International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

A Survey on improving performance of Information Retrieval System using Adaptive Genetic Algorithm

Prajakta Mitkal[1], Prof. Ms. D.V. Gore[2]

[1]Modern College of Engineering Shivajinagar, Pune

[2]Modern College of Engineering Shivajinagar, Pune

Abstract

Today information retrieval, most probably relevant information retrieval is the centric issue worldwide. Information retrieval can be of any type such as image retrieval, text retrieval and multimedia retrieval and can be retrieved using various search engines. Text retrieval is the common thing in today's era, but the issue to be discussed is relevance and the performance of the Information Retrieval System. There are numerous techniques and algorithms available for improving the performance of Information Retrieval System such as Ranking Algorithm, relevance feedback, tokenization, etc. Genetic Algorithm is the famous for efficient optimization and robustness motivated from Darwin's Principle of natural selection and survival using fitness function. Genetic Algorithm can be further extended to Adaptive Genetic Algorithm with Cosine Similarity and Horng & Yeh approach and can be applied on Hadoop Distributed File System. Hadoop Distributed File System with Adaptive Genetic Algorithm may definitely increase the performance of Information Retrieval System.

Keywords: Information Retrieval System (IRS), Genetic Algorithm (GA), Adaptive Genetic Algorithm (AGA), Hadoop Distributed File System (HDFS)

1. INTRODUCTION

1.1Information Retrieval and IRS

Information Retrieval is a method of searching information in documents; documents themselves or metadata that describes these documents. This can be search in the local database or on the internet for text, image or multimedia [2]. For retrieving the information the process begins with the entry of query by user, which can be simple formal statement of natural language. An object is the entity that represents the information in the database. Query not only has the single object in the collection, but it may have several objects that are matching the query, with different degree of relevance.

IRS can compute a numeric score on how well each object in the database matches the query and rank the objects according to these values. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

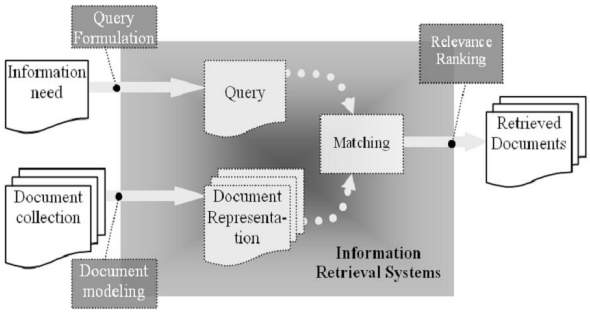

Figure.1 Formal IRS structure

There are three components of IRS that are Documentary database, Query subsystem and Matching function.

- Query subsystem is a system that allows user to form the query and present the retrieved relevant documents by the system.
- Matching function compares both query and documents in the database and gives the value which measures the similarity between the document and query.
- Documentary database is the storage space where all the documents are stored, along with documents it also represents their information content present in the documents.

Issues of IRS are Effectiveness, Efficiency and Execution time.

- Effectiveness means retrieving the most relevant information from big data.
- Efficiency deals with the amount of documents retrieved in small amount of time.

2. TEXT INFORMATION RETRIEVAL USING AGA

The text information retrieval using AGA can work in following manner:

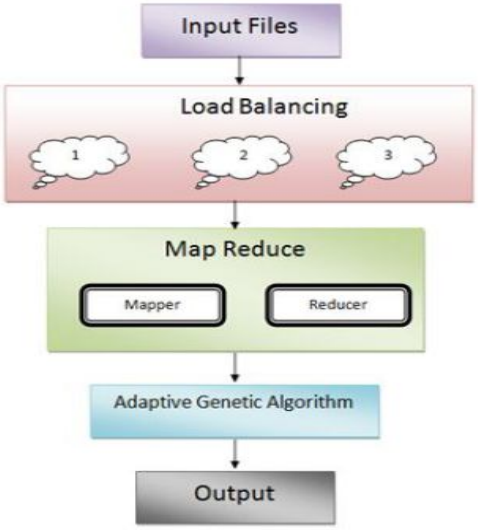

Figure.2 Information retrieval using AGA

2.1Input Files

The corpus of text is collected from various resources such as Pubmed, abstracts from IEEE explore, etc.

2.2Data Repository

Hadoop distributed file system can be used for information retrieval system [4]. It is part which stores all the collected abstracts at one location, which will act as central server. Here the input files are pushed on the central server. This central server is present in Hadoop cluster which act as Master and remaining NameNode act as slaves. Hadoop cluster consist of copying, creation and insertion operation with the help of Map-Reduce approach [4].

2.3HDFS

Apache Software Foundation has designed the project HDFS to store data reliably even in the presence of failures, which provides fault tolerance file system to run on commodity hardware. These failures may include Namenode failure, DataNode failure and Network partition. HDFS uses the master/slave policy in which one device is considered as master server which manages the one or more devices called as slaves. HDFS cluster consists of Namenode and a master server which manages the file system namespace and regulates access to various files. HDFS consist of Map/Reduce software framework, designed for processing large and distributed data set.

2.4Online IRS

Firstly the query is inserted using online information retrieval system by user [2]. After that tokenization operation is performed, this has following steps:

- Extraction of all the words from each document.
- Elimination of the stop-words.
- Stemming the remaining words using the porter stemmer, this is most commonly used.
- Inverted file indexing

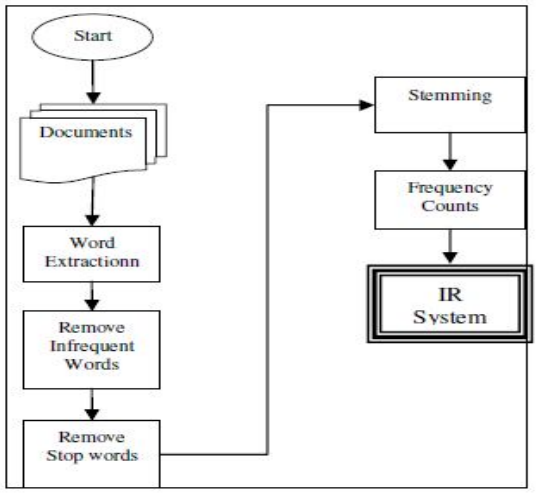

Figure.3 Tokenization in IRS

- After determining the terms that described the documents, the weights are assigned and term frequency and inverse document frequency is calculated.[5]
- Evaluate the retrieved document using average Recall and Precision formula [5].

2.5Genetic Algorithm

Genetic Algorithm (GA) is a probabilistic algorithm simulating the mechanism of natural selection of living organisms and is often used to solve problems having expensive solutions.

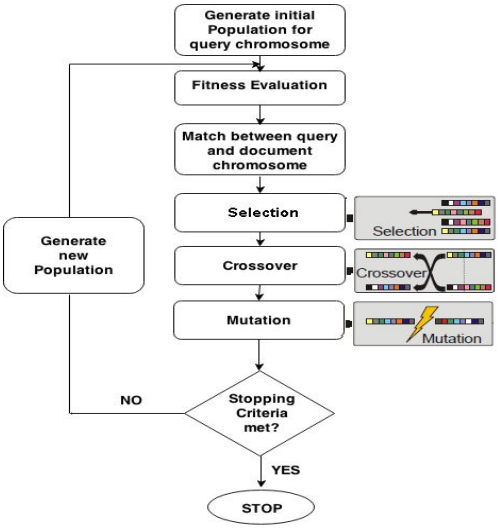

Figure.4 Genetic Algorithm

In GA, the search space is collection of candidate solutions to the problem; each represented by a string which is termed as chromosome. Each chromosome has an objective function value, called as fitness. A collection of chromosomes associated with fitness is referred as Population. This population, at a given iteration of the GA is called as generation [2]. When GA is applied to solve a problem, the first step is to define a representation that describes the problem states. The most commonly used representation that describes the problem states. The most commonly used representation is the bit string. An initial population is then defined, and three genetic operations

that are selection, crossover and mutation are performed to generate the next generation. This process is repeated until the termination criterion is satisfied. This is also called as Simple Genetic Algorithm.

2.6Adaptive Genetic Algorithm

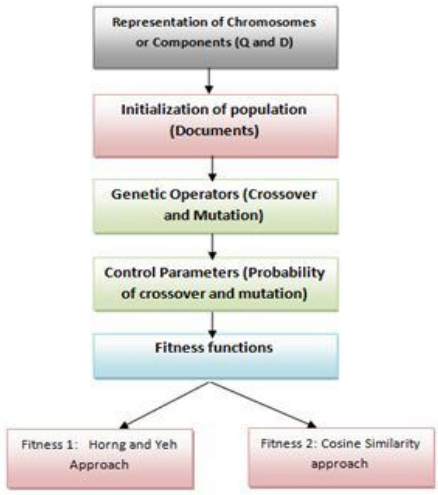

Figure.5 Adaptive Genetic Algorithm

Adaptive Genetic Algorithm is used to optimize and adapt the relevance feedback, which include best performance by using crossover and mutation operators with variable probabilities, where as the traditional genetic algorithm (GA) uses fixed values of those, and remains unchanged during execution. This developed Genetic Algorithm supports adaptive adjustment of mutation and crossover probabilities, which allows faster attainment of better solution, and then we describe two different fitness functions [5].

2.6.1Representation of the Chromosomes

Chromosomes are represented in binary format and are converted to a real representation by using a random function [6]. AGA work with chromosomes using weights of terms representation, and have the same number of genes (components) as the query and the documents have terms with non-zero weights [5].

2.6.2The population (Selection operator)

In Adaptive Genetic Algorithm an initial population is received which contains the chromosomes corresponding to the relevant documents and this population is represented by terms of weight [5].

2.6.3Genetic Operators

The two operators are used as genetic operators that are crossover and Mutation. Crossover is one of the basic operators of Genetic Algorithm. In crossover two or more parent chromosomes are selected and pair of genes are interchanged with each others. While mutation operator is a process in which gene of the chromosome is changed.

2.6.4Control parameters

The values of the control parameters crossover probability (pc) and mutation probability (pm) are variables that plays an important role in Genetic Algorithm [8].

2.6.5Fitness function

Fitness function is a performance measure or reward function, which evaluates how each solution is good [6].

Sr No.	Similarity Measure	Weighted Term Vector
1	Cosine Similarity	$F=\frac{\sum_{i=1}^{t}(d_{ik}\cdot q_{k})}{\sqrt{\sum_{i=1}^{t}d_{ik}^{2}\cdot\sum_{i=1}^{t}q_{k}^{2}}}$
2	Horng and Yeh	$F=\frac{1}{\lvert D\rvert}\sum_{i=1}^{\lvert D\rvert}\left(r(di)\sum_{j=1}^{\lvert D\rvert}\frac{1}{j}\right)$

Table.1 Fitness Function

Fitness function is a performance measure or reward function, which evaluates how each solution is good [6]. Adaptive Genetic Algorithm uses two different functions to determine the fitness values that are Cosine Similarity and Horng and Yeh formula. In comparative study of both fitness functions Horng and Yeh formula gives effective result for retrieving information.

2.7Output

The list of top ranked relevant documents will be generated as output to the users query.

3. RESEARCH REVIEW

Research Paper	Publication and Author	Techniques	Conclusion
An improved Matching functions for Information Retrieval using GA	IEEE, 2013 Prof. Anuradha D. Thakare and Dr. C. A. Dhote	Genetic Algorithm, IRS, Recall and Precision	Using GA in IRS increases the performance of IRS.
Genetic Algorithm for Information Retrieval	IEEE, 2009 Philomina Simon and S. Siva Sathya	Genetic Algorithm, IRS	GA in IRS gives optimized result.
An Effective optimized Genetic Algorithm for scalable information retrieval from cloud using big data	Science Publication, 2014 Palson Kennedy and T V Gopal	Load balancing techniques, HDFS, Genetic Algorithm, Cloud	Genetic Algorithm can be applied in HDFS environment .
Improving the Effectiveness of IRS using AGA	IJCSIT, October, 2013 Wafa Maitah, Mamoun Al-Rababaa and Ghasan Kannan	IRS, AGA	Applying AGA on different models such as VSM, EBM gives more efficient result as compare to GA.

4. EXISTING RESULTS REVIEW

Adaptive Genetic Algorithm uses two different fitness functions that are Cosine Similarity and Horng & Yeh formula. These two approaches are applied on different models such as Vector Space Model, Extended Boolean Model and Language Model and the results are as shown below:

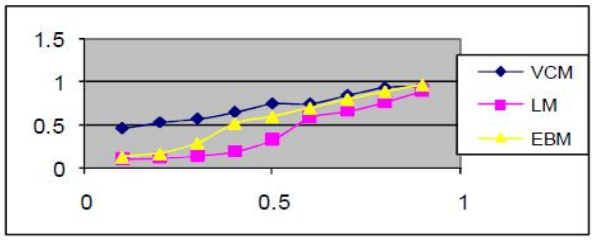

Figure.6 Average Recall and Precision values for queries by applying Adaptive Genetic Algorithm with Cosine Similarity fitness

Figure.6 shows the comaprision between Vector Space Model, Extended Boolean Model and Language Model with Cosine as fitness. From figure we can conclude that Vector Space Model represent the best strategy over Extended Boolean Model and Language Model.

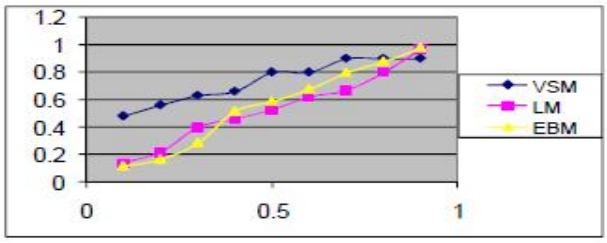

Figure.7 Average Recall and Precision values for queries by applying Adaptive Genetic Algorithm with Horng & Yeh fitness

Figure.7 shows the comaprision between Vector Space Model, Extended Boolean Model and Language Model with Cosine as fitness. From figure we can conclude that Vector Space Model represent the best strategy over Extended Boolean Model and Language Model.

Recall	average Recall and Precision		AGA Improvement %
	GA	AGA2	
0.1	0.17	0.13	-0.307
0.2	0.17	0.17	0.00
0.3	0.18	0.29	0.611
0.4	0.19	0.52	1.736
0.5	0.2	0.6	2.00
0.6	0.23	0.7	2.043
0.7	0.24	0.8	2.33
0.8	0.26	0.89	2.423
0.9	0.27	0.97	2.592
Average	0.212	0.563	1.434

Table.2 Comparison of GA and AGA using Horng and Yeh function

Table 2 shows that Adaptive Genetic Algorithm definitely shows the improvements in retrieving relevant document.

Recall	average Recall and Precision		AGA Improvement %
	GA	AGA1	
0.1	0.16	0.48	2.00
0.2	0.17	0.56	2.29
0.3	0.18	0.63	2.50
0.4	0.18	0.66	2.66
0.5	0.19	0.8	3.21
0.6	0.2	0.9	3.50
0.7	0.2	0.8	2.00
0.8	0.24	0.9	3.30
0.9	0.25	0.9	2.60
Average	0.196	0.736	2.924

Table.3 Comparison of GA and AGA using Cosine Similarity function

Table 3 shows results of Genetic Algorithm and Adaptive Genetic Algorithm using recall and precision and it also shows that adaptive genetic algorithm works effectively rather that traditional genetic algorithm. The results discussed here are the comparative study of Genetic Algorithm and Adaptive Genetic Algorithm and it also shows improvement of Adaptive Genetic Algorithm. But in proposed approach the same approach will be compared but in Hadoop Distributed File System, which definitely shows more improvement in existing results.

5. CONCLUSION

Genetic Algorithm recommends search ability which is domain independent. It is capable to improve the performance of information retrieval system. Using Adaptive Genetic Algorithm, which is improved approach of existing traditional Genetic Algorithm in Hadoop Distributed File system also shows the performance improvement in Information retrieval and this performance is calculated by using evaluation measures precision and recall.

References

[1] Prof. Anuradha D. Thakare and Dr. C.A. Dhote, "An Improved Matching Functions for Information Retrieval Using Genetic Algorithm," ICACCI 2013, IEEE, pp. 770-774.

[2] Philomina Simon and S. Siva Sathya, "Genetic Algorithm for Information Retrieval," IAMA 2009, IEEE.

[3] Venkata Udaya Sameer and Rakesh Chandra Balabantaray, "Improving ranking of webpages using user behaviour, a Genetic algorithm approach," First International Conference on Networks & Soft Computing @ 2014, IEEE.

[4] Palson Kennedy R. and T.V. Gopal, "An Effective optimized Genetic Algorithm for scalable information retrieval from cloud using big data," Journal of computer science, Science Publication@ 2014, pp. 1026-1035.

[5] Wafa Maitah, Mamoun Al-Rababaa and Ghasan Kannan, "Improving the Effectiveness of Information Retrieval System using Adaptive Genetic Algorithm," International Journal of Computer Science & Technology(IJCSIT) Vol 5, No. 5, October 2013.

[6] Ahmed A.A Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali and Osman A. Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems," Proceedings of world academy of science engineering and technology, vol 17 December 2006, ISSN 1307-6884.

[7] Dana Vrajitoru, "Genetic Programming Operators Applied to Genetic Algorithms".

[8] Sean Luke and Lee Spector, "A comparision of Crossover and Mutation in Genetic Programming," Proceedings of the Second Annual Conference on Genetic Programming, 1997.

[9] Manoj Chahal and Jaswinder Singh, "Effective Information Retrieval Using Similarity function: Horng and Yeh Coefficient," International Journal of advanced research in Computer Science and Software Engineering Vol 3, Issue 8 August, 2013.

AUTHOR

P**rajakta Kantilal Mitkal** received the Bachelor's of Engineering degree (B.E) in Computer Science and Engineering in 2010 BMIT, Solapur. She is now pursuing Master's degree in Computer Engineering at P.E.S's Modern College of Engineering, Pune. Her current research interests include Information Retrieval and Data Mining.

Prof. Ms. D.V. Gore is currently working as Assistant Professor in Computer Engineering Department at P.E.S's Modern College of Engineering, Pune (India). She has completed her Postgraduate studies at D.Y.Patil's College of Engineering, Akurdi, pune