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) can be defined broadly 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]. The major information retrieval models 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 reflect the query. 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 [8], and Recall which is the system's ability to retrieve all related documents of a query [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. A (GA) is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics [3]. The Genetic algorithm (GA) approach has gained importance and popularity, as evident in the number of studies that have used it to improve different optimization procedures to be able to find a global solution in many problems.

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, the correct answer for each query (relevant documents) is also known in advanced.

Arabic is the official language of over than twenty one Arab countries, and it is the religious language of more than one billion Muslims around the world. The Arabic language is unique and difficult language; the difficulty comes from several sources; amongst them: it differs syntactically, morphologically, and semantically from other Indo-European languages [13]. Compared to English, Arabic language is more sparsed, 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 uniqueness and the special properties for the Arabic language, its differences from the English and the other languages, and the lack of similar studies in the literature 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 from mutation criteria to another. 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 based on English data collection. Cordn et al. [19] introduced MOGA-P, an algorithm to deal with search query optimization as a multi-objective optimization problem and compared their approach 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. In [17] the researchers used Genetic Algorithms to improve performance of Arabic information retrieval system, which based on vector space 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)

A GA is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics [3]. 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 a problem. GAs exploits the idea of the survival of the fittest and an interbreeding population to create a novel and innovative search strategy. A population of strings, representing solutions to a specified problem, is maintained by the GA. The GA then iteratively creates new populations from the old by ranking the strings and interbreeding the fittest to create new strings, which are hopefully closer to the optimum solution to the problem at hand. So in each generation, the GA creates a set of strings from the bits and pieces of the previous strings. The idea of survival of the fittest is of great importance to genetic algorithms. GAs use what is termed as a fitness function in order to select the fittest string that will be used to create new, and conceivably better, populations of strings. The only thing that the fitness function must do is to rank the strings in some way by producing the fitness value. These values are then used to select the fittest strings. 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. Reproduction: the selection of the fittest individuals based on the fitness function.

B. Crossover: is the exchange of genes between two individual chromosomes that are reproducing. In one point cross over [3] a chunk of connected

genes will be swapped between two chromosomes.

C. Mutation: is the process of randomly altering the genes in a particular chromosome. There are two types of mutation:

1) Point mutation: in which a single gene is changed.
2) Chromosomal mutation: where some number of genes is changed completely.

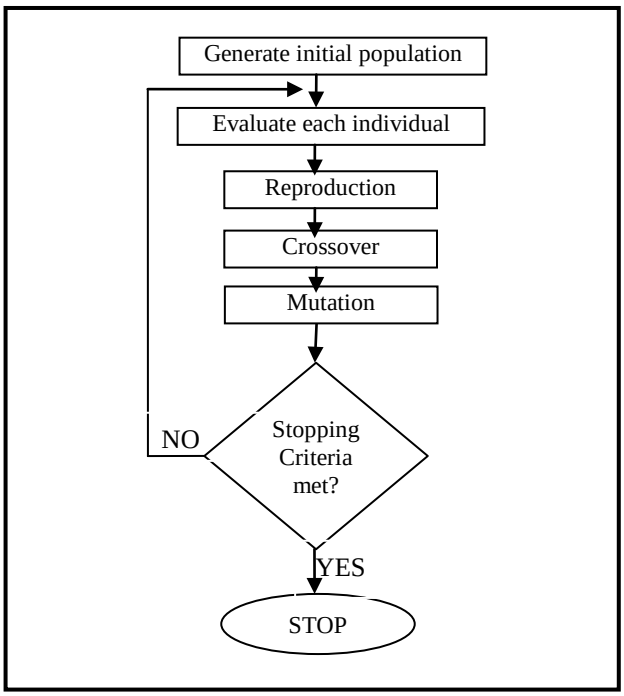

Fig. 1 Flowchart for Typical Genetic Algorithm.

5. Experiment (GA)

In this study we used IR system based on Boolean model and Fuzzy set model that was built and implemented by Hanandeh [6] to handle the 242 Arabic abstracts collected from the Proceedings of the Saudi Arabian National Conference [24]. The study was conducted as following:

1) Select the highest 15 terms frequency from the top 10 documents retrieved using the original query used by Hanandeh [6].
2) Construct 10 Queries from the selected terms.
3) Represent queries as a tree and calculate the fitness function which is either precision or recall for each query.
4) Select the best two queries.
5) Perform Crossover (one point crossover is used).
6) Perform Mutation (three different mutation techniques are used; for more details see the next section).
7) Update Population by replacing the new two queries with the worst two queries of 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) Number of generation: the number of iteration can be determined by predefined scale of accepted error, or can be defined before the GA starts. In this study the number of iterations used is 75 iterations.
2) Fitness Function Operator: Fitness function is a performance measure or reward function which evaluates how each solution is good. In this study precision and recall are used as two fitness functions.

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

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

Where rd is the number of relevance documents and fd is the number of retrieved document and α and β are arbitrary weights. In this study and based on previous studies [32] the value of α, β used is 0.25, 1.0 respectively.

3) Selection operator: In this study we used a single point crossover strategy with crossover probability Pc =0.8. The best two individuals with best fitness values are chosen from a population, and represented as 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 operators: In this experiment the mutation operator works as the most important operator for the learning of query. Each node from the new offsprings may be mutated; that depends on mutation probability (pm=0.2), this

mutation is applied if random number is less than probability of mutation. Different types of mutations are used in this study:

a) Mutation on Boolean operator: randomly exchanging one operator to another.
b) Mutation on term node (leaf node): in Boolean model one term is selected randomly from the offspring and replace by any other one from the terms in a given collection of documents. But in fuzzy model the term is not replaced, only the term weight is changed.
c) Mutation by inserting or deleting operator between two nodes in the offsprings.

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

1) GA(M1,Precision): GA that use mutation on the operator and the precision as a Fitness Function.
2) GA(M2,Precision): GA that use Mutation on the term node (leaf node) and the precision as a Fitness Function.
3) GA(M3,Precision): GA that use Mutation by inserting or deleting operator between two nodes and the precision as a Fitness Function.
4) GA(M1,Recall): GA that use Mutation on the operator and the recall as a Fitness Function.
5) GA(M2,Recall): GA that use Mutation on the term node (leaf node) and the recall as a Fitness Function.
6) GA(M3,Recall): GA that use Mutation by inserting or deleting operator between two nodes and the recall as a Fitness Function.

6. Experiment Results

The results for the GA strategies based on Boolean model are shown in Table 1, Table 2. From those tables we notice that GA(M2,Precision), GA(M3,Precision) and GA(M2,Recall) give a high improvement than user query while GA(M1,Precision), GA(M1,Recall) and GA(M3,Recall) gives a low improvement than user query. We can also notice that GA(M2,Precision) gives the highest improvement over the user query in the Boolean model. The results for our experiments can be improved for the Boolean model by increasing the number of iterations 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	GA(M1)	GA(M2)	GA(M3)
0.1	0.156	0.161	0.169	0.146
0.2	0.162	0.162	0.173	0.187
0,3	0.166	0.167	0.179	0.174
0.4	0.178	0.169	0.191	0.172
0.5	0.188	0.178	0.203	0.182
0.6	0.221	0.213	0.232	0.219
0.7	0.223	0.219	0.243	0.225
0.8	0.241	0.236	0.256	0.233
0.9	0.245	0.239	0.265	0.239
Average	0.19777	0.19377	0.21233	0.19744

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

Recall	Traditional	GA(M1)	GA(M2)	GA(M3)
0.1	0.156	0.152	0.166	0.145
0.2	0.162	0.159	0.169	0.157
0.3	0.166	0.167	0.173	0.168
0.4	0.178	0.176	0.187	0.176
0.5	0.188	0.187	0.194	0.179
0.6	0.221	0.219	0.228	0.211
0.7	0.223	0.226	0.238	0.216
0.8	0.241	0.245	0.251	0.232
0.9	0.245	0.243	0.261	0.241
Average	0.197778	0.19711	0.20744	0.19166

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] Kushchu, I., "Web-Based Evolutionary And Adaptive Information Retrieval", Evolutionary Computation, IEEE Transactions, Volume: 9, Issue: 2, ISSN: 1089-778X, pp. 117 – 125, 2005.
[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.