROCESS

The following process was followed to implement GA Generate.

Matching function variants: For each individual matching function we assigned a randomly chosen weight (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. Figure 2 explains the flow of the work.

Matching function variants fitness evaluation: For each individual in the population an Overall matching score is calculated for each document and documents in the collection are arranged in the decreasing order of this score. 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. It involves four stages

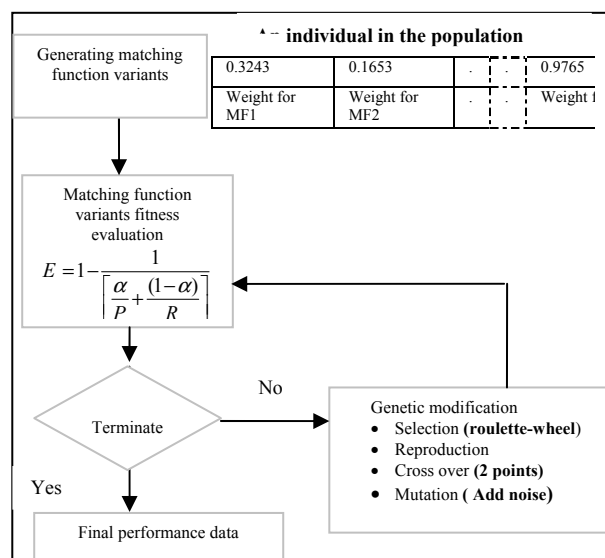


Figure 2. Genetic Operation

.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, this is 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 preset number of generations (60).

VI. IMPLEMENTATION

The documents which used is an Arabic document collection, 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, then the weighted function in equation (1) is being calculated using the values calculated in the similarity measures, then the precision is calculated before the starting of the genetic operation. The genetic operation (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.

VII. RESULTS AND DISCUSSION

Table1. Holds a sample of 20 queries from the used 50 queries.

Query	PMF1	PMF2	PMF3	PMF4	WF	WFGM
استخدام الحاسب الآلي	0.7425	0.5754	0.5532	0.6298	0.8325	0.9225
التعليم بمساعدة الحاسب	0.6573	0.6103	0.5404	0.6065	0.7473	0.8372
الحاسب الآلي	0.6231	0.6206	0.7639	0.5413	0.8539	0.9439
الحاسبات الصغيرة	0.7860	0.6598	0.5414	0.5981	0.8760	0.9660
الحاسبات المتناهية الصغر	0.5773	0.5704	0.6471	0.5302	0.7370	0.8270
الحرف العربي	0.5293	0.5323	0.5207	0.5397	0.6223	0.7123
الخطة الوطنية للمعلومات	0.6841	0.5814	0.6495	0.6193	0.7741	0.8641
الخليج العربي	0.6834	0.6145	0.6763	0.7507	0.7734	0.8634
الذكاء الاصطناعي	0.5467	0.6698	0.6257	0.6794	0.7598	0.8498
الذكاء الآلي	0.5445	0.6168	0.7072	0.7268	0.7972	0.8872
العالم العربي	0.6508	0.6795	0.6603	0.6377	0.7695	0.8595
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
النص العربي	0.6191	0.7288	0.5278	0.7319	0.8188	0.9088
انظمة الحاسبات الالوية	0.6017	0.6852	0.6137	0.5316	0.7751	0.8651
برامج الحاسب الآلي	0.6111	0.6995	0.6151	0.5299	0.7895	0.8795
برمجة الحاسبات الالوية	0.6498	0.7258	0.5852	0.6902	0.8158	0.9058
تدريس مواد الحاسب	0.6385	0.5810	0.6177	0.5644	0.7285	0.8185
تركيب الجملة العربية	0.7769	0.6109	0.6543	0.7383	0.8669	0.9569

The first column represents the sample query, the second column represents PMF1 which is the precision of the first matching function, the first matching function was the dot measure, the MF2 is the second matching function which is the Cosine, the Third is the Jaccard and the fourth is the Dice function, the columns form the first to the fourth represents the precision of the sample query using that matching function. The weighted matching function calculated by equation (1) precision is found on the fifth columns (WF), this is the value of the precision of the weighted function before using the genetic optimization, note that the precision of WF achieves a better precision from any individual function by its own, this is due to the cooperation between the values of the fitness function, each fitness function of the classic approaches participates in increasing the precision so there combination will lead the a better precision as illustrated in figure 3 .

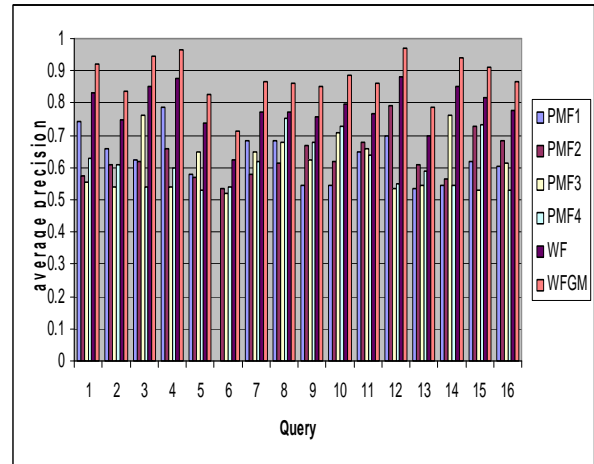


Figure 3. Matching Function with Genetic Algorithm Adaptation

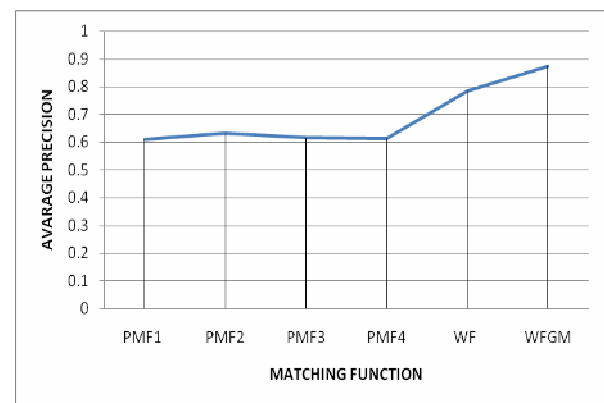


Figure 4. Average precision enhancement by Genetic algorithm

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. We note the overall performance enhancement in the retrieval operation achieved by the genetic algorithm.

VIII. CONCLUSION

Since the first introductions of information retrieval systems in the late 70's; performance enhancement has been a very attractive trend for researchers, as long as we have new methodologies and approaches we will work on improving the performance of the suggested methodologies.

The genetic algorithms are one of the widely used approaches in improving the performance of the non-linear systems where the Arabic information retrieval systems is one of them, we used the genetic algorithms to optimize a weighted function combines of the four similarity measures of the vector space model (Dot, Cosine, Jaccard and Dice), the genetic algorithm was used to seek for the best achievable combination of weights and function, we used 200 Arabic documents in our collection along with 50 queries.

The usage of the genetic algorithms has shown a better performance in precision for the document collection

through the optimization of the weighted function suggested.

REFERENCES

- [1] J. H. Holland, "Adaptation In Natural And Artificial Systems", University Of Michigan Press, Ann Arbor, 1975.
- [2] K. A. Dejong, "An Analysis Of The Behavior Of A Class Of Genetic Adaptive Systems", Ph.D. Thesis, University Of Michigan, 1975.
- [3] D. E. Goldberg, "Genetic Algorithms In Search, Optimization, And Machine Learning", Addison-Wesley, Reading, MA., 1989.
- [4] P. Pathak, M.Gordon and W.Fan, "Effective Information Retrieval Using Genetic Algorithms Based Matching Functions Adaptation", Proceeding Of The 33rd Hawaii International Conference On System Sciences 2000 .
- [5] William P. Jones, George W. Furnas, "Pictures Of Relevance: A Geometric Analysis Of Similarity Measures", Journal Of The American Society Of Information Science 38(6): 420-442,1987
- [6] Vicente P. Guerrero Bote ,F_Elix De Moya Aneg, " A Test Of Genetic Algorithms In Relevance Feedback", 2001 Published By Elsevier Science Ltd.
- [7] Hsinchun Chen," Machine Learning For Information Retrieval: Neural Networks, Symbolic Learning, And Genetic Algorithms", Journal Of The American Society Of Information Science. 46(3):194-216, 1995
- [8] Richard K. Belew," Adaptive Information Retrieval: Using A Connectionist Representation To Retrieve And Learn About Documents", AIR Stands For Adaptive Information Retrieval 1989 ACM,
- [9] Ellen M. Voorhees," Evaluation By Highly Relevant Documents", *SIGIR'01*, September 9-12, 2001, New Orleans, Louisiana, USA. ACM
- [10] NIR .OREN, "Reexamining *Tf.Idf* Based Information Retrieval With Genetic Programming", Proceedings Of SAICSIT 2002, Pages 224–234.
- [11] C. Carpinto, G. Romano and V.Gianninni," Improving Retrieval Feedback With", ACM Transactions On Information Systems, Vol. 20, No. 3, July 2002, Pages 259–290.
- [12] C.David, A.BLAIR and E.Maron," An Evaluation of Retrieval Effectiveness for Full-Text Document Retrieval Systems", *Communications Of The ACM*, March 1985 Volume 28.
- [13] Gerard Salton; Chris Buckley," Improving Retrieval Performance By Relevance Feedback", *Journal Of The American Society For Information Science (1986-1998)*; Jun 1990; 41, 4; ABI/INFORM Global Pg. 288
- [14] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, And Osman A. Sadek," Using Genetic Algorithm To Improve Information Retrieval Systems", Transactions on Engineering, Computing and Methodology, Volume 17 December 2006 ISSN 1305-5313
- [15] Salton, G. *The SMART Retrieval System - Experiments In Automatic Document Processing*. Perntice-Hall, Inc., Englewood Cliffs, NJ, (1971).
- [16] S. Kato, An Image Retrieval Method Based On A Genetic Algorithm, 13th International Conference On Information Networking (ICOIN'98) , January 1998
- [17] M. Shokouhi, P. Chuba and Z. Raeesy , Enhancing Focused Crawling With Genetic Algorithms, International Conference On Information Technology: Coding And Computing (ITCC'05) - Volume II , April 2005
- [18] K. Abe, T. Taketa and H. Nunokawa , An Efficient Information Retrieval Method In WWW Using Genetic Algorithms, 1999 International Conference On Parallel Processing Workshops (ICPPW'99), September 1999
- [19] W. Fan, Michael D. Gordon, P. Pathak, W. Xi and Edward A. Fox, Ranking Function Optimization For Effective Web Search By Genetic Programming: An Empirical Study, Proceedings Of The 37th Annual Hawaii International Conference On System Sciences (HICSS'04) - Track 4 , January 2004
- [20] Kushchu, I., 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 2005
- [21] Praveen Pathak, Michael Gordon, Weiguo Fan , Effective Information Retrieval Using Genetic Algorithms Based Matching Functions Adaptation, 33rd Hawaii International Conference On System Sciences-Volume 2 , January 2000
- [22] Jia-Long Wu, Alice M. Agogino , Automating Keyphrase Extraction With Multi-Objective Genetic Algorithms Found In: Proceedings Of The 37th Annual Hawaii International Conference On System Sciences (HICSS'04)-Track4,January2004
- [23] Jialun Qin, Hsinchun Chen , Using Genetic Algorithm In Building Domain-Specific Collections: An Experiment In The Nanotechnology Domain, Proceedings Of The 38th Annual Hawaii International Conference On System Sciences (HICSS'05) - Track 4 , January 2005
- [24] Bartell, B. T., Cottrell, G. W., And Belew, R. K. (1992). Latent Semantic Indexing Is An Optimal Special Case Of Multidimensional Scaling. In Belkin, N., Editor, Proc. 15th Annual Intl. ACM SIGIR Conf.