Improving the User Query for the Boolean Model Using Genetic Algorithms

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

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

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

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

Abstract

The Use of genetic algorithms in the Information retrieval (IR) area, especially in optimizing a user query in Arabic data collections is presented in this paper. Very little research has been carried out on Arabic text collections. Boolean model have been used in this research. To optimize the query using GA we used different fitness functions, different mutation strategies to find which is the best strategy and fitness function that can be used with Boolean model when the data collection is the Arabic language. Our results show that the best GA strategy for the Boolean model is the GA (M2, Precision) method.

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

1. Introduction

The resource discovery problem is concerned with how to find information interest among the vast and growing amount of information available, this resource discovery problem is one of the most pressing issues with the explosive growth of the Internet [7]. Information retrieval (IR) defined as the study of how to determine and retrieve from collection of stored documents the parts that are relevant to particular information needs [1]. The major information retrieval models are the vector space model, Boolean model, Fuzzy model and the probabilistic model. These models are used to find the similarity between the query and the documents in order to retrieve the documents that reflect the query. The similarity then used to evaluate the effectiveness of IR system using two measures: Precision [8], and Recall [2].

The problem with the IR models is that it may converge to a result that is only locally optimal, which means it may lead to form a query that is better than the original form but significantly poorer than another undetected form, so Genetic Algorithm (GA) can be used to solve this problem by finding the global solution for the given problem. The genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution [3].

In this paper, we will work on Boolean IR model to optimize the user query using different genetic algorithms settings (different mutation techniques, different fitness functions). As a test bed; we are going to use an Arabic data collection which was presented for the first time by [24]; this data set is composed from 242 documents and 59 queries.

syntactical, morphological, and semantic differences are found in Arabic language compared to other Indo-European languages [13]. Arabic language text is more sparsed than that of English, which means that for the same text length English words are repeated more often than Arabic words [14, 15]. Sparseness may negatively affect the retrieval quality in Arabic language because Arabic terms will get less weight compared to English. In written Arabic, most letters take many forms of writing. Also, there is a punctuation associated with some letters that may change the meaning of two identical words. Finally; comparing to English roots, Arabic roots are more complex. The same Arabic root, depending on the context, may be derived from multiple Arabic words.

Finally, we can say that the lack of similar studies in the literature based on Arabic data collections was our motivator to conduct a deep and rich comparative study that apply different Genetic algorithm (GA) strategies using different mutation techniques and different fitness functions on the output of traditional IR system based on Boolean model in order to improve the user query.

2. Previous Studies

Query optimization is an active research area in IR, many studies have been conducted in this area based on English

data collections [4,8,9,10,11,12,16,17,18,19,20,21,22,23]. Vaclav S, Dusan H [4] deals with Genetic algorithms to optimize the Boolean query in information retrieval system based on English data collection, in this study the authors used three different mutation criteria, they found that GA improves the performance compared to traditional approach, and the improvement is different when using different mutation criteria. Masaharu et al. [8] employed a few number of query terms and concept categories with Boolean expressions; they use only the words that exist in the original query for reformulating the Boolean query. Morgan and Kilgour employ GAs to choose search terms from a thesaurus and dictionary [12]. Unlike [8, 12]; in our study we used terms not only from the original query; but also from the retrieved documents. The authors in [9, 10, 11] examine GAs for information retrieval and they suggested new crossover and mutation operators, all of them used English data collections.

Other contributions towards evolutionary optimization of search queries were introduced by Kraft et al. [18]; they used genetic programming to optimize Boolean search queries only, and their work was based on English data collection. Cordn et al. in [19] introduced what he called MOGA-P, it is an algorithm to deal with search query optimization as a multi-objective optimization problem, the authors conduct a comparison between their approach and with several other methods including Kraft's. Yoshioka and Haraguchi [20] introduced query reformulation interface to transform Boolean search queries into more efficient search expressions. Finally the researchers in [23] investigate evolutionary algorithms as a tool for the optimization of user queries and seek for its good settings.

Using GA to improve the performance of Arabic information system is rare in the literature when compared to other languages. In [16, 17] the researchers used Genetic Algorithms to improve performance of Arabic information retrieval system, the system was based on vector space model. In this research we are going to use the Boolean model.

3. Boolean model

Retrieval systems based on Boolean logic have long served as the cornerstone of the commercial document retrieval system market and remain very important because of the relative simplicity of the query language and the ease with which it can be understood and implemented [5]. The most common use for a Boolean expression is to state what characteristics must be present in material to be retrieved in a system that retrieves and presents to users bibliographic records or full-text. A second use of Boolean expressions, likely to increase in importance over the next decade, is in rules incorporated into document and email filtering systems. Boolean expressions typically use three operators: AND, OR, and NOT.

4. Genetic Algorithms (GA)

GAs exploits the idea of the survival of the fittest and an interbreeding population to create an innovative search strategy. The GA algorithm flowchart is illustrated in Figure 1. Figure 1 shows the Genetic algorithm operations such as Reproduction which can be implemented to choose the fittest individuals using something called the fitness function. The second operation found in figure 1 is the crossover operation, the cross over operation details can be found in [3]. The last operation in GA is called mutation which include the alteration of genes randomly for a particular chromosome.

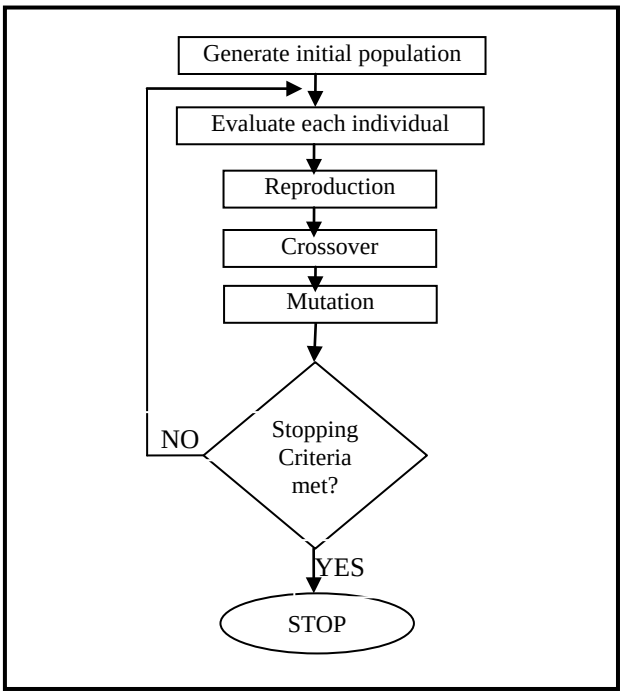

Fig. 1 Flowchart for Typical Genetic Algorithm.

5. Experiment

In this study we used IR system based on Boolean model that was built and implemented by Hanandeh [6] to deal with the 242 Arabic documents from the Proceedings of the Saudi Arabian National Conference [24]. This study

was conducted in a way that is similar to [16] with differences related to the mutation strategies as following:

1) We select the highest 15 terms frequency from the best 10 documents retrieved using the original query used which was used by Hanandeh syetem [6].
2) Based on the 15 terms; we construct 10 Queries.
3) Each query from the previous 10 queries was represented as a tree, and we calculate either precision or recall and used them as the fitness function for each query.
4) We then select the best two queries from the results.
5) One point crossover is performed between the two queries.
6) Six different mutation techniques were used (for more details see the next section).
7) Finally the Population is updated by replacing the resulted new queries (the number is two queries) with the worst two queries in the 10 Queries selected in step 2.
8) Go to step 3.

In order to use GA a set of parameters must be determined, these parameters are:

1) We decide to use 75 iterations to control the maximum number of iterations that the GA allowed to perform.
2) precision and recall are used as two fitness functions in this experiment. fitness functions are used to decide how good the solution is?

$$\mathrm{Re}call = \frac{\sum_d [rd \times fd]}{\sum_d [rd]}$$

$$\Pr ecision = \frac{\alpha \sum_d [rd \times fd]}{\sum_d [rd]} + \frac{\beta \sum_d [rd \times fd]}{\sum_d [rd]}$$

Where rd is the number of relevance documents, fd is the number of retrieved document, and α ,β are arbitrary weights. In this study and based on [24], we choose the value of α, β to be 0.25, 1.0 respectively.

3) In this study we used a single point crossover strategy with crossover probability Pc =0.8. to conduct the crossover; the best two genes (that have the best fitness values) are chosen from the population, and then represented as two trees. When the one point crossover is applied (i.e. if Random number < Probability of crossover) the two trees will exchange sub tree between them.

4) mutation now will performed. Each node from the new offsprings (the new generated genes) may be mutated based on mutation probability which equals to(pm=0.2), this mutation is implemented if the generated random number is less than probability of mutation. This paper used three mutation operators from [16] which are: a, b, and c; and proposed another three types of mutations:

   a) Boolean operator Mutation: randomly we exchange one operator (AND,OR,NOT) with another with (p=0.2).
   b) leaf node mutation: this is implemented by selecting a term randomly from the resulted offspring and replaced by any other term from the terms in the collection of documents used in the experiment with (p=0.2).
   c) The adding/deleting Mutation: this mutation is performed by inserting or deleting operator (AND,OR,NOT) between two nodes in the offsprings with (p=0.2).
   d) Mixed mutation (double mutation): this type takes into consideration the idea of multi point mutation; where two mutations are conducted over the offsprings. We have proposed three mixed types:
      1- First mutation is by inserting or deleting operator between the two nodes with (p=0.1), the second mutation is on the operator with (p=0.1).
      2- First mutation is by inserting or deleting operator between two nodes with (p=0.1), the second mutation is on the leaf node with (p=0.1).
      3- First mutation is on the operator with (p=0.1), the second mutation is on the term node (leaf node) (p=0.1).

As a result we create 12 different GA strategies for the Boolean and fuzzy models, those strategies are as following:

1) GA(M1,Precision): GA with mutation on the operator, and precision as a Fitness Function.
2) GA(M2,Precision): GA with Mutation on the leaf node, and the precision as a Fitness Function.
3) GA(M3,Precision): GA with Mutation that insert or delete operator between two nodes, and the precision as a Fitness Function.

4) GA(M4,Precision): GA that use Two mutations in the same time. first: by inserting or deleting operator between two nodes with (p=0.1), second: mutation on the operator with (p=0.1), and the precision as a Fitness Function.
5) GA(M5,Precision): GA that use Two mutations in the same time. first: by inserting or deleting operator between two nodes with (p=0.1), second: Mutation on the term node (p=0.1), and the precision as a Fitness Function.
6) GA(M6,Precision): GA that use Two mutations in the same time. first: mutation on the operator with (p=0.1), second: Mutation on the term node (leaf node) (p=0.1), and the precision as a Fitness Function.

7) GA(M1,Recall): GA with Mutation on the operator, and the recall as a Fitness Function.
8) GA(M2,Recall): GA with Mutation on the term leaf node, and recall as a Fitness Function.
9) GA(M3,Recall): GA with Mutation by inserting or deleting operator between the two nodes, and the recall as a Fitness Function.
10) GA(M4, Recall): GA that use Two mutations in the same time. first: by inserting or deleting operator between two nodes with (p=0.1), second: mutation on the operator with (p=0.1), and the recall as a Fitness Function.
11) GA(M5, Recall): GA that use Two mutations in the same time. first: by inserting or deleting operator between two nodes with (p=0.1), second: Mutation on the term node (leaf node) (p=0.1), and the recall as a Fitness Function.
12) GA(M6, Recall): GA that use Two mutations in the same time. first: mutation on the operator with (p=0.1), second: Mutation on the term node (leaf node) (p=0.1), and the recall as a Fitness Function.

6. Experiment Results

The results for all of the GA strategies are shown in Table 1, Table 2. From those tables we notice that GA(M2,Precision), GA(M5,Precision) and GA(M2,Recall) give a high improvement than user query while all other proposed methods gave us results lower than the original user query. Finally the GA(M2,Precision) achieves the highest improvement between all methods over the user query. The results for this experiment can be improved if we increase the number of iterations (more than 75) for the GA, in one hand increasing the number of iterations will improve the performance, but in the other hand this will lead to increase the run time.

Table 1: Results when Precision was used as a Fitness Function in the Boolean Model.

Recall	Traditional	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
0.1	0.156	0.161	0.169	0.146	0.149	0.156	0.152
0.2	0.162	0.162	0.173	0.187	0.154	0.169	0.159
0,3	0.166	0.167	0.179	0.174	0.163	0.184	0.167
0.4	0.178	0.169	0.191	0.172	0.170	0.189	0.175
0.5	0.188	0.178	0.203	0.182	0.179	0.200	0.179
0.6	0.221	0.213	0.232	0.219	0.208	0.222	0.218
0.7	0.223	0.219	0.243	0.225	0.218	0.231	0.225
0.8	0.241	0.236	0.256	0.233	0.230	0.237	0.235
0.9	0.245	0.239	0.265	0.239	0.239	0.245	0.245
Average	0.1977	0.1937	0.2123	0.1974	0.1900	0.2036	0.1950

Table 2: Results when Recall was used as a Fitness Function in the Boolean Model.

Recall	Traditional	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
0.1	0.156	0.152	0.166	0.145	0.144	0.156	0.145
0.2	0.162	0.159	0.169	0.157	0.151	0.158	0.159
0.3	0.166	0.167	0.173	0.168	0.160	0.160	0.167
0.4	0.178	0.176	0.187	0.176	0.168	0.176	0.177
0.5	0.188	0.187	0.194	0.179	0.174	0.182	0.183
0.6	0.221	0.219	0.228	0.211	0.201	0.210	0.206
0.7	0.223	0.226	0.238	0.216	0.212	0.221	0.217
0.8	0.241	0.245	0.251	0.232	0.220	0.248	0.229
0.9	0.245	0.243	0.261	0.241	0.232	0.252	0.238
Average	0.1977	0.1971	0.2074	0.1916	0.1846	0.1958	0.1912

References

[1] Baeza-Yates, and Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.
[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] Vaclav, S., and Dusan, H., "Using Genetic Algorithms for Boolean Queries Optimization", Ninth IASTED International Conference on Internet and Multimedia Systems and Application, ISBN 0-88986-510-8, 2005.

[5] Salton, G., McGill,M., Introduction to Modern Information Retrieval, McGraw-Hill, New-York, 1983.
[6] Hananda E, "Evaluation of Different Information Retrieval models and Different indexing methods on Arabic Documents", Phd Thesis, ARAB Academy, 2008.
[7] Yuwono, B., and Lee, D. L "WISE: A World Wide Web Resource Database System," IEEE Transaction on Knowledge and Data Engineering, ISSN: 1041-4347, Volume: 8 Issue:4, pp. 548-554, 1996.
[8] Masaharu, Y., and Makoto, H, "An Appropriate Boolean Query Reformulation Interface for Information Retrieval Based on Adaptive Generalization", International Conference on Data Engineering, pp. 148-153, 2005.
[9] M. Boughanem, C. Chrisment, and L. Tamine, "On using genetic algorithms for multimodal relevance optimization in information retrieval", Journal of the American Society for Information Science and Technology, 53(11), pp. 934–942, 2002.
[10] J. T. Horng, and C. C. Yeh, "Applying genetic algorithms to query optimization in document retrieval", Information Processing and Management, 36(5), pp. 737–759, 2000.
[11] D. Vrajitoru, "Crossover improvement for the genetic algorithm in information retrieval", Information Processing& Management, 34(4), pp. 405–415, 1998.
[12] J. Morgan, and A. Kilgour, "Personalising on-line information retrieval support with a genetic algorithm". In A. Moscardini, & P. Smith (Eds.), PolyModel 16: Applications of artificial intelligence, pp. 142–149, 1996.
[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] Mashba F, "Evaluate the Effectiveness of Genetic Algorithm (GA) in Information Retrieval Based on Arabic Documents ", Phd Thesis, Arab Academy, 2009.
[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, May 4-5, pp. 37-41, 2009.
[18] D. H. Kraft, F. E. Petry, B. P. Buckles, and T. Sadasivan, "Genetic Algorithms for Query Optimization in Information Retrieval: Relevance Feedback", In E. Sanchez, T. Shibata, and L.A. Zadeh, editors, Genetic Algorithms and Fuzzy Logic Systems Soft Computing Perspectives, Singapore, pp. 155-173, 1997.
[19] Oscar Cordn, Flix de Moya, and Carmen Zarco, "Fuzzy logic and multiobjective evolutionary algorithms as soft computing tools for persistent query learning in text retrieval environments", In IEEE International Conference on Fuzzy Systems 2004, ISBN: 0-7803-8353-2, pp. 571-576, Budapest, Hungary, 2004.
[20] Masaharu Yoshioka, and Makoto Haraguchi, "An Appropriate Boolean Query Reformulation Interface for Information Retrieval Based on Adaptive Generalization", WIRI '05 Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration, ISBN: 0-7695-2414-1, pages 145-150, 2005.
[21] Owais, S., Kromer, P., and Snasel, V., "Implementing GP on Optimizing Boolean and Extended Boolean Queries in IRs With Respect to Users Profiles", ISBN: 1-4244-0271-9 , pp. 412 – 417, 2006 .
[22] Simon, P., and Sathya, S.S., "Genetic algorithm for information retrieval", International Conference on Intelligent Agent & Multi-Agent Systems (IAMA 2009) , ISBN: 978-1-4244-4710-7, pp. 1 – 6, 2009.
[23] Kromer, P., Snásel, V., Platos, J., and Abraham, A., "Evolutionary improvement of search queries and its parameters", 2010 10th International Conference on Hybrid Intelligent Systems (HIS), pp. 147 - 152 ISBN: 978-1-4244-7363-2, 2010.
[24] 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.

First 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).

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. 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.