ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Improving the Performance of Information Retrieval System using AGA in Distributed Environment

Prajakta Mitkal, Prof. (Ms.) Deipali Gore

M.E. Student, Dept. of Computer Engineering, P.E.S's MCOE, Savitribai Phule University, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, P.E.S's MCOE, Savitribai Phule University, Pune, Maharashtra, India

ABSTRACT: Today's era of technology and advancements, the centric issue is in focus of Data Mining, which mainly include Information Retrieval (IR) or rather it can be said as Relevant Information Retrieval. Information Retrieval has become a challenging area of researcher's, because there is rapid growth of computer and Internet technologies. This rapid growth generates huge amount of text-based data, for example, web pages, e-mails, news articles, etc. Manually handling this large amount of corpus is quiet complicated, expensive and in feasible. To handle and organize this large amount of data many techniques are available such as Text classifiers, Text categorization, etc. The main issue is Information retrieval from this categorized or classified text. There are many approaches or algorithms for IR such as clustering, fusion, tokenization, Virtual Center based Matching function, K-Nearest Neighbor, Naive Bayes algorithm, Genetic Algorithm, Decision Tree, etc. The main objective of the researcher is extract relevant information from large amount of data which should be beneficial for increasing effectiveness, scalability, reliability and efficiency of the Information Retrieval System. The preliminary focus is to load the corpus on HDFS using Map-Reduce technique. After Map-Reduce technique, the preprocessing of the documents is done that includes tokenization, stop-word removal and stemming of the documents. The performance of IRS is calculated by fitness values of Genetic algorithm and Adaptive Genetic Algorithm.

KEYWORDS: Information retrieval system (IRS), Genetic Algorithm (GA), Adaptive Genetic Algorithm (AGA), Map-Reduce.

I. INTRODUCTION

The idea of Information Retrieval evolved from the word information, while information is an extension of data. Data raw, unorganized facts whose processing is must. Actually data may be something simple, apparently unsystematic and useless, until it is organized. Information is a data which is processed, organized, structured or presented in a given context so as to make it useful. Information is presented or classified data which has some meaningful values for the receiver. Information Retrieval is a method of searching information in documents; document themselves or metadata that describes these documents [1]. This can be search in the local database or on the internet for text, images, sound or data. Actually the process begins when any user enters a query into the system, which can be just the formal statements. It is not just that query has a single object in the collection; it may have several objects that match the query, with different degree of relevance. An object is the entity that is represented by information in database. Generally information system can compute a numeric score on how well each object in the database matches the query and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query. The ranking of objects can be improved by using Genetic algorithm [2].

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. In Genetic Algorithm, the search space is collection of candidate solutions to the problem; each represented by a string is termed as a chromosome. Each chromosome has an objective function value, called fitness. A collection represented in set of chromosomes together with their associated

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

fitness is called the population. This population, at a given iteration of the genetic algorithm, is called a generation [3]. When genetic algorithm 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 [4].

Adaptive Genetic Algorithm has been optimized and adapted for relevance feedback. As characteristic of Adaptive Genetic Algorithm has chosen for having the best performance by using crossover and mutation operators with variable probabilities, where as the traditional genetic algorithm uses fixed values of those and remains unchanged during execution [4]. Developed Genetic Algorithm supports adaptive adjustment of mutation and crossover probabilities; this allows faster attainment of better solutions and then we describe fitness functions, based on the order of retrieval, which we used to guide the algorithm in the search process. This fitness function is Cosine Similarity which is applied on different models such as Vector Space Model, Extended Boolean Model and Language model [5].

One of the most frequently mentioned, and most highly rated, issue is efficiency. Many systems have an impact on efficiency, and metrics such as response time of query and indexing speed are major concern of virtually every company involved with text based systems. The goal of Apache Software Foundation is to design the project to provide fault tolerant file system to run on commodity hardware. HDFS is the sub-project of Apache Hadoop Foundation whose objective is to store data reliably even in the presence of failures. These failures may include NameNode failures, DataNode failures and Network partitions. HDFS uses the master/slave approach in which one device is considered as master, which controls one or more devices which are referred as slaves. HDFS cluster consists of Namenode and a master server which manages the file system namespace and regulates assess to various files [6, 7]. Hadoop Map/Reduce is a software framework for processing large and distributed data set on compute clusters of commodity hardware. It handles scheduling of tasks, monitoring the tasks and re-executing the failed tasks [8].

II. RELATED WORK

Paper [1], relates with solution for searching information in a given set of objects. Objects here are documents which can be text, image, video or audio. It also helps to understand how genetic algorithm is used in Information Retrieval and under which conditions. Our main aim is to obtain optimized result of given query and Genetic algorithm is basically used to solve optimization problems.

Paper [3], include the review of information retrieval using Genetic Algorithm. Firstly here unstructured information is considered as input then by applying Load balancing techniques the input is categorized into different clouds, then mapper and reducer algorithm is used, so that load can be well balanced. Mapper and Reducer generally act with the concept of key-value pair and finally Genetic algorithm is applied to retrieve effective and efficient information using genetic operators that are crossover and mutation.

Paper [4], depicts the concept of information retrieval using Adaptive Genetic Algorithm. Adaptive Genetic Algorithm is one step ahead to Genetic Algorithm, which is use of Cosine Similarity and Horng and Yeh approach. In this paper Adaptive Genetic Algorithm is applied on three different models that are: Vector Space model, Extended Boolean Model and Language Model. The results that obtained are Vector space model represent a best strategy with Horng and Yeh approach rather than Cosine Similarity approach, Extended Boolean Model represents a best strategy with Cosine similarity rather than Horng and Yeh approach and Language Model represents best strategy with Horng and Yeh approach rather than Cosine Similarity. Adaptive Genetic Algorithm follows the steps such as Selection, Crossover and mutation.

III. PROPOSED ALGORITHM

A. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is to improve the performance of Information Retrieval System using adaptive genetic algorithm, in terms of effectiveness of the relevant document and efficiency. The proposed algorithm is consists of three main steps.

Step 1: Input Files and Map-Reduce Techniques:

The corpus used for information retrieval is abstracts from IEEE papers of year 2010-15 as training dataset. This input data is loaded on HDFS through user. HDFS submits the job to the JobScheduler that consist of

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Map-Reduce. Using Map-Reduce input data is converted in key-value pair and give the output data to the user that can be extracted from output HDFS.

The corpus used for information retrieval as testing dataset is news articles of Times of India newspaper of year 2013. These documents are tested without HDFS, using GA and AGA algorithms.

Step 2: Pre-processing:

The preprocessing of input data is carried to avoid the irrelevant information available. The preprocessing of input data includes the steps as follows:

1) Tokenization: It includes extraction of word token (index terms) from running text. For example, if the text is 'Places to be visited in Mumbai', then the output generated after tokenization is [places, to, be, visited, in, Mumbai].
2) Stopword Removal: Words that do not have any disambiguation power such as articles, prepositions, etc are removed. This step is occurred after the tokenization. The given list of tokens by tokenizer is striped as [places, visited, Mumbai].
3) Stemming: The remaining tokens are then stemmed to the root form. Hence the final list becomes [place, visit, Mumbai], using Porters algorithm.
4) Term weighting: The TF-IDF is the best known weighting scheme in information retrieval, which depicts the relevance of term in the document. The TF-IDF can be calculated by using eq. (1) as:

$$w_{t,d} = \log(1 + tf_{t,d}) \times \log_{10}(N / idf_t) \quad \text{eq. (1)}$$

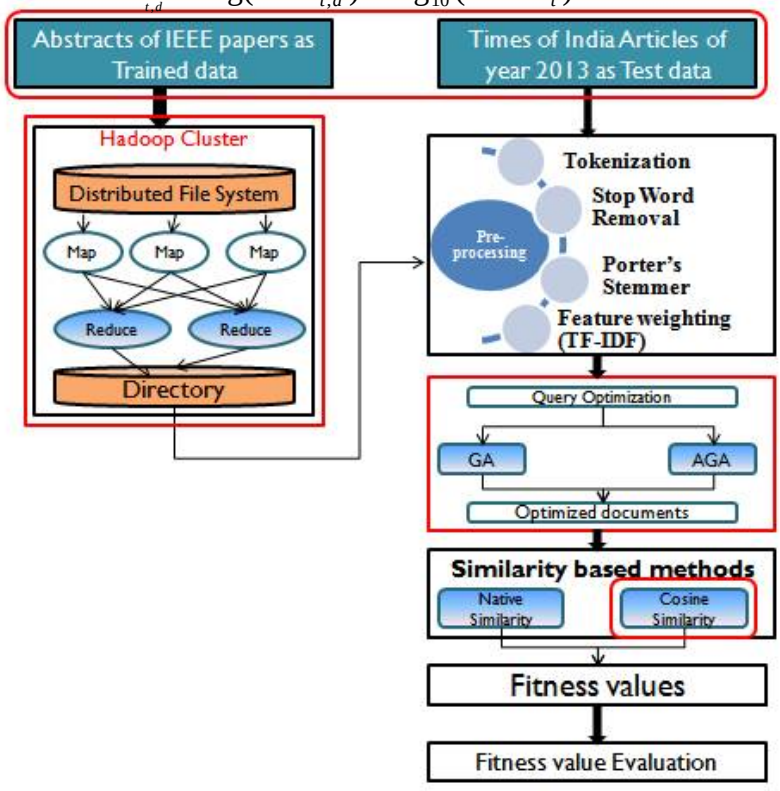

Fig. 1. Proposed System Architecture

Step 3: Fitness Evaluation:

Improvement in performance of IRS can be evaluated by comparing the result obtained by GA with AGA. Genetic Algorithms extends powerful search capability in a multidimensional space. In Information Retrieval many issues such as optimization and searching can be solved using techniques of Genetic Algorithm. This algorithm also helps to improve the performance of retrieval system [9]. Genetic algorithm is a combination of various computational models based on principles of evolution and natural selection. Actually this algorithm converts the problem in a specific domain into a model by using chromosomes using Selection, crossover and mutation operators [10]. Genetic algorithm was used to optimize the data nodes that are to find the optimal

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

data nodes to increase the efficiency which is the issue of Information retrieval system [11]. Steps included in GA are:

1) Representation of chromosomes.
2) Initialization of population
3) Genetic operators(Crossover and Mutation)
4) Repeat till criteria meets

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. The steps included in AGA are:

1) Representation of chromosomes.
2) Initialization of population
3) Control parameters(Probability of crossover and Probability of mutation)
4) Fitness function, eq. (2) (Cosine Similarity) [11, 12].

$$F = \frac{\sum_{k=1}^{t} d_{ik} * q_k}{\sqrt{\sum_{k=1}^{t} (d_{ik})^2 * \sum_{k=1}^{t} (q_k)^2}} \qquad \text{eq. (2)}$$

IV. MATHEMATICAL MODEL

Mathematical Model for proposed approach is as follows:
Let S, be the proposed system which can be represented as

$$S = \{\{I\}, \{F\}, \{O\}\}$$

Where,
I -> Input document collection (for Training and Testing)
F -> Functions used
O -> Matched document with its fitness value
Where,

$$F = \{f1, f2, f3, f4\}$$

f1 → Term Weighting (TF-IDF)
f2 → Optimization methods (AGA)
f3 → Similarity Based Methods (Cosine Similarity)
f4 → Evaluation Parameters (Precision, Recall and Accuracy)

V. SIMULATION RESULTS

The testing and training dataset used for the proposed algorithm is abstracts of IEEE papers from year 2010 to 2015.
Training dataset = 20 domains, 1303 documents (with HDFS)
Testing dataset = 5 domains, 251 documents (without HDFS)
Hadoop Architecture used for proposed algorithm is Single Node architecture that includes, Namenode, Datanode, TaskTracker, JobScheduler, ResourceManager.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

The details of dataset are as follows:

Sr. No	Domain	Training Dataset	Testing Dataset
1	Advanced Computing	94	90
2	Algorithms	100	100
3	Biometrics	90	61
4	Computer Graphics	88	
5	Distributed System	100	
6	Data Mining	100	
7	Ecommerce	100	
8	E-learning	100	
9	Embedded System & Software	100	
10	Genetic Algorithm	100	
11	Geographical Information System	64	
12	Image Processing	40	
13	Information Retrieval	55	
14	Mobile Computing	39	
15	Network Security	45	
16	Pattern Recognization	31	
17	Sensor Network	27	
18	Software testing	11	
19	System Software	12	
20	Web and Internet Computing	7	
	Total	**1303**	**251**

Table 1. Number of documents of each domain as Training and Testing Dataset

The performance of IRS is evaluated in terms of performance parameters that are Precision, Recall and Accuracy. Precision, Recall and Accuracy can be calculated with the help of confusion matrix [10].

	Document belonging to the Category	Document not belonging to the Category
Matched documents retrieved	True Positive (TP_i)	False Positive (FP_i)
Matched document rejected while retrieval	False Negative (FN_i)	True Negative (TN_i)

Table 2. Confusion Matrix

$$\text{Precision} = \frac{TPi}{\text{TPi+FPi}} \quad \text{eq. (3)}$$

$$\text{Recall} = \frac{TPi}{\text{TPi+FNi}} \quad \text{eq. (4)}$$

$$\text{Accuracy} = \frac{TPi+FNi}{\text{N}} \quad \text{eq. (5)}$$

The above table describes the performance parameters required to calculate the performance of IRS. The results carried out will be calculated on the basis of the conditions:

1) TP (True-Positive) is the number of matched documents retrieved correctly.
2) FP (False-Positive) is the number of matched documents retrieved incorrectly.
3) FN (False-Negative) is the number of unmatched documents not retrieved.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

4) TN (True-Negative) is the number of matched documents not retrieved.

The results obtained for both GA and AGA in distributed environments (HDFS) and with HDFS are shown below:

1) Performance of IRS for **Advance computing domain**
   a) With HDFS, **Time Required= 1.48 secs**
   b) Without HDFS, **Time Required= 58 secs**

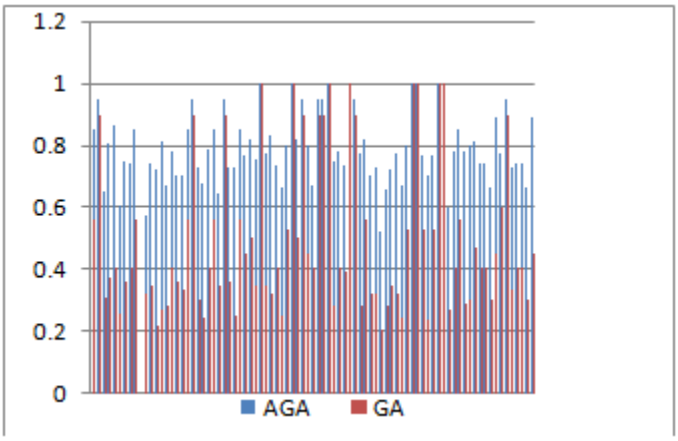

Fig 2. Performance of IRS using GA and AGA with HDFS

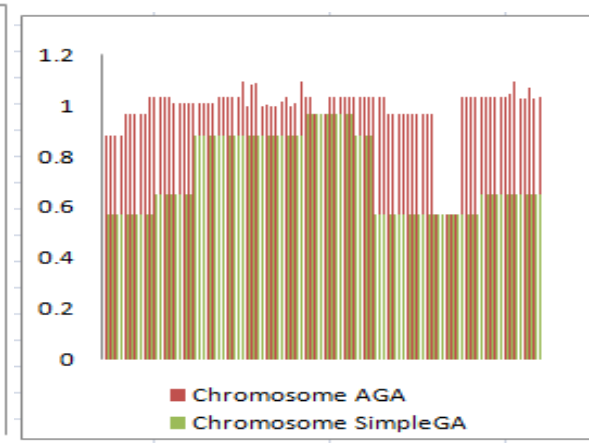

Fig 3. Performance of IRS using GA and AGA without HDFS

2) Performance of IRS for **Algorithms domain**
   a) With HDFS, **Time Required= 1.54 secs**
   b) Without HDFS, **Time Required= 1 min**

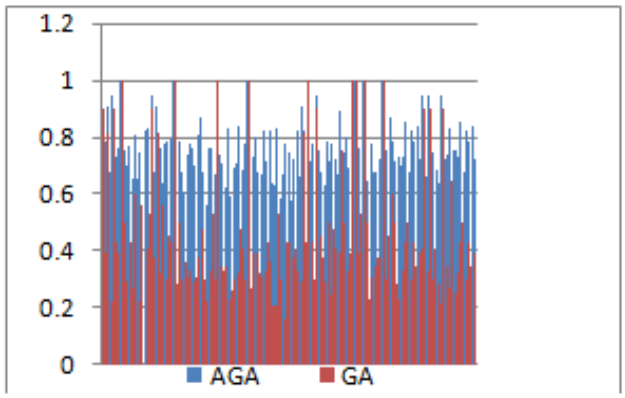

Fig 4. Performance of IRS using GA and AGA with HDFS

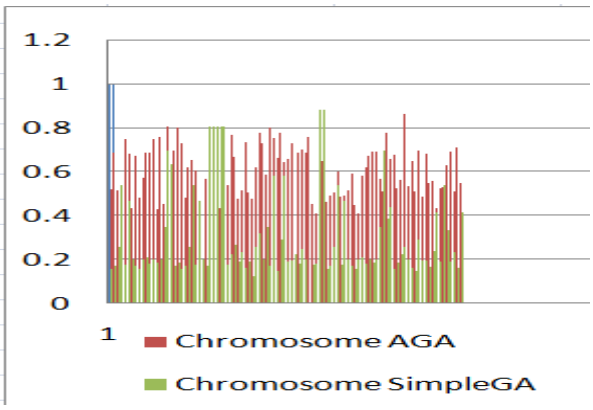

Fig 5. Performance of IRS using GA and AGA without HDFS

The above graph show the documents retrieved for two domains of dataset advanced computing and algorithms. The graph is showing the number of documents retrieved using GA and AGA using Hadoop environment and without Hadoop environment. The proposed algorithm is for evaluating the performance of IRS by comparing GA and AGA in Distributed environment, that with HDFS. The above graphs are to test the results of Testing and Training dataset results.

The Genetic Algorithm consists of the operators that are crossover and mutation. For the proposed algorithm the crossover and mutation are:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Sr.No	Algorithm	Crossover	Mutation
1	Genetic Algorithm	0.65	0.01
2	Adaptive Genetic Algorithm	1.0	0.5

Table 3. Genetic operators for GA and AGA

The performance of IRS within distributed environment has shown below in graph. The graph shows the improvement in performance of IRS using AGA as compare to GA. The effectiveness of each domain is good using AGA as compare to GA.

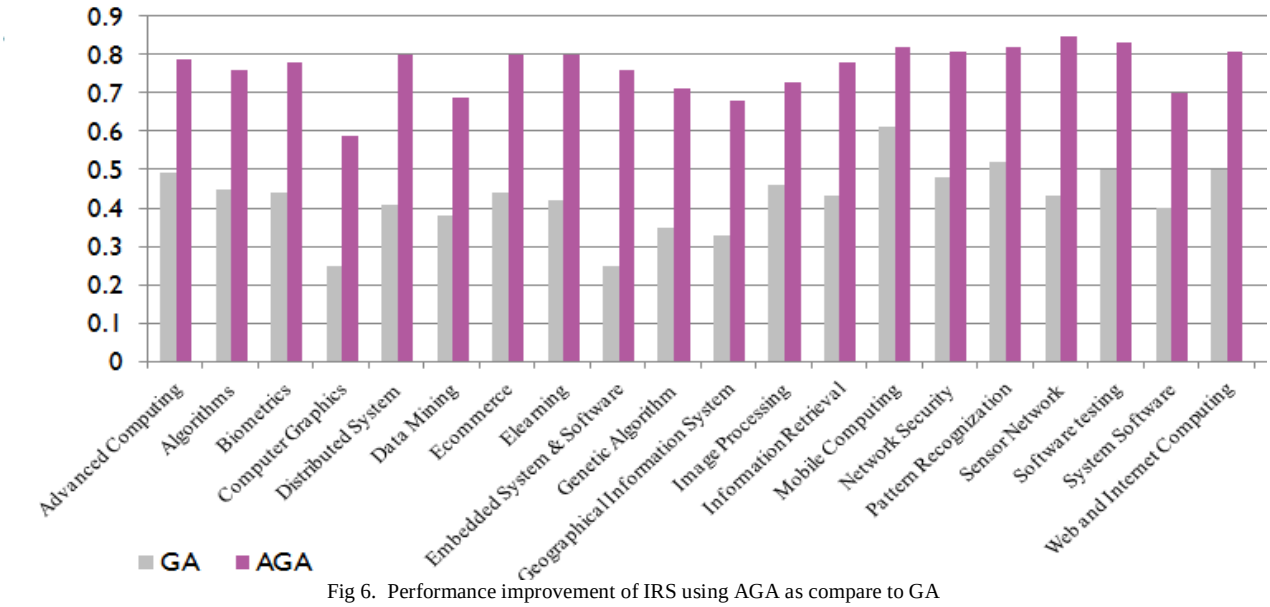

Fig 6. Performance improvement of IRS using AGA as compare to GA

The above graph shows the number of documents retrieved of each domain using GA and AGA.
The matching documents retrieval result of GA and AGA has shown below, in form of confusion matrix.

	Document belonging to the category		Document not belonging to the category	
	GA	AGA	GA	AGA
Matched Documents retrieved	1036	1153	147	81
Matched Documents rejected while retrieval	109	56	11	03

Table 4. Confusion Matrix of Information Retrieval System using GA and AGA

The above table shows the number of matched documents retrieved and rejected as well as documents retrieved and rejected that are belonging to category and not belonging to category. The performance of IRS is calculated in terms of Precision, Recall and Accuracy. The results obtained by proposed system are as follows:

Method Used	Precision	Recall	Accuracy
Genetic Algorithm	0.87	0.90	0.80
Adaptive Genetic Algorithm	0.93	0.95	**0.89**

Table 5. Performance improvement of IRS using AGA in Distributed Environment

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

The above table shows the result obtained in terms of precision, recall and accuracy. According to above table there is **a 9% improvement** in the performance of IRS using the proposed algorithm AGA.

VI. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better as compare to the existing algorithm. The performance of IRS has been improved by **9%** by using the proposed algorithm in Distributed Environment. The proposed algorithm provides effective as well as efficient result in terms of precision, recall and accuracy. The main advantage of efficiency in terms of time is also improved because of use HDFS for information retrieval. The performance of the proposed algorithm is analyzed on Single Node and multi-node Hadoop Cluster, and the variation of results can be obtained in terms of time where the better results obtained are in multi-node Hadoop cluster. In future more algorithms can be included in the proposed algorithm such as Ranking algorithms, so the again the performance of IRS can be improved and the user need to be meet. We have used 3 and 5 nodes Hadoop Cluster, as we have used less dataset as compare to the word Big Data. In future, we can increase the number of nodes, increases the dataset and can analyze the performance.

REFERENCES

1. Philomina Simon and S Siva Sathya, "Genetic Algorithm for Information Retrieval", IEEE, IAMA, 2009.
2. Venkata Sameer and Rakesh Balabantaray, "Improving ranking of webpage's using user behavior, a Genetic algorithm approach", IEEE, 2014.
3. A.S Siva Sathya, B Philomina Simon, "A Document Retrieval system with combination terms using genetic algorithm", IJCEE, vol. 2. No. 1, February, 2010.
4. Palson Kennedy and T V Gopal, "An Effective Optimized Genetic Algorithm for scalable information retrieval from cloud using Big Data", Journal of Computer Science., January 2014.
5. Wafa Maitah Al-Rababaa and Ghasun Kannan, "Improving the effectiveness of Information Retrieval system using Adaptive Genetic Algorithm", IJSCIT, 2013.
6. Harshawardhan Bhosale and Prof. Devendra Gadekar, "A Review paper on Big Data and Hadoop", IJSRP, Vol. 4 Issue 10, 2014.
7. Thanh Cuong Nguyen, Wenfeng Shen, Jiwei Jiang and Weimin Xu, "A Novel Data Encryption in HDFS", ICGCC, IEEE, pp.~2183-2187, 2013.
8. Richard McCreadie, Craig Macdonald and Ladh Ounis, "MapReduce indexing strategies: Studying scalability and efficiency", ELSEVIER,IPMIJ, 2010.
9. Prof. Anuradha Thakare and Dr. C A Dhote,"An Improved Matching Functions for Information Retrieval Using Genetic Algorithm", IEEE, ICACCI, 2013.
10. Ahmed Radwan, Bahgat Adbel Latef, Abdel Mgeid & Osman Sadek, "Using Genetic Algorithm to improve Information Retrieval Systems", ISSRI, vol 2 issue 5, 2008.
11. Moheb Ramzy Girgis, Abdelmgelid Amin Aly & Fatima Mohy Eldin Azzam, "The Effect of Similarity Measures on Genetic Algorithm-Based Information Retrieval", IJCSEITR, vol. 4, no. 5, 2014.
12. Prof. Anuradha Thakare and Dr. C A Dhote, "New Unification matching scheme for efficient information retrieval using Genetic Algorithm", IEEE, ICACCI, pp~ 1936-1941, 2014.

BIOGRAPHY

Prajakta Kantilal Mitkal, received the Bachelor's of Engineering degree (B.E.) in Computer Science and Engineering in 2010, from 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, Data Mining and Big Data.