AN EFFICIENT ADAPTIVE GENETIC ALGORITHM FOR VECTOR SPACE MODEL

[1]WAFA' ALMA'AITAH, [2]KHALED ALMAKADMEH
[1]Department of Basic Science, Hashemite University
[2]Department of Software Engineering, Hashemite University
E-mail: [1]wafaa_maitah@hu.edu.jo, [2]khaled.almakadmeh@hu.edu.jo

ABSTRACT

In this study, we propose to use an Adaptive Genetic Algorithm (AGA) aimed to enhance the performance of information retrieval under Vector Space Model (VSM) in both (Cosine and Dice similarity). Using the algorithm is aimed to improve the quality of the results of user's query and generate improved queries that fit searcher's needs. Furthermore, we investigate and evaluate different fitness functions that reduce the search space and reduce the number of iterations needed to generate an optimized query.

Traditional Genetic Algorithms (GA) use fixed values for crossover and mutation operators; and such values remained fixed during the execution of the algorithm. The adaptive genetic algorithm uses crossover and mutation operators with variable probability; which allows for faster attainment of improved query results. The proposed approach is verified using (242) proceedings abstracts (in Arabic) collected from the Saudi Arabian National conference.

Keywords: *Information Retrieval; Adaptive Genetic Algorithm; Vector Space Model; Query Optimization.*

1. INTRODUCTION

Information retrieval (IR) handles the representation, storage, organization, and access to information items [1]. In IR one of the main problems is to determine which documents are relevant for the user's needs. In practice, the IR problem is usually mentioned as a top ranking problem, which aims to retrieve information in accordance according to the degree of relevance (similarity) between the searched document and the user query [1] [2].

Genetic algorithms started to be applied in information retrieval in order to optimize search query by using a genetic search, algorithm. A good query can be defined as a set of terms that express accurately the information need while being usable within collection corpus; the last part of this definition is critical in order to produce an efficient matching process. Traditional Genetic Algorithms used in previous research studies depends on fixed control parameters; and in particular crossover and mutation probabilities. In this research study, we propose to use an adaptive genetic algorithm depends on variable crossover and mutation probabilities to improve performance of information retrieval.

According to the natural evolution process; the use of analogies of natural action led to the development of Genetic Algorithms (GAs) of four main elements [3][4]:

- Representation of an individual's as possible solutions;
- Fitness function assigned a fitness score and indicates how good an individual is;
- Reproduction method; and
- Selection criteria that selects highly fit individuals to reproduce the offspring by crossover and mutation techniques.

2. VECTOR SPACE MODEL(VSM)

In Vector Space Model, a document is viewed as a vector in n-dimensional document space (where n is the number of distinguishing terms used to describe contents of the document in a collection). Each term represents one dimension in the document space [1], [5].

Document retrieval is based on the measurement of the similarity between the query and the documents by using different similarity measures (as shown in Table 2.1). This means that documents with a higher similarity to the query are judged to be more relevant to it and should be retrieved by the IRS in a higher position in the list of retrieved documents. In This method, the retrieved documents can be orderly presented to the user with respect to their relevance to the query [6].

$$\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$
$$\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

Table 2.1: Different Similarity Measures

Similarity Measure	Evaluation for Weighted Term Vector
Cosine	$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i}^2}}$
Dice	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w}$
Jaccard	$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1} w_{i,j}^2 + \sum_{i=1} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j}}$

3. RELATED WORK

1. Eman Al Mashagba, Feras Al Mashagba and Mohammad Othman Nassar [7]:

This research study used different similarity measures (Dice, Inner Product) in the VSM, for each similarity measure we analyzed ten different GA approaches based on:

- Different fitness functions;
- Different mutations; and
- Different crossover strategies.

This is aimed to find the best strategy and fitness function that can be used when the data collection is in Arabic language. Our results show that the different GA approaches have differences in their results; the best IR system found is the one that uses the Inner Product similarity as fitness with one-point crossover operator.

2. Abdelmgeid A. Aly [8]:

This research study presented an adaptive method using genetic algorithm to modify user's queries based on relevance judgments. This method is tested for the three well-known documents collections (CISI, NLP and CACM). The results showed that the method is applicable to large text collections; where more relevant documents presented to users in the genetic modification. The study showed the effects of applying GA to improve the effectiveness of queries in IR systems. This study is based on Vector Space Model (VSM) in which both documents and queries were represented as vectors; the weights are assigned to terms proposed by Salton and Buckle, and the system is evaluated by the precision and the recall formula.

3. Andrew T. [9]:

The author presented in this study the applicability of several techniques, like connectionist Hopfield network; symbolic ID3/ID5R; evolution-based genetic algorithms; symbolic ID3 Algorithm; evolution-based genetic algorithms; simulated annealing; neural networks; and genetic programming. The study reported that these techniques are promising in their ability to analyze user queries, information needs, and suggesting alternatives for the user search queries.

It is worth mentioning that the genetic algorithm used in [7][8][9] have only used fixed control parameters especially crossover and mutation probabilities; in which such parameters remain unchanged during execution.

4. EXPERIMENTATION

This study used an Adaptive Genetic Algorithm (AGA) that has been optimized and adapted for relevance feedback. In this section, we describe the characteristics of such algorithm; these characteristics are chosen to obtain the best performance by using crossover and mutation operators with variable probabilities where as the traditional Genetic Algorithms (GA) uses fixed values of such operators, and they remain unchanged during the algorithm execution. GA that supports adaptive adjustment of mutation and crossover probabilities allows faster attainment of search results.

Next, describe two different fitness functions; both are based on the order of retrieval in which used to guide the algorithm in the search process. We detail the characteristics of an AGA that give the best performance; these characteristics guide the algorithm in the search process in the following manner:

1. Representation of the chromosomes: the chromosomes represented using binary representation; these chromosomes have the same number of genes (components) as there are terms with nonzero weights in the query and in the documents of the feedback. The set of different terms contained in those documents and in the query are calculated firstly and the size of the chromosomes is equal to the number of terms in that set.

2. Population: Our AGA receives an initial population consisting of the chromosomes corresponding to the top 15 documents retrieved from traditional IR with respect to that query.

3. Selection: the selection process select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected).

4. Genetic operators: used two-point crossover as the crossover operator.

5. Control parameters: Crossover probability P_c and mutation probability Pm play an important role in GA. Crossover causes a randomized exchange of genetic material between chromosomes. Crossover occurs only with some probability P_c which controls the rate at which chromosome is subjected to crossover [10], [11]. The larger value P_c is, the faster is the new chromosome introduced into the population. The smaller value P_c is, the lower the searching process is leading to stagnation. Typical initial value of P_c is in the range 0.5 to1.0. The mutation probability P_m is varied according to the generations. The initial P_m is larger for the global search, and in some generations it is smaller for the local search. Finally, it is larger again for avoidance of local optimum. Typical initial value of P_m is in the range 0.005 to 0.05.

We put forward adaptive varied values of P_c and Pm as follows [11], [12]:

$$p_c = \begin{cases} p_{c1} - \dfrac{(p_{c1} - p_{c2}) * (f' - f_{avg})}{f_{\max} - f_{avg}}, f' \geq f_{avg}, \\ p_{c1}, f' \prec f_{avg} \end{cases}$$

where, f_{max} is the maximum fitness function of current generation, f_{avg} is the average fitness function of current generation, is larger fitness function of the two crossover chromosomes selected, f is the fitness function of mutation chromosome selected, p_{c1}; p_{c2} is crossover probability, and pm1, pm2 is mutation probability. The study experimental parameters include: $p_{c1} = 0:9$; $p_{c2} = 0:6$, $p_{m1} = 0:1$; $p_{m2} = 0:001$.

6. The Fitness Functions: the study ran the AGA described above with different order based fitness functions.

Fitness 1: This fitness function, due to Horng and Yeh [12], is very innovative. As well as taking into account the number of relevant and of irrelevant documents, it also takes account of the order of their appearance.

One calculates the similarity of the query vector with all the documents, and sorts the documents into decreasing order of similarity. Finally, one calculates the fitness value of the chromosome with the following formula:

$$F = \frac{1}{|D|} \sum_{i=1}^{|D|} \left(r(di) \sum_{j=1}^{|D|} \frac{1}{j} \right)$$

Where, $|D|$ is the total number of documents retrieved, and r(di) is the function that returns the relevance of document d, giving a 1 if the document is relevant and a 0 otherwise. We shall refer to this fitness function as fitness 1.

Fitness 2: cosine similarity

$$\frac{\sum_{k=1}^{t} (d_{ik} \bullet q_k)}{\sqrt{\sum_{k=1}^{t} d_{ik}^2 \bullet \sum_{k=1}^{t} q_k^2}}$$

Where, d_{ik} is the weight of term i in document k and q_k is the weight of term i in the query.

5. EXPEREMANTL RESULTS

The traditional information retrieval systems were built and implemented to handle the Arabic collection using C# NET. The following two IR systems were built and implemented:

- VSM1: System that used Vector Space Model with Cosine similarity.
- VSM2: System that used Vector Space Model with Dice similarity.

Different AGA strategies were used in this study. Those strategies are as the following:

- AGA1: AGA that use cosine similarity as fitness.
- AGA2: AGA that use Horng & Yeh formula as fitness.

5.1 Applying AGA Strategy Using Cosine Similarity

Table 5.1 show the comparison between vector space model with Cosine as fitness VSM1 (AGA1) and vector space model with Horng as fitness VSM1 (AGA2), from this table we notice that the VSM1 (AGA2) represent the best strategy over VSM1 (AGA1).

Table 5.1 Average Precision Values for 59 Queries by Applying AGAs on Vector Space Model

Recall	Average Precision	
	VSM1 (AGA1)	VSM1 (AGA2)
0.1	0.46	0.48
0.2	0.53	0.56
0.3	0.57	0.63
0.4	0.65	0.66
0.5	0.75	0.8
0.6	0.75	0.8
0.7	0.85	0.9
0.8	0.9	0.94
0.9	0.9	0.95
Average	0.651	0.736

5.2 Applying AGA Strategy Using Dice Similarity

Table 5.2 show the comparison between vector space model with Cosine as fitness VSM2 (AGA1) and vector space model with Horng as fitness VSM2 (AGA2), from this table we notice that the VSM2 (AGA2) represent the best strategy over VSM2 (AGA1).

Table 5.2 Average Precision Values for 59 Queries by Applying AGAs on Vector Space Model

Recall	Average Precision	
	VSM2 (AGA1)	VSM2 (AGA2)
0.1	0.31	0.34
0.2	0.38	0.38
0.3	0.46	0.49
0.4	0.46	0.53
0.5	0.56	0.68
0.6	0.67	0.69
0.7	0.73	0.75
0.8	0.78	0.83
0.9	0.79	0.83
Average	0.57	0.61

5.3 Comparison Between The Best AGA's Strategies

From tables 5.1 and 5.2 we notice that the VSM1 (AGA2) is better than VSM2 (AGA2) in all recall levels. Which means that VSM1(AGA2) that use cosine similarity and use cosine similarity as a fitness function represent the best IR system in VSM to be used with the Arabic data collection.

5.4 Comparison Between Best Agas Strategy With Traditional Gas

In this comparison best AGAs strategy compares with traditional GAs that based on the study done by feras Mashakbeh [13]. The results for the AGAs are shown in tables 5.3,5.4 using the average Recall and Precision relationship. From the tables 5.3 and 5.4, we notice that VSM using cosine similarity and Horng as fitness VSM1 (AGA2) compare with VSM with Cosine as fitness under traditional Genetic Algorithm VSM 1(AGA2`) gives the highest improvement over VSM1 (GA) with 55.1%. and VSM using Dice similarity and Horng as fitness VSM2(AGA2) compare with VSM using Dice similarity with precision as fitness under traditional Genetic Algorithm VSM2 (AGA2) gives the highest improvement over VSM2 (GA) with 26.4%.

Table 5.3: Comparison Between VSM1 (AGA2) and VSM (GA)

Recall	Average Precision		AGA Improvement %
	VSM (GA)	VSM1 (AGA2)	
0.1	0.16	0.48	32
0.2	0.17	0.56	39
0.3	0.18	0.63	45
0.4	0.18	0.66	48
0.5	0.19	0.8	61
0.6	0.2	0.9	7
0.7	0.2	0.9	7
0.8	0.24	0.9	66
0.9	0.25	0.9	65
Average	0.196	0.736	55.1%

Table 5.4: Comparison between VSM2 (AGA2) and VSM (GA)

Recall	Average Precision		AGA Improvement %
	VSM (GA)	VSM2 (AGA2)	
0.1	0.14	0.34	20
0.2	0.15	0.38	23
0.3	0.29	0.49	20
0.4	0.27	0.53	26
0.5	0.40	0.68	28
0.6	0.40	0.69	29
0.7	0.39	0.75	16
0.8	0.41	0.83	42
0.9	0.49	0.83	34
Average	0.28	0.61	26.4%

6. CONCLUSION

This study presented the application of an Adaptive Genetic Algorithm (AGA) with different fitness functions (Cosine and Horng) and variable operator's rate (crossover and mutation) on similarity measure (Dice and cosine) in vector space model. In this paper, we compared different adaptive genetic algorithm strategies by calculating evaluation using average recall formula and different similarity measure, then by calculating the improvement of each approach over the traditional IR system. The experimental results show that the vector space model using cosine similarity with Horng as fitness provides a better search strategy. Future work includes conducting further experimentation with different data sets that varies in size. This paper conducted an experiment by applying an adaptive strategy using the Vector Space Model; more experimentation can be conducted using different models like Fuzzy set model and probability model. Further future work may include the experimentation of the Adaptive Generic Algorithm using different operators such as type of operators, adaptive fitness, and variable size of chromosome.

REFRENCES:

[1] Christopher D. Manning ,Prabhakar Raghavan and Hinrich Schütze , "An Introduction To Information Retrieval", Book , Cambridge University Press , February 16, 2008.

[2] Cristina Lopez-Pujalte, Vicente P. Guerrero-Bote, Felix de Moya-Anegon, "Genetic algorithms in relevance feedback: a second test and new contributions", Proceedings in Information Processing and Management 39, 2003.

[3] Djoerd Hiemstra, "Using language models for information retrieval", 2000.

[4] Xiaohua Zhou, "Semantics-based Language Models for Information Retrieval and Text Mining", PhD thesis, Drexel University, 2008.

[5] Graham Bennett Falk Scholer Alexandra Uitdenbogerd," A Comparative Study of Probabilistic and Language Models for Information Retrieval", Proceedings of Nineteenth Australasian Database Conference (ADC), Vol. 75, pp 1- 10, 2008

[6] Kanaan G, hanandeh E, "Evaluation of Different Information Retrieval models and Different indexing methods on Arabic Documents", PhD Thesis, ARAB Academy, 2008.

[7] Eman Al Mashagba, Feras Al Mashagbaand Mohammad Othman Nassar, " Query Optimization Using Genetic Algorithms in the Vector Space Model", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5,no 3, September 2011.

[8] Abdelmgeid A. Aly, "Applying Genetic Algorithm In Query Improvement Problem" International Journal "Information Technologies and Knowledge", Vol.1, pp.309-316, 2007.

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

[10] Wang Lei, Shen Tingzhi, "An Improved Adaptive Genetic Algorithm and its application to image segmentation", Proceeding of 5th International Conference on Artificial Neural Network and Genetic Algorithms, pp.112–119, 2004.

[11] Mauro Annunziato, Stefano Pizzuti," Adaptive Parameterization of Evolutionary Algorithms Driven by Reproduction and Competition", Proceedings of Genetic and evolutionary computation conference, pp.1597- 1598, 2005.

[12] Horng & C.C Yeh, "Applying genetic algorithms to query optimisation in document retrieval", Proceedings of the Information Processing and Management, Vol 36, pp 737-759, 2000

[13] Feras AL-Mashakbeh ,“Evaluate the Effectiveness of Genetic Algorithm in Information Retrieval Based on Arabic Documents ”, Unpublished ph.D Thesis , Faculty of Information System and Technology, The Arab Academy for Banking and Financial Science, Amman, Jordan, 2008.