
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN: 2321-2152 www.ijmece .com
Vol 7, Issue.3 july 2019

Vector Space Model for Data Retrieval with Genetic Algorithms

MS.J.SWATHI[1], MS.G.LAKSHMI[2], SD.MEER SUBAN ALI[3]

ABSTRACT: *In order to better serve users' needs and guide them in their quest to find the precise information they're looking for among the ever-expanding body of data, information retrieval systems (IRs) frequently employ genetic algorithms to improve the information retrieval process and boost the efficiency of the optimal information retrieval. The development of adaptive evolutionary algorithms aids in the correct retrieval of user-requested information by paring down the number of retrieved relevant files and excluding unnecessary ones. The researcher in this study selected tests from the mathematics section of the Cranfield English Corpus to investigate the underlying issues, such as the selection of mutation probability and fitness function. In 1960, Cyrial Cleverdon compiled 1,400 documents and 225 questions for use in simulations at the University of Cranfield. The study also applied two suggested adaptive fitness functions, mutation operators, and adaptive crossover, and the similarity between the query and documents was calculated using cosine similarity and jaccards. Method used to evaluate the quality of outcomes using the criteria of accuracy and recall. The research indicated that applying adaptive evolutionary algorithms may result in a number of enhancements.*

KEYWORDS*: Vector space model, precision, recall, information retrieval, adaptive genetic algorithm*

INTRODUCTION

Information retrieval systems are useful because of the vast quantities of data and documents posted online by millions of authors and organizations. However, there are a number of issues that consumers may face when using an information retrieval system. Researchers have sought to improve the system's efficacy and precision via a number of studies that focus on these issues that have an impact on accuracy [13].

Through the use of adaptive evolutionary algorithms, this research demonstrated some improvements in the performance of an information retrieval system by executing various queries, utilizing many approaches to collect relevant information, and then ranking these queries according to their degree of similarity.

It should now be evident that the research intends to look at the models used for retrieving information. The researcher in this study employed two models to calculate the similarity between the query and the documents: the Vector Space Model and the Extended Boolean Model [5].

By comparing the first fitness function (Cosine) with variable probability mutations and variable probability crossover, and the second fitness function (Jaccard's) with the same, the researcher was able to achieve better results by employing a variable ratio of mutation operators [5].

The research uses a corpus of 1,400 English-language mathematical papers and 255 questions to assess the efficacy of the findings in terms of accuracy and recall [4] [7].

ASSOCIATE PROFESSOR[1,2,3]
COMPUTER SCIENCE ENGINEERING
TRINITY COLLEGE OF ENGINEERING AND TECHNOLOGY, PEDDAPALLY
(jillaswathi@gmail.com),(lakshmiraj2006feb@gmail.com), (meersubanali@gmail.com)

INFORMATION RETRIEVAL

One, the foundational processes of information retrieval systems revolve on locating documents that are similar to a user's query in order to provide the most relevant results. In this chapter [9], we will show some of the core techniques involved in identifying a matched pair, given that both documents and queries are vectors in the Vector Space Model (VSM). Two, the procedures involved in a data retrieval system are as follows:

Third, we will be clearing all documents of punctuation.

4- Eliminating all unnecessary words from each manuscript, including common filler terms like prepositions and articles [7]. [8].

Five-word stemming using the most popular stemmer, porter [8].

Sixth-document identifiers (doc IDs) are assigned and linked to elements in any document using an inverted index [7].

GENETIC ALGORITHM

The genetic algorithm is a powerful tool in data-driven random searches for optimal solutions [8].An evolutionary algorithm, or genetic algorithm, is used to improve upon a pool of potential answers to an optimization issue. Traditionally, solutions are represented in the binary system as strings of 0s and 1s, although alternative encodings are also feasible. Each candidate solution has a collection of attributes (its chromosomes) that may be modified and changed.[2]

$$Pc$$

The chromosomes of living organisms are used as a model for the encoded representation of solutions in a genetic algorithm. An analogy between a chromosome and a possible answer to a problem is made [7].

Proposed Fitness Function

A proposed fitness function after modification:

F (cosine) =

((

-.....(1)

- Where $w_{i,j}$ =weights term I in document j.
- $W_{i.k}$ = weights term I in query k.

The researcher modification of traditional fitness function (Cosine) was through adding new equation

$$\left(\sum_{i=1}^{t} [wi, \frac{q}{i}]\right)$$

to traditional cosine equation as

2- F (jaccard's) =

$$\left(\sum_{i=1}^{t} [wi, \frac{q}{i}]\right)$$

- Where $w_{i,j}$ =weights term I in document j.
- $W_{i.k}$ = weights term I in query k.

The researcher modification of traditional fitness function (Jaccard's) was through adding new

equation to traditional jaccard's equation as

Equation of crossover probability

$$f \geq f_{avg}$$

The researcher employed the probability crossover (Pc) equation (3) to calculate a variable value (not fixed) of the Crossover probability based on numerous ratio probabilities in order to construct a new chromosome (the original offspring) by swapping certain genes. [11].

The probability of a crossover occurs at a point when k1 = 0.5 and k2 = 0.9, respectively.

Fmax = maximum value of the fitness function on the selected chromosome.

F' = chromocome function fitness.

Favg = average chromosomal fitness function avarage.

Adaptive mutation

The crossover (pc) and mutation (pm) probabilities greatly affect the accuracy of

solutions and the convergence speed of genetic algorithms. By carefully choosing the ratios used, the researcher was able to provide the optimum results. Researchers used a broad variety of ratios for mutation probabilities as a result, but we chose on (0.2, 0.001) based on our in-depth familiarity with the system. We then used the fitness function to evaluate the probability associated with the ratios, ultimately selecting the most robust sample for future research [2, 5].

Equation of mutation operator probability

A good ratio is obtained by applying mutation to a chromosome produced from crossover as new chromium (mutated offspring) by flipping some genes, in order to get a new chromosome (mutated offspring) better than the original chromosome (original offspring) [11]. This was accomplished by using the probability mutation (Pm) equation (4) to determine the variable value (not fixed) of mutation probability.

$$Pm = \begin{cases} \dfrac{K_3(f_{max} - f')}{f_{max} - f_{avg}} & f' < f_{avg} \\ K_4, & f' \geq f_{avg} \end{cases} \quad ,$$

-.....(4)

2. LITERATURE REVIEW

In (2013), Wafa Zaal: She used the Vector Space Model, the logical model, and the language model to modify the genetic algorithm for her research. Instead of utilizing a constant number, as is done in conventional genetic algorithms, she made the crossover and mutation probabilities vary. And she got the greatest outcomes by increasing crossover and Mutation. Arabic text was utilized for this study. According to her findings, a 13.0% increase in performance may be achieved by using a Vector Space Model with cosine similarity [10].

In (2013), Korejo and Khuhro: Adaptive mutation was investigated, and four operators were provided for use in a genetic algorithm to determine operator mutation, notwithstanding the difficulty of putting this study to use. The author posited a solution in which the mutation rate was modified. The seed population for the following generation was chosen with each operator mutation in mind. Finally, our findings showed that adaptive mutation led to the best results at work [2].

In (2012), Ammar Sami: Through the use of crossover, mutation operator, and specialized fitness function, he proposes a genetic algorithm-based research methodology to enhance online information retrieval systems, and to apply information retrieval through a genetic algorithm to split the work into two units, document indexing unit and genetic algorithm unit. At long last, it achieved a 90% success rate [13].

In (2009), Noha Marwan: She employed Jaccard's and Ochiai's fitness functions and worked with four different Islands to get the data. She evaluated the islands using both sets of criteria separately. She used an enlarged query, presented the outcomes, and contrasted the two sets of findings for the four islands. She demonstrated that Jaccard's method was superior to Ochiai's in terms of random selection and that Ochiai's method was superior to Jaccard's in terms of unbiased model tournament selection [14].

EXPERIMENTAL RESULTS

Results

There were 10 questions asked in this investigation. There were typically 8 statements in each query. 80 findings were obtained by the researcher. After collecting data, the researcher performed statistical analysis using predetermined criteria. The researcher decided to just provide the results from the first inquiry.

• Table 1 shows that the Vector Space Model (VSM) and the cosine fitness function are employed by IRs that use the adaptive genetic algorithm (AGA). Recall rose but accuracy fell in all circumstances due to the rising number of samples, and the results of using cosine in table 1 were better than the result of using suggested cosine with VSM in table 2.

Table 1: value of precision and Recall for query number1 by using (VSM) and (AGA) using the cosine fitness function.

Recall	Precision (%)
0.1	85
0.2	75
0.3	72
0.4	64
0.5	50
0.6	38
0.7	32
0.8	22
0.9	20

• In table 2, it was shown that the Vector Space Model (VSM) and the cosine fitness function were employed by the adaptive genetic algorithm (AGA) in IRs. It has shown papers in an order that makes sense; the researcher observed superior assessment findings, a higher degree of similarity, and a higher average accuracy than with the standard cosine. Since the AGA gave more weight to term query and has superior crossover and mutation probability, this result had topped the table with a 91% accuracy rate and a recall value of 0.1.

Table 2: value of precision and Recall for query number1 by using (VSM) with (AGA) using the proposed cosine fitness function.

Recall	Precision (%)
0.1	91
0.2	85
0.3	79
0.4	65
0.5	54
0.6	45
0.7	39
0.8	28
0.9	22

• Table 3 demonstrates that the adaptive genetic algorithm (AGA) with the Vector Space Model (VSM) and jaccard's fitness function is utilized in IRs. Table 4 shows that when the researcher used the recommended jaccard's and VSM, the outcomes improved. In addition, the researcher could observe that recalls went up as accuracy went down across the board in this table. However, there was a poor outcome with an 80% accuracy rate and a recall value of 0.1 compared to the following example.

Recall	Precision (%)
0.1	80
0.2	70
0.3	60
0.4	58
0.5	43
0.6	33
0.7	29
0.8	21
0.9	19

• In table 4, we see that the adaptive genetic algorithm (AGA) with the Vector Space Model (VSM), and the suggested fitness function of jaccard, is applied in IRs. Researchers observed improved assessment outcomes and a higher degree of similarity when applying the suggested jaccard's and jaccard's fitness function. By using the recommended jaccard's fitness function, the researcher improved his or her average accuracy. However, employing the recommended cosine was the best scenario out of the four considered in (VSM).

Table 4: value of precision and Recall for query number1 by using (VSM) with (AGA) using proposed jaccard's fitness function.

Recall	Precision (%)
0.1	87
0.2	76
0.3	71
0.4	64
0.5	50
0.6	35
0.7	30
0.8	25
0.9	21

Table 5: Average value of precision for all queries using VSM with cosine

Recall	Average precision Cosine (%)	Average precision proposed cosine (%)	AGA Improvement (%)
0.1	85	92	7
0.2	76	85	9
0.3	72	80	8
0.4	65	68	3
0.5	51	55	4
0.6	39	45	6
0.7	33	38	5
0.8	23	28	5
0.9	20	23	3
average	51.5	57.1	5.6

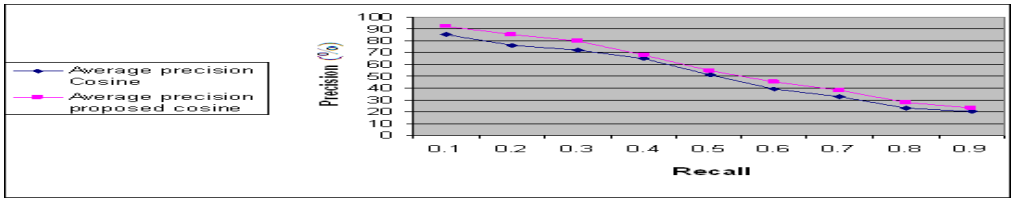

Figure 1: Average value of precision for all queries using VSM-cosine

• Cosine and a suggested cosine were employed in this instance of the Vector Space Model. The average accuracy while using suggested cosine was higher than when employing cosine, as shown, and the precision decreased as the recall increased. In addition, the degree of enhancement was satisfactory in terms of the best recall value, as a precision rate and recall value of 0.1 are excellent for all queries while using a small sample size. Unfortunately, the recall value was just 0.9. This is a positive outcome since adaptive crossover and mutation were used to improve the ratio probability, and the fitness function was modified to provide more weight to each query phrase. However, as can be shown in Figure 1, utilizing VSM with cosine yielded better results than Jaccard's fitness function. Table 6: Average value of precision for all queries using VSM with jaccard's

Recall	Average precision Jaccard's (%)	Average precision proposed Jaccard's (%)	AGA Improvement (%)
0.1	80	87	7
0.2	71	76	5
0.3	62	70	8
0.4	57	65	8
0.5	42	47	5
0.6	32	35	3
0.7	30	31	1
0.8	21	25	4
0.9	19	20	1
average	46	50.6	4.6

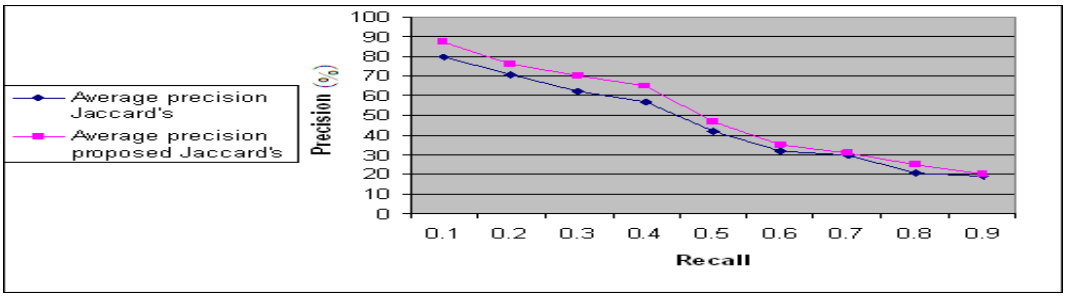

Figure 2: Average value of precision for all queries using VSM-jaccard's

• A Vector Space Model with both the original and suggested Jaccard dimensions was employed here. Compared to traditional jaccard's, the suggested version's average accuracy is clearly higher. In addition, recall was improving but accuracy was going down. All queries had an excellent degree of improvement, good top recall and average accuracy rate and a good recall value of 0.1, but a mediocre recall value of 0.9. Adaptive crossover and mutation, together with a revised fitness function, led to this positive outcome. In contrast to jaccard's, however, VSM with cosine led to a more significant improvement.

Table 7: using Vector Space Model with fitness function option.

option	Cosine	Proposed cosine	Jaccard's	Proposed Jaccard's
Average Precision(%)	51.5	57.1	46	50.6

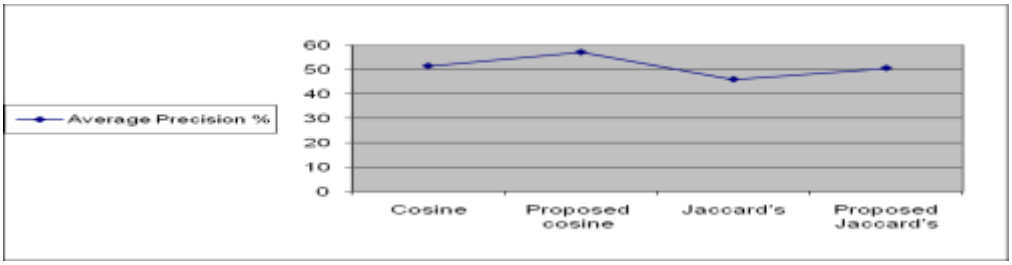

Figure 3: Using VSM with function fitness option

After experimenting with several fitness function modifications and the probability crossover and mutation operator, the researcher found the optimal solution.

Results comparing the efficacy of various improvement techniques for each IR model with each fitness function were shown in table 8.

Table 8: Compare between improvements

	Average Improvement Proposed cosine (%)	Average Improvement Proposed jaccard's (%)
VSM	5.6	4.6

3. CONCLUSION

By combining the Vector Space Model (VSM) with a number of fitness functions (cosine, suggested cosine, jaccard's, and proposed jaccard's), the researcher presented an adaptive genetic algorithm (AGA) to improve IRs. Therefore, the following is what the researcher found:

In order to recover relevant documents with an average accuracy of (57.1%), the best result occurred when combining the adaptive crossover and mutation operator with the VSM-proposed cosine.
The best outcome for locating relevant documents was achieved with a superior initial population generation.

Third, the researcher found a significant degree of similarity between all relevant documents when applying the cosine given by the VSM.
The recommended cosine yielded the most accurate VSM results (57.1) on average.
5. The greatest result emerged when utilizing suggested cosine with an improvement degree of 5.6% when compared to the use of cosine with an average accuracy of (51.5%) and the use of proposed cosine with an average precision of (57.1%) accompanying VSM.
The greatest result showed when utilizing suggested jaccard's with an improvement degree of 4.6% when comparing the usage of jaccard's with an average precision of (46%) and the use of jaccard's addition with an average accuracy of (50.6% accompanied VSM).

1. FUTURE WORK

1. The researcher concluded the following after completing the study, which aimed primarily at enhancing information retrieval utilizing genetic algorithm and adaptive methods:

2. First, future research may make use of models other than the Vector Space Model, such as the Extended Boolean Model and the Probabilistic Model.

3. Future research may use strategies besides than adaptive crossover and adaptive mutation.

.

REFERENCES

According to [1] "Web Information Retrieval Using Genetic algorithm Particle Swarm Optimization," written by Priya Borkar and Leena Patil and published in the worldwide Journal of Future Computer and Communication, Volume 2, Issue 6, Pages 595–599.
According to University of Sindh, Vol.45, pp.41-48 (2013), "Genetic Algorithm Using an Adaptive Mutation Operator for Numerical Optimization Functions" by Korejo and Khuhro.
According to Eldos (2013, Vol. 3, No. 2, pp111-124) "Mutative Genetic Algorithms" in the Journal of Computations & Modelling.

Based on the work of Mohammad Nassar, Feras AL Mshagba, and Eman AL mshagba ("Improving the User Query for the Boolean Model Using Genetic Algorithms"), published in the IJCSI International Journal of Computer Science Issues, Volume 8, Issue 5, Number 1 (pp66-70), 2011.
According to [5] "A Comparative Study of Adaptive Mutation Operators for Genetic Algorithms" by Imtiaz Korejo, Shengxiang Yang, and ChangheLi, published in The VIII Metaheuristics International Conference.
Huifang Cheng of Handan, China, "Improved Genetic Programming algorithm," International Asia Symposium on Intelligent Interaction and Affective Computing, ieee, pp168-177, 2009.
From the World Academy of Science, Engineering, and Technology, pp1021-1027 (2008), comes "Using Genetic Algorithm to Improve Information Retrieval Systems" by Ahmed Radwan, Bahgat Abdel Latef, and Abdel Ali, Osman Sadek.
According to [8] Detelin Luchev, "APPLYING GENETIC ALGORITHM IN QUESTION IMPROVEMENT PROBLEM," International Journal "Information Technologies and Knowledge," Vol.1,No. 1, pp309- 216, (2007).
[9] NIR OREN, "Reexamining tf.idf based information retrieval with Genetic Programming", University of the Witwatersrand, paper, pp1-10, (2002).
Reference: [10] Wafa. Maitah, Mamoun. Al-Rababaa, and Ghasan. Kannan, "IMPROVING THE EFFECTIVENESS OF INFORMATION RETRIEVAL SYSTEM USING ADAPTIVE GENETIC ALGORITHM", International Journal of Computer Science & Information Technology (IJCSIT), Vol 5, No. 5, pp91-105, (2013).
In 2004, researchers from Istanbul Technical University's Electrical and Electronics Faculty's Department of Computer Engineering published "A Gene Based Adaptive Mutation Strategy for Genetic Algorithms" in the journal LNCS (3103), pages 271-281.
[12] Sima Etaner and Gulsen Cebiroglu, "An Adaptive Mutation Scheme in Genetic Algorithms for Fastening the Convergence to the Optimum", Istanbul Technical University, Computer Engineering Department, 2005.
Retrieved from "Information Retrieval" (pdf, pages. 1-11) at http://www.dsoergel.com/NewPublications/HCIEncyclopediaIRShortEFForDS.
Based on the work of Kalayanasaravan and Thangamani ("Document Retrieval System Using Genetic Algorithm"), published in Kongu Engineering College, Perundurai, Volume 2, Issue 10, Pages 943-946, 2013.