A Framework for the evolutionary clustering of dynamic heterogeneous information network based on Hadoop MapReduce

L. Visalatchi[1], M. Balamurugan[2]
[1]Associate professor,
Dr. Umayal Ramanathan College for Women, Karaikudi, Tamil Nadu, India,
visaramki@gmail.com
[2]Associate professor,School of Computer Science, Engineering and Applications,
Bharathidasan University, Tiruchy, Tamil Nadu, India,
mbala@bdu.ac.in, mmbalmurugan@gmail.com

Abstract - Evolutionary **clustering is a way of dynamic clustering which takes the time period into account which is the most suitable one for clustering the dynamic social networks and information networks. Comparing with the traditional static clustering, evolutionary clustering will provide the information instantly considering the time stamps, which will be helpful to diagnose the entire dynamic network and to retrieve the data as per the user need. The dynamic networks are huge in size which makes it difficult and time consuming to cluster the data whether it may be supervised or unsupervised clustering. In this paper, we have proposed a framework to overcome this difficulty by using hadoop map reduce. This framework combines the evolutionary algorithm and the mapreduce frame work. The map reduce module process the input and splits the data into fragments, and reduces the data size. The evolutionary algorithm, groups the data, satisfying the given criteria in parallel. The map reduce model benefits the clustering technique not only in reducing the size of data and storage but also in processing the data simultaneously among the cluster of computers which facilitates in producing the output even faster.**

Keywords: Evolutionary clustering, heterogeneous network, hadoop, mapreduce

I. INTRODUCTION

A network that consists of multiple type of objects are known as heterogeneous network. Internet, Research Collaboration networks, Biological networks are some of the information networks containing multi-type objects. Various analytical techniques have been proposed for better understanding of heterogeneous networks and their evolving properties. Evolutionary clustering [1] proposed by Ravikumar et.al is a technique which includes the time stamp of the data to be clustered dynamically, therefore producing dynamic clusters. The significant advantage of evolutionary clustering is cluster correspondence across the time periods which will help us in analyzing the characteristics of an information network qualitatively. Since the size of the information network is huge, the process of clustering takes more time. In this paper, we propose a frame work to reduce the time consumption [14][15] of evolutionary clustering process using Hadoop map reduce[11][12[13]. The experiments were performed and the results were analyzed to verify the reduced time consumption.

II. RELATED WORK

Several algorithms have been proposed for evolutionary clustering; we reviewed the algorithms such as Evolutionary spectral clustering by incorporating temporal smoothness [2][3], Evolutionary clustering and analysis of bibliographic networks[8][9], Evolutionary patterns of themes in the text stream[4][5][10], Particle and density based evolutionary clustering method[6][7], based on the issues of handling huge data and time consumption. Most of these algorithms encounter a drawback, that is, as the data size increases, the search space for each time period is getting increased[20] and the parameters for temporal smoothness need to be changed. The execution time is increased and it becomes insignificant for them to handle dynamic variation[21] of objects. In this study, we adopt the hadoop map reduce technique for determining the sequence of clusters[16][19] over the time period without specifying the number of clusters[17][18] to be produced and therefore results in less time consuming, robust and reduced search space frame work for evolutionary clustering.

III. EVOLUTIONARY CLUSTERING OF DYNAMIC INFORMATION NETWORKS BASED ON HADOOP MAP REDUCE

The framework consists of following steps:

Step 0: Store the input dataset in HDFS.

Step 1: The sorted key, value pairs are distributed to mappers which apply map function.

Step 2: The output of map function is consolidated per key using reduce.

Step 3: The similarity function is applied to determine the clusters based on the labels provided.

Step 4: The output clusters are evaluated with the evolutionary patterns

IV. EXPERIMENTS AND RESULTS

Experiments were performed with the four area dataset of DBLP network. Papers from the area DB, DM, IR, ML from the year 2010 to 2016 are used for experiment. This is called as four area dataset. Our experiment is done with 10K papers of four area dataset, initially. It consists of data objects such as author, paper, conference and terms. Later we increase the size of dataset progressively 20K,30K,40K,50K. In the first step the input is stored in HDFS in the form of key, value pairs;<First Author,Paper Title>. The sorted key value pairs are distributed to mappers by job tracker, the map() will determine, count of papers written by certain author and the reduce function as per parameter provided sum the total number of papers written by an author removing redundant records. To determine the clusters of papers belonging to identical research area based on the labels such as DB, DM, a similarity function is applied and the output clusters are written to a stable storage.

This experiment is performed in hadoop architecture of version 1.1.0, consists of one master node and two slave nodes each of which have 7.5GB memory. In many of the evolutionary algorithms the running time is directly proportional to the data size. However, the hadoop framework , by subdividing the work among multiple nodes increases the speed of execution time even though it encounters the overhead of job tracking and task tracking.

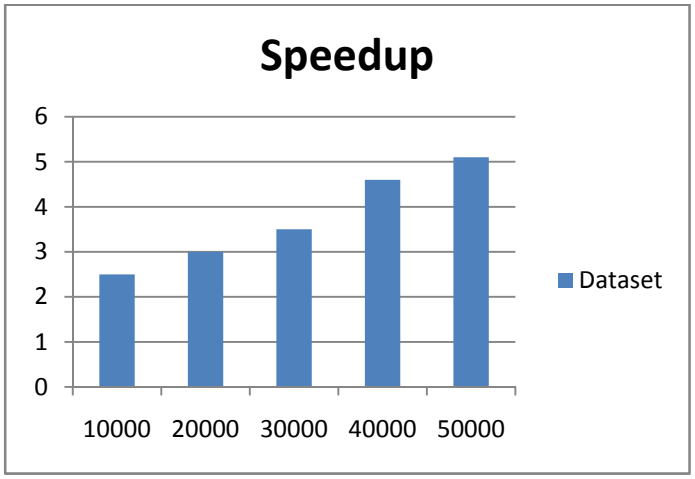

Fig.1 The execution speed up with the increase in data size

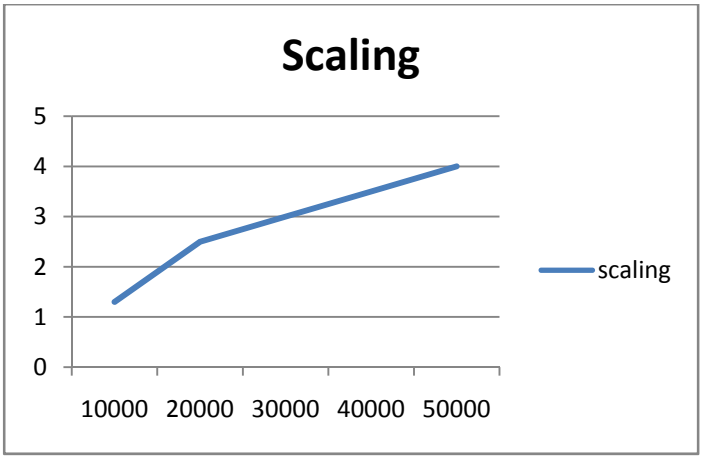

Fig. 2 The horizontal scaling of hadoop with increase in data size

Figure1 shows the running time of the different size of datasets. The execution time of 50K dataset is reduced comparing with the 10K of dataset. This shows that the framework is suitable for handling large datasets and the execution speed will be increased when the size of the dataset increases. Figure2 shows the measure of scaling for our experiment. As the result shows, the performance increases as the job size increases. This is the result of distributed architecture of storing and processing data in hadoop. The results prove that the framework based on hadoop mapreduce for evolutionary clustering of heterogeneous dynamic network is suitable for huge social networks.

V. CONCLUSION

In the broad area of evolutionary clustering various techniques were proposed so far. However, to handle the huge data set of social information networks need a technique to reduce the execution time and space. In this paper, we propose a framework to implement the evolutionary clustering based on hadoop mapreduce. The results show that the execution time is much reduced as a result of splitting the work among multiple nodes and the generation of clusters over time periods instantly provides a promising effect in the field of evolutionary clustering.

REFERENCES

[1] D.Chakrabarti, R.Kumar, and A. Tomkins. “Evolutionary clustering “ KDD(2006)
[2] Y.Chi, X.Song, D.Zhou, K.Hino and B. Tseng. “Evolutionary spectral clustering” KDD(2006)
[3] Xu X, Yuruk N, Feng Z, Schweiger T A J. SCAN: A Structural Clustering algorithm for networks, Proceedings of the 13th ACM SIGKDD international Conference on Knowledge discovery and data mining, KDD’07, 2007, pp. 824-33.
[4] Lin Y R,ChiY,ZhuS,SundaramH, Tseng B L, FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. In Proceedingsof the 17th international Conference on world wide web, WWW’08, 2008, pp. 685-94.
[5] TangL,LiuH,ZhangJ, NazeriZ. Community evolution in dynamic multimode networks. In Proceedingsof the 14th ACM SIGKDD international Conference on Knowledge discovery and data mining, KDD’08, 2008,pp. 677-85.
[6] KimM S,HanJ. A Particle and Density based Evolutionary clustering method for dynamic networks,Proceedings of the VLDBEndowernment, 2009, 2(1), pp. 622-33.
[7] SunY,HanJ,ZhaoP,YinZ,ChengH,WuT.Rankclus: Integrating clustering with Ranking for heterogeneous information network analysis,Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, EDBT’09, 2009,pp. 565-76.
[8] SunY, HanJ, Yu Y. Ranking-Based Clustering of Heterogeneous Information Networks with star network schema,Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD’09, 2009, pp. 797-806.
[9] Gupta M, Agarwaal C C, Han J,SunY. Evolutionary Clustering and Analysis of bibliographic networks, 2011 International Conference on Advances in Social Network Analysis and Mining (ASONAM),Kaohsiung, 2011, pp. 63-70
[10] Donn Morrison and Ian McLoughlin and Alice Hogan and Conor Hayes “Evolutionary clustering and analysis of user behavior in online forums” AAAI(2012)
[11] Y.Sun,Jie Tang,J.Han, Cheng chen and Manish Gupta. “Co-Evolution of multi-typed objects in dynamic star networks” IEEE(2013)
[12] Lammel, R.: Google’s MapReduce Programming Model - Revisited. Science of Computer Programming 70, 1–30 (2008)
[13] Hadoop: Open source implementation of MapReduce, http://lucene.apache.org/hadoop/
[14] Ghemawat, S., Gobioff, H., Leung, S.: The Google File System. In: Symposium on Operating Systems Principles, pp. 29–43 (2003)
[15] Borthakur, D.: The Hadoop Distributed File System: Architecture and Design (2007)
[16] S. Vijaykumar, M. Balamurugan, S.G. Saravanakumar, Unique Sense: Smart Computing Prototype, Procedia Computer Science, Volume 50, 2015, Pages 223-228, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.04.056.
[17] Vijaykumar S, Dr. M. Balamurugan, Ranjani K, Big Data: Hadoop Cluster Deployment on ARM Architecture, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 4, Special Issue 1, June 2015, ISSN 2278-1021 & 2319-5940.
[18] Visalatchi. L,Dr.M.Balamurugan, An Efficient Evolutionary Clustering using Strengthened Density based Algorithm, IJARTET, Volume 3, Special Issue 20, April 2016,ISSN:2394-3777
[19] Visalatchi. L,Dr.M.Balamurugan, Self Constructing Clusters in dynamic multi typed heterogeneous networks, IEEE Xplore, IEEE, ISBN:978-1-4673-7807-9
[20] Prasanka.R,Dr.M.Balamurugan, The Comprehensive analysis for Mining the knowledge using Feature Reduction techniques in Medical Data, IJCTA, 8(5), 2015, pp. 2119-2126
[21] Prasanka.R,Dr.M.Balamurugan, “Comparison of Data Mining Techniques for Feature Reduction using Evolutionary Search”, , IJARTET, Volume 3, Special Issue 20, April 2016,ISSN:2394-3777