Journal of Computer Science 10 (6): 1026-1035, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.1026.1035 Published Online 10 (6) 2014 (http://www.thescipub.com/jcs.toc)

AN EFFECTIVE OPTIMIZED GENETIC ALGORITHM FOR SCALABLE INFORMATION RETRIEVAL FROM CLOUD USING BIG DATA

[1]Palson Kennedy. R. and [2]T.V. Gopal

[1]Faculty of I and C, Anna University Chennai 25, India and Department of CSE, PMREC.Chennai-95.
[2]Department of CSE, Faculty of I and C, CEG, Anna University, Chennai-25, India

Received 2013-12-14; Revised 2013-12-19; Accepted 2014-01-31

ABSTRACT

The distributed computations are broadly used in the current world for processing large scale jobs. For data intensive applications with big data, it has recently received a very good attention. A simple programming model that allows easy development of scalable parallel applications to process big data on large clusters was required. In our proposed work the input files will be subjected to load balancing. In load balancing process the files will be separated and are stored in the clouds. Load balancing is done to handle the big data. Then the stored files will be subjected to map reduce process. In mapping process the files are mapped and a key value will be assigned to the files and then the files are reduced. The map reduce process is to be done by assigning mappers and reducers to the cloud servers. After the mapreduce process the files will be optimized using genetic algorithm. If the node data size increases the efficiency reduces, for increasing the efficiency we have optimized the node data size using genetic algorithm. The experimental results will show the enhance in the node of the data size has done efficiently and the overall efficiency increased to considerable level with the node increments. The proposed method implemented using JAVA.

Keywords: MapReduce, Optimization, Load Balancing, Genetic Algorithm, Distibuted Computations

1. INTRODUCTION

The IRS is responsible for collecting taxes and the interpretation and enforcement of the Internal Revenue Code in USA. Ontological classification of unstructured data is critically important in the managing of initiatives, programs and strategies. Internal Revenue service has been attempting to ratchet up its enforcement efforts in the last few years. While this is no surprise given the spate of abusive tax shelter transactions peddled by lawyers and accountants in late 1990's, it should strike a measure of fear into the heart of the general taxpaying public (Lavoie, 2008) When the specialized requirements of the IRS and their requirements to stay in compliance to government regulations are included, the complexity of their roles and the need for accuracy, auditability and transparency are critically important (Kennedy, 2010). The increased importance of Business intelligence infrastructures reflects three interacting trends: More turbulent, global business environments, additional pressures to unveil valid risk and performance indicators to stakeholders and aggravated challenges of effectively managing the more and more densely interwoven processes (Baars and Kemper, 2008) for years, critics had claimed that IRS managers aggressively required their subordinates to maximize the taxes they extracted from citizens and to intensify the enforcement actions to do so (Riccucci *et al*., 2006).

Information retrieval technology has been central to the success of the Web. Web based indexing and search systems such as Google and Yahoo have profoundly changed the way we access information (Finin *et al*., 2005). It is the method of searching information in documents, documents themselves or metadata that

Corresponding Author: Palson Kennedy, R., Faculty of I and C, Anna University Chennai 25, India and Department of CSE, PMREC.Chennai-95

describes these documents. This can be a search in the local database or in the Internet for text, images, sound, or data. The retrieval formats should be flexible and should provide users to manipulate the search process and results by retrieving search history, adjusting search strategies, editing and sorting search results and choosing preferable delivery formats (Naik and Rao, 2011). Many retrieval systems today reference documents from disparate sources or collections. These sources may reside on the same computer as the retrieval system; some of them may be dispersed over a local or wide area network, whereas others may reside at various Internet locations (Losee and Church, 2004). Scaling-down a distributed IR system will maintain the behavior of the whole system and, at the same time, the computer requirements will be softened. This allows the use of virtualization tools to build a large-scale distributed system using just a small cluster of computers (Cacheda *et al*., 2010).

An integrated approach for searching and browsing in heterogeneous peer-to-peer networks" a Service Oriented Architecture (SOA) for supporting searching and browsing in a hierarchical network. It is characterized by an high extensibility and reusability. More than proposing new algorithms, their paper proposes a very good architecture exploiting in a highly optimized fashion state-of-the-art technology in both P2P and IR (Baraglia *et al*., 2005). The information retrieval problem is a more complex operation than traditional search techniques based on object identity or filenames, currently being used in P2P systems. The Information Retrieval (IR) community has over the years developed algorithms for precise document retrieval in static data environments (Zeinalipour-Yazti *et al*., 2004). The network bandwidth in a P2P network is fully utilized. As nodes in the network are interconnected, there is no single point of failure. P2P networks are characterized by high processing power and storage without the overhead of high cost hardware (Renuga and Sudhasadasivam, 2005). Present day summarization technologies fall short of delivering fully informative summaries of documents. Largely, this is due to shortcomings of the state-of-the-art in natural language processing (Boguraev *et al*., 1998).

The sensor network is a collection of small-size, low-power, low-cost sensor nodes that have some computation, communication, storage and even movement capabilities. These nodes can operate unattended, sensing the environment, generating data, processing data and providing the data to users (Subramanian *et al*., 2007). Scalable management and self-organizational capabilities are emerging as central requirements for a generation of large-scale, highly dynamic, distributed applications. An entirely new distributed information management system called Astrolabe. Astrolabe collects large-scale system state; permitting rapid updates and providing on-the-fly attribute aggregation. Large-scale web and text retrieval systems deal with amounts of data that greatly exceed the capacity of any single machine. To handle the necessary data volumes and query throughput rates, parallel systems are used, in which the document and index data are split across tightly-clustered distributed computing systems (Moffat *et al*., 2006). When sampled documents are long, storage costs can be large; a method of pruning long documents is used to reduce storage costs. It demonstrate that the building resource descriptions and centralized sample databases from the pruned contents of sampled documents can reduce storage costs by 54-93% while causing only minor losses in the accuracy of distributed information retrieval (Lu and Callan, 2002).

2. RELATED WORK

A handful of researches have been presented in the literature for unstructured data analysis using various methods. Recently, the use of public networks and the deployment of powerful personal computing units by end users have brought a shift from the traditional Client-Server computing model to the Peer-to-Peer (P2P) computing model.: Analyses have received a great deal of attention among researchers. A brief review of some recent researches is presented here.

Subramanian *et al*. (2007) proposed three schemes for securing distributed data storage and retrieval in sensor networks. All the schemes have the following properties: (i) only authorized entities can access data stored in the sensor network; (ii) the schemes were resilient to a large number of sensor node compromises. The second and the third schemes do not involve any centralized entity except for a few initialization or renewal operations and thus support secure, distributed data storage and retrieval. The third scheme further provides high scalability and flexibility and hence it was most suitable in real applications. The effectiveness and efficiency of the proposed schemes have also been verified through extensive analysis and TOSSIM-based simulations.

Baars and Kemper (2008) suggested that the evolution of management support towards corporate wide Business Intelligence infrastructures, the integration of components for handling unstructured data comes into focus. The study contains three types of approaches for tackling the respective challenges were distinguished. The approaches were mapped to a three layer BI framework and discussed regarding challenges and business potential. The application of the framework was exemplified for the domains of Competitive Intelligence and Customer Relationship Management.

The scientific challenge of today was to advance our understanding of how economic value was created through innovation and knowledge appropriation. New data on innovation and knowledge appropriation were needed to represent modern business activity and to guide policy makers in the 21st century economy. Cyber-enabled transaction data grounded in theoretically driven micro-level measures of innovation within organizations offer expanded potential for scientists to meet those needs.

Chaudhuri *et al*. (2013) presented a domain independent platform for data cleaning developed as part of the Data Cleaning project at Microsoft Research. Their platform consists of a set of core primitives and design tools that allow a programmer to develop sophisticated data cleaning solutions with minimal programming effort. Their primitives were designed to allow rich domain and application specific customizations and can efficiently handle large inputs. Their data cleaning technology have significant impact on Microsoft products and services and it have been successfully used in several real-world data cleaning applications.

Naik and Rao (2011) indicated that the a digital library comprises diverse collections of digital objects representing text, sound, maps, videos, photos and a working environment, technology and services. The main objective of any Digital Library (DL) was to fulfill the needs of its users. A general problem for a user was information search and retrieval in the Internet world. Their paper discusses the information search and retrieval system its models and uses in digital libraries.

Chotayakul *et al*. (2013) are designed as a multi-echelon inventory problem with single-item capacitated lot-sizing to minimize total costs of running ATM network. In this study, they formulate the problem as a Mixed Integer Program (MIP) and develop an approach based on reformulating the model as a shortest path formulation for finding a near-optimal solution of the problem.

Siham, *et al*. (2013) present passive clustering mechanisms and the main clustering protocols proposed for wireless sensor networks; they introduce a new protocol designated for mobile nodes in wireless sensor network that is based on the APC-T. This mechanism provides the stability of clusters after each departs of cluster-head and allows balanced energy consumption among the sensor nodes. Comparison with the existing schemes such as APC-T and Geographically Repulsive Insomnious Distributed Sensors (GRIDS) proves that the mechanism for selecting a backup of cluster-head nodes, which is the most important factor influencing the clustering performance, can significantly improves the network lifetime.

3. MOTIVATION OF THE RESEARCH

The natural tendency to apply reductionism to the area of unstructured content analysis needs to be countered with a holistic foundation within which the entire ecosystem of unstructured content within an organization can be defined. Ecosystems are by nature more oriented towards reciprocity and a continual rejuvenation of content and are therefore holistic in nature. Unstructured content analysis emanating from these areas of development lack shared structured data schema which are critical for the development of a holistic ecosystem. In ecosystem XML parsing and reliance on XSLT style sheet definitions are providing taxonomy-based personalization for structured content yet is unproven for unstructured content use. A TinyOS mote Simulator (TOSSIM) is used to ease the development of sensor network applications. It then captures the data stored in the sensor nodes; if the data are encrypted, it will attempt to figure out the key and thus decrypt the data. Due to the lack of physical protection for sensor nodes, the compromise of sensor nodes and the capturing of data from the compromised sensor nodes cannot be fully prevented.

Initial experience with the Astrolabe aggregation mechanisms demonstrates that the system is extremely powerful despite its limits. Managers of an application might use the technology to monitor and control a distributed application using aggregates that summarize the overall state within the network as a whole and also within the domains (scopes) of which it is composed. The disadvantage is that this strategy does rely on clocks being approximately synchronized. The trade-off can be made by the applications. Digital libraries rely on effective retrieval methods with easy access to the information. Thus, the success of digital libraries is depends on the quality of retrieval. Accordingly, research in the IR has traditionally been important in the research pertaining to the digital libraries. Some difficulties in this context are the variable IR methods and interfaces like representation and relevance ranking, as well as the scope of the search results. Another difficulty is the merging of the individual results and the inter-system ranking.

Boolean approach is used to implement and it is computationally efficient for the current large scale and operational retrieval systems. It enables users to express structural and conceptual constraints to describe important linguistic features. Users find that synonym specifications and phrases are useful in the formulation of queries. The disadvantage is the users find it difficult to construct effective Boolean queries for several reasons. Users are using the natural language terms AND, OR or NOT that have a different meaning when used in a query. Thus, users will make errors when they form a Boolean query, because they resort to their knowledge of English. It is to improve the efficiency of the information retrieval system by using peer to peer techniques. Peer-to-peer networks have become popular of late for a range of applications, including the well known file sharing systems such as Gnutella and Bit Torrent. It has the tendency to detect the queries in the detected files.

4. AN UNSTRUCTURED DATA ANALYSIS AND CLASSIFICATION SYSTEM BASED ON AN EFFECTIVE OPTIMIZATION ALGORITHM

In our proposed work the input files will be subjected to load balancing.The Basic Architecture of our proposed methodology in shown in **Fig. 1**. In load balancing process the files(unstructured data) will be separated and are stored in the clouds. Load balancing is done to handle the big data. Then the stored files will be subjected to map reduce process. In mapping process the files are mapped and a key value will be assigned to the files and then the files are reduced. The map reduce process is to be done by assigning mappers and reducers to the cloud servers. After the mapreduce process the files will be optimized using genetic algorithm. If the node data size increases the efficiency reduces, for increasing the efficiency we have optimized the node data size using genetic algorithm. The experimental results will show the increase in the node of the data size has done efficiently and the overall efficiency increased with the node increments.

4.1. Load Balancing Technique

In the load balancing process the files are separated and are stored in the cloud servers. There are five services considered.1.Server-id, 2.Node-id, 3.Processing time, 4.Memory rate, 5.CPU rate. The process of genetic algorithm is given in **Fig. 4**.

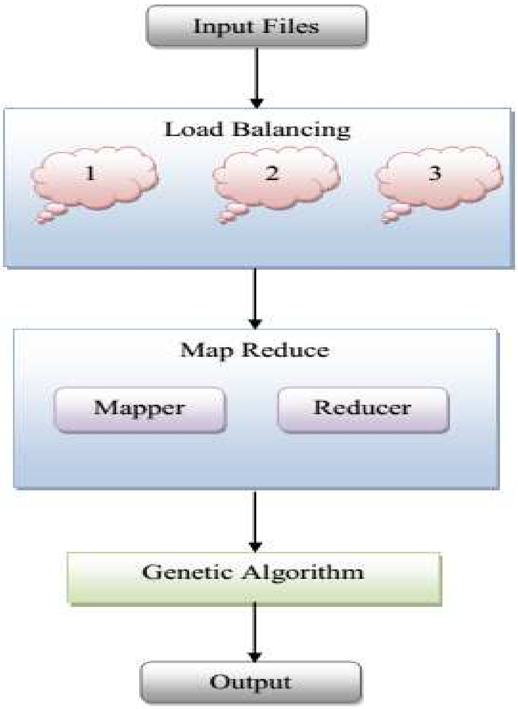


Fig. 1. Architecture for our proposed methodology

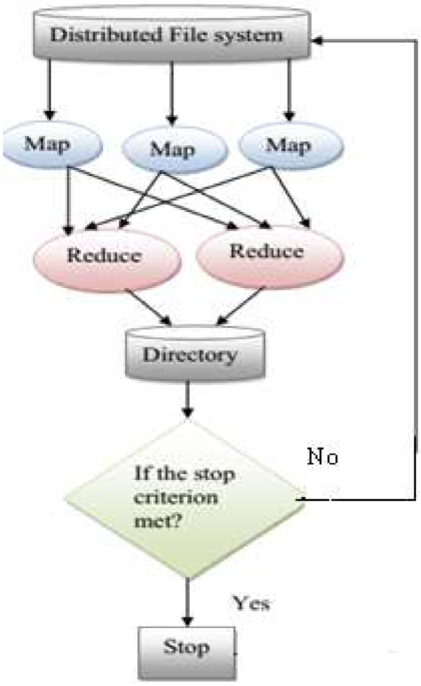


Fig. 2. MapReduce architecture

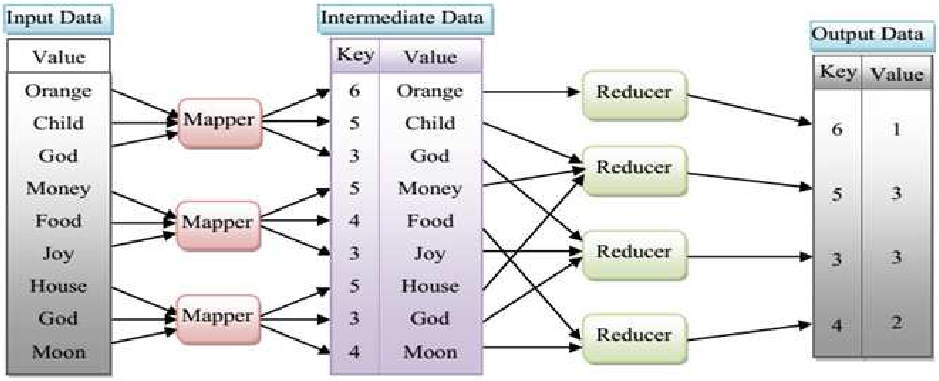


Fig. 3. Example for the MapReduce model

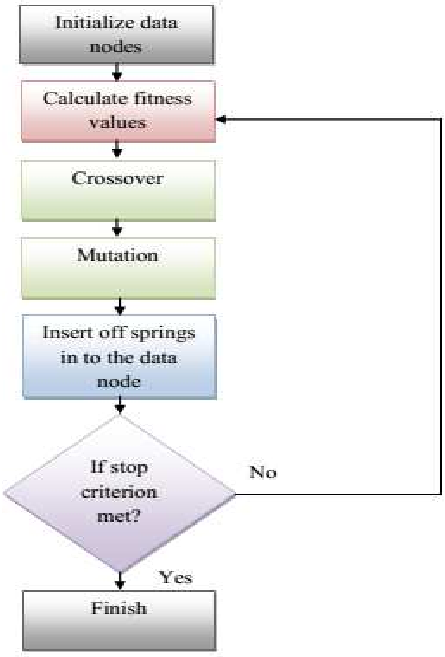


Fig. 4. Process steps in genetic algorithm

1. Server-id

It refers to the identification that was assigned to the server.

2. Node-id

It refers to the identification that was assigned to the node.

3. Processing Time

The time required for the processing of the input.

4. Memory Rate

Memory rate refers to the memory usage. Here it is varied from 100 to 10000.

5. CPU Rate

CPU rate refers to the CPU usage. Here it is varied from 100 to 10000.

The services considered will be assigned a service-id. For retrieving the files from the server the user have to provide the service-id as request. After the load balancing process map reduce is done.

4.2. Map Reduce Algorithm

The basic Map reduce Architecture shwon in **Fig. 2** and the Model of the same is given in **Fig. 3**. MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm. A MapReduce program is composed of Map()-performs filtering and sorting Reduce()-performs a summary operation.

Mapping

The master node takes the input, divides it into smaller sub-problems and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem and passes the answer back to its master node.

Reducing

The master node then collects the answers to all the sub-problems and combines them in some way to form the output.

Mappers and reducers are assigned separately to the servers in the cloud. The map reduce operations were carried out separately and then thee nodes were optimized using genetic algorithm.

4.3. Optimization of Data Nodes Using Genetic Algorithm

Genetic algorithm is a family of computational models based on principles of evolution and natural selection. These algorithms convert the problem in a specific domain into a model by using a chromosome-like data structure and evolve the chromosomes using selection, crossover and mutation operators. Genetic algorithm was used to optimize the data nodes. That is to find the optimal data nodes to increase the efficiency.

4.4. Selection Process in Genetic Algorithm

Selection is the first stage in genetic algorithm. Selection is process in which the association rules are chosen for later processing (crossover) from the population. A selection procedure is given as follows: For the data nodes the fitness function is calculated using the below formula Equation (1):

$$f\left(x_1, x_2, \ldots\ldots\ldots, x_n\right) = \sum_{i=1}^{n} x_i^2 \quad (1)$$

where, $x_{1,}\, x_2, \ldots\ldots\ldots\ldots, x_n$ = data nodes.

The above flow chart in Fig 4 represents the process that was done in the genetic algorithm. The first step is the initialization of genes (data nodes). The genes are assigned to the chromosomes based on our requirement. Then the fitness values are calculated in order to select the genes. With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents. With a mutation probability mutate new offspring at each locus (position in chromosome). After that the mutation is done. Then the offspring is produced and it is inserted in to the population. If the criteria is met then the process is stopped, otherwise the process is repeated.

4.5. Crossover in Genetic Algorithm

It is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next. It is analogous to reproduction and biological crossover, upon which genetic algorithms are based. Cross over is a process of taking more than one parent solutions and producing a child solution from them Equation (2). The best solutions were selected using the fitness function calculation. Then the crossover rate is assigned. Crossover rate should be selected more than 5:

$$\text{Crossover points} = \text{Crossover rate} \times \text{Length of the chromosomes} \quad (2)$$

where, Length of the chromosomes = Number of genes = 10.

4.6. Mutation in Genetic Algorithm

Mutation is the final stage in genetic algorithm. Mutation is a genetic operator which is used to maintain genetic diversity from one generation of a population of algorithm chromosomes to the next. A genetic operator is an operator used in genetic algorithms to maintain genetic diversity. Then the Euclidean distance between the services were calculated. During the computation, the distances among vectors can be measured by Equation (3):

Euclid Distance which is expressed as:

$$d(x, y) = \|x - y\| = \sqrt{\sum\nolimits_{i=1}^{n} (xi - yi)_2} \quad (3)$$

Based on the fitness values non optimal nodes are eliminated. Thus the optimal data nodes are identified using a threshold value. Thus the input data was analyzed using the effective optimization algorithm.

5. RESULTS

In this section we have given the results of our proposed work and have analyzed their performance. The implementation is done in JAVA for unstructured data analysis using NETBEANS version 1.7.2 (jdk 1.7). **Table 1** Shows the samples values of the services used in our work. **Table 2** shows the fitness values calculated for services using Euclidean distances.

Figure 5 shows the input for the proposed methodology. There were three sections named as initial scheduling, map reduced output and scheduling based on GA. **Figure 6** shows the request size allocation for CPU usage and memory usage. **Figure 7** shows the initial scheduling process. **Figure 8** shows the map reduce output. **Figure 9** shows the final output based on genetic algorithm.

Table 1. Sample values for services used

Server ID	Node ID	Processing time	Memory rate	CPU rate
1	3	1	11	7
2	7	1	6	9
2	9	6	6	7
3	1	5	7	1
1	4	5	1	1

Table 2. Shows the euclidean distance calculated between the services used

Server id	Node id	Processing time	Memory rate	CPU rate	Server id	Node id	Processing time	Memory rate	CPU rate
2	4	1	8	10	2	7	2	13	2
1	4	4	21	6	2	5	3	3	7
1	5	1	3	11	3	5	6	2	20
3	2	6	6	8	1	3	7	12	12
2	3	6	12	8	2	4	3	6	8
2	9	3	2	5	1	4	7	4	3
1	6	6	5	1	1	6	8	9	6
2	6	6	4	6	3	8	7	11	10
3	6	5	22	24	2	8	7	3	10
Fitness value = 27.748438335462886					Fitness value = 43.515584764523496				

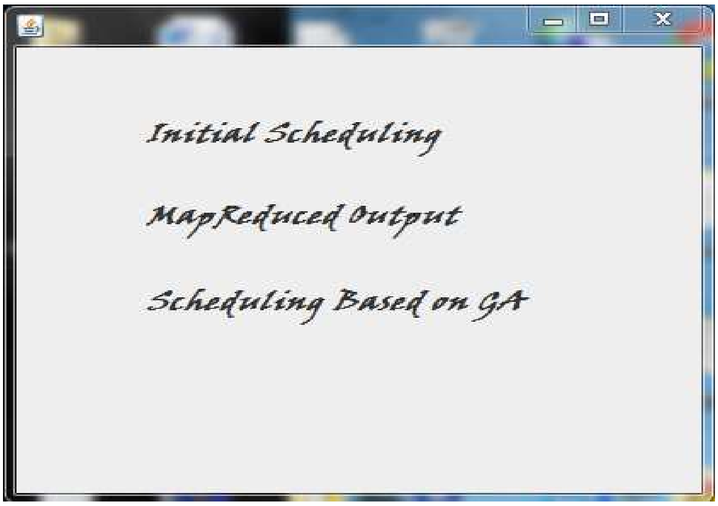
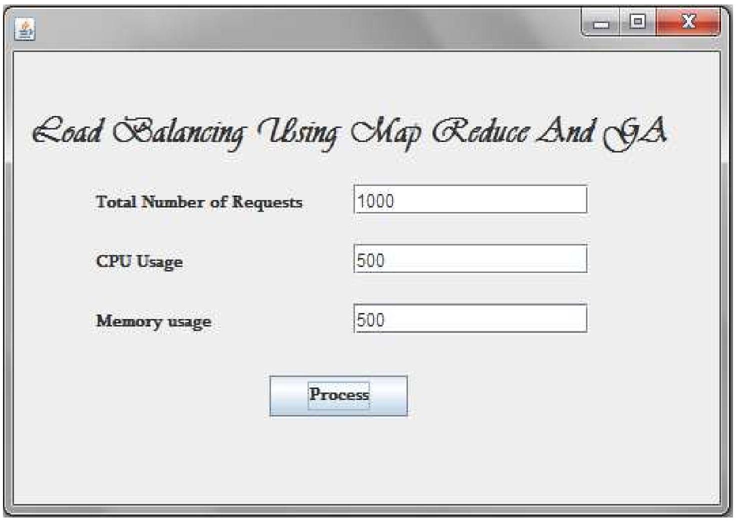


Fig. 5. Input for the proposed methodology

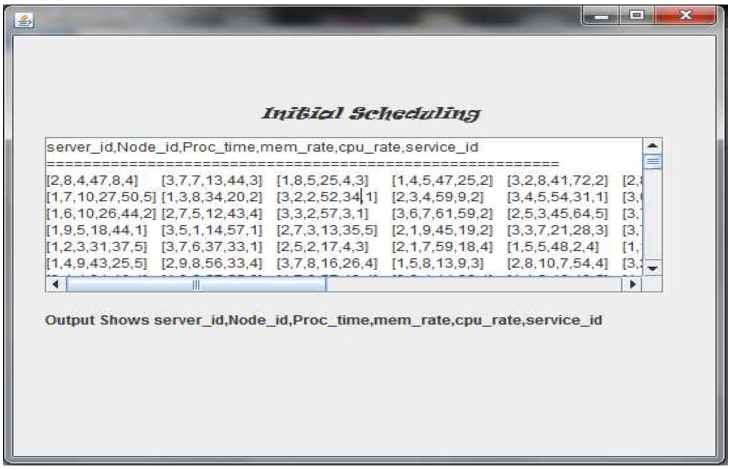

Load Balancing Using Map Reduce And GA

Total Number of Requests: 1000

CPU Usage: 500

Memory usage: 500

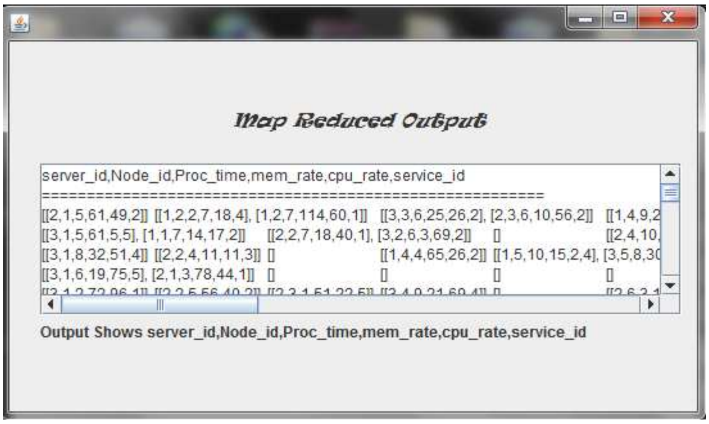
Fig. 6. Screenshot for request size allocation



```
server_id,Node_id,Proc_time,mem_rate,cpu_rate,service_id
========================================================
[2,8,4,47,8,4]     [3,7,7,13,44,3]   [1,8,5,25,4,3]    [1,4,5,47,25,2]   [3,2,8,41,72,2]   [2,
[1,7,10,27,50,5] [1,3,8,34,20,2]   [3,2,2,52,34,1]   [2,3,4,59,9,2]    [3,4,5,54,31,1]   [3,
[1,6,10,26,44,2] [2,7,5,12,43,4]   [3,3,2,57,3,1]    [3,6,7,61,59,2]   [2,5,3,45,64,5]   [3,
[1,9,5,18,44,1]    [3,5,1,14,57,1]   [2,7,3,13,35,5]   [2,1,9,45,19,2]   [3,3,7,21,28,3]   [3,
[1,2,3,31,37,5]    [3,7,6,37,33,1]   [2,5,2,17,4,3]    [2,1,7,59,18,4]   [1,5,5,48,2,4]    [1,
[1,4,9,43,25,5]    [2,9,8,56,33,4]   [3,7,8,16,26,4]   [1,5,8,13,9,3]    [2,8,10,7,54,4]   [3,
```

Output Shows server_id,Node_id,Proc_time,mem_rate,cpu_rate,service_id

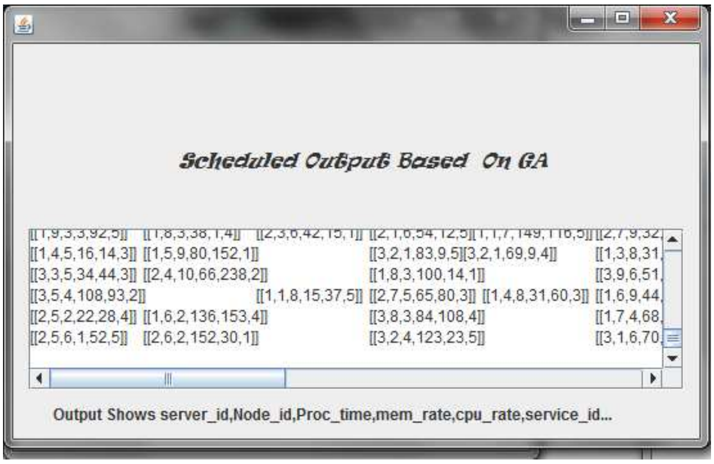
Fig. 7. Initial scheduling process



```
server_id,Node_id,Proc_time,mem_rate,cpu_rate,service_id
========================================================
[[2,1,5,61,49,2]] [[1,2,2,7,18,4], [1,2,7,114,60,1]]  [[3,3,6,25,26,2], [2,3,6,10,56,2]]   [[1,4,9,2
[[3,1,5,61,5,5], [1,1,7,14,17,2]]     [[2,2,7,18,40,1], [3,2,6,3,69,2]]     []                [[2,4,10,
[[3,1,8,32,51,4]] [[2,2,4,11,11,3]] []                 [[1,4,4,65,26,2]] [[1,5,10,15,2,4], [3,5,8,30
[[3,1,6,19,75,5], [2,1,3,78,44,1]]  []                 []                 []                []
```

Output Shows server_id,Node_id,Proc_time,mem_rate,cpu_rate,service_id

Fig. 8. Output for mapreduce process



```
[[1,9,3,3,92,5]]   [[1,8,3,38,1,4]]   [[2,3,6,42,15,1]] [[2,1,6,54,12,5][1,1,7,149,116,5]] [[2,7,9,32,
[[1,4,5,16,14,3]] [[1,5,9,80,152,1]]                     [[3,2,1,83,9,5][3,2,1,69,9,4]]         [[1,3,8,31,
[[3,3,5,34,44,3]] [[2,4,10,66,238,2]]                    [[1,8,3,100,14,1]]                     [[3,9,6,51,
[[3,5,4,108,93,2]]                    [[1,1,8,15,37,5]] [[2,7,5,65,80,3]] [[1,4,8,31,60,3]] [[1,6,9,44,
[[2,5,2,22,28,4]] [[1,6,2,136,153,4]]                    [[3,8,3,84,108,4]]                     [[1,7,4,68,
[[2,5,6,1,52,5]]   [[2,6,2,152,30,1]]                    [[3,2,4,123,23,5]]                     [[3,1,6,70,
```

Output Shows server_id,Node_id,Proc_time,mem_rate,cpu_rate,service_id...

Fig. 9. Output for the proposed methodology

6 DISCUSSION

Figure 10 shows the plot for execution time based on the request size. **Figure 11** shows the plot for fitness values based on the request size. **Figure 12** shows the plot for convergence fitness based on the request size. The comparison graphs plotted clearly shows that our proposed method is efficient.

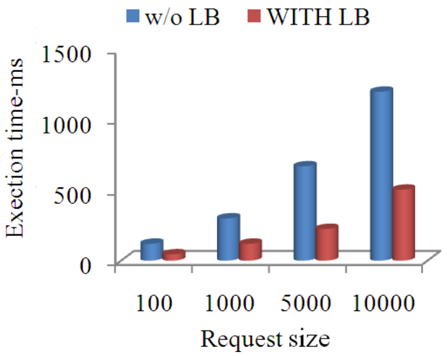


Fig. 10. Plot for comparison of R size and E time

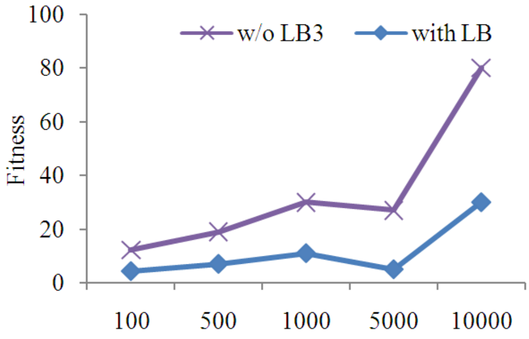


Fig. 11. Plot for comparison of R size and F values

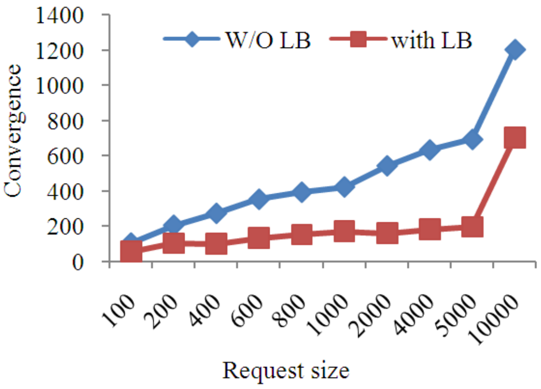


Fig. 12. Plot for comparison of request size and convergence fitness

The algorithm is working efficiently and the convergence is good as long as the request size below 5000. Therefore the maximum Confidence Interval is 5000.∓ 0.5.

7. CONCLUSION

In this paper, we have proposed map reduce algorithm with the support of load balancing and genetic algorithm. The proposed system was implemented and a set of files were utilized to analyze the outcomes of the proposed big data analysis method. The experimental results showed that our proposed method was efficient. This study presents a genetic algorithm based load balancing algorithm for Map Reduce environments in support of data intensive distributed applications. Simulation results have shown the effectiveness of the algorithm in balancing workload among Map Reduce nodes. The algorithm speeds up the computation process of SVD and maintains the high level of accuracy in information retrieval. Even though the experimental and simulation results the algorithm shows satisfied performance, it is clear that still a variety of opportunities available, for example: Determining the best fitness value of rank k that is used in GA can be investigated further to gain the most efficient computation. The experimental code shows certain level of accuracy, it can be improved further by using a better model. In the future this map reduce framework can be extended using different optimization algorithms for better results. The execution time can be further reduced than our proposed method.

8. REFERENCES

Baars, H. and H.G. Kemper, 2008. Management support with structured and unstructured data-an integrated business intelligence framework. Inform. Syst. Manag., 25: 132-148. DOI: 10.1080/10580530801941058

Baraglia, R., D. Laforenza and F. Silvestri, 2005. The SIGIR heterogeneous and distributed information retrieval workshop. ACM SIGIR Forum, 39: 19-24. DOI: 10.1145/1113343.1113348

Boguraev, B.,C. Kennedy, R.B. Uamy, S. Brawer and Y.Y. Wong *et al*., 1998. Dynamic presentation of document content for rapid on-line skimming. Comput. Inform. Sci., pp: 109-118.

Cacheda, F., V. Carneiro, D. Fernández and V. Formoso, 2010. Performance evaluation of large-scale Information Retrieval systems scaling down. Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval, (SIR' 10).

Chaudhuri, S., Z. Chen, K. Ganjam, R. Kaushik and V.R. Narasayya, 2013. Towards a domain independent platform for data cleaning. IEEE Data Eng., 34: 43-50.

Chotayakul, S., P. Charnsetthikul, J. Pichitlamken and J. Kobza, 2013. An optimization-based heuristic for a capacitated lot-sizing model in an automated teller machines network. J. Math. Stat., 9:pp.283-288. DOI: 10.3844/jmssp.2013.283.288

Finin, T., J. Mayfield, A. Joshi, R.S. Cost and C. Fink, 2005. Information retrieval and the semantic web. Proceedings of the 38th International Conference on System Sciences, Jan. 3-6, IEEE Xplore Press, pp: 113a-113a. DOI: 10.1109/HICSS.2005.319

Kennedy, R.P., 2010. Unstructured content analysis and classification system for the IRS. Int. J. Comput. Applic., 1: 32-37. DOI: 10.5120/105-216

Lavoie, R., 2008. Analyzing the schizoid agency: Achieving the proper balance in enforcing the internal revenue code. Akron Tax J.

Losee, R.M. and L. Church, 2004. Information retrieval with distributed databases: Analytic models of performance. IEEE Trans. Parallel Distrib. Syst., 14: 18-27. DOI: 10.1109/TPDS.2004.1264782

Lu, J. and J.P. Callan, 2012. Pruning long documents for distributed information retrieval. Proceedings of the 11th International Conference on Information and Knowledge Management, Nov. 04-09, ACM New York, NY, USA., pp: 332-339. DOI: 10.1145/584792.584847

Moffat, A., W. Webber and J. Zobel, 2006. Load balancing for term-distributed parallel retrieval. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 06-10, ACM New York, NY, USA., pp: 348-355. DOI: 10.1145/1148170.1148232

Naik, N.R. and A.M. Rao, 2012. Information search and retrieval system in libraries. Proceedigns of the 8th International Caliber held in Goa University, (CGU' 11), Goa, pp: 16-27.

Renuga, A.R. and R.G. Sudhasadasivam, 2005. P2P information retrival framework for digital library system. J. Theoretical Applied Inform. Technol., 5: 301-306.

Riccucci, N.M. H.G. Rainey and J. Thompson, 2006. Leadership and the transformation of a major institution: Charles rossotti and the internal revenue service. Public Administrat. Rev., 66: 596-604.DOI:10.1111/j.1540-6210.2006.00619.x

Siham, A., M. Abdelillah and E.G. Driss, 2013. Advanced energy efficient passive clustering mobility in wireless sensor networks. Am. J. Applied Sci., 10: 1558- 1569. DOI: 10.3844/ajassp.2013.1558.1569

Subramanian, N., C. Yang and W. Zhang, 2007. Securing distributed data storage and retrieval in sensor networks. Proceedings of the 5th Annual IEEE International Conference on Pervasive Computing and Communications, Mar. 19-23, IEEE Xplore Press, White Plains, NY, pp: 191-200. DOI: 10.1109/PERCOM.2007.29

Zeinalipour-Yazti, D., V. Kalogeraki and D. Gunopulos, 2004. Information retrieval in peer-to-peer networks. J. Comput. Sci. Eng., 6: 20-26. DOI: 10.1109/MCSE.2004.12