Detecting Communities over Large Scale Graph Structure Data using MapReduce

Harsha J. Kolhe **Prof. Amitkumar Manekar**

Abstract — **With the appearances of the internet there is growing interest in executing analysis tasks over large scale graph structure data. This task includes processing of sub- graph or multi-hop neighborhoods in graph. Examples of these graphs include identifying social circles, modified recommendation, Anomaly finding, link prediction and so on. These works are not well served by the vertex centric approach and by using this approach there is a problem of high communication, scheduling and memory expenses. There is an NSCALE approach that is a novel end-to-end graph processing approach for distributed execution of complex subgraph analytics over the large scale graph in the cloud. The phase GEL, graph extraction and loading of the NSCALE approach extracts only the appropriate portions of the graph to minimize the communication cost.**

Keywords: **subgraph extraction, cloud computing, Vertex centric framework, graph partitioning.**

I. INTRODUCTION

Graphs are very attractive when it comes to modeling real world data, because they are flexible more than tables and rows in a RDBMS. Expressing information network data as a graph is most natural with nodes representing the entities and edges denoting the interaction between them. There is rising need for executing complex analytics over such graph data to get valuable insights into the networks functional abilities, for scientific discovery, for event or anomaly finding and so on. As the world is moving towards a growth of Big Data, representing such data in the forms of graphs and performing graph analytics over such large volumes of graph data has become a crucial task.

Developing distributed graph processing frameworks for such tasks is being adopted widely everywhere.

Consider a simple example of social network which uses graphs for its representation. Social networks are normally modeled as graphs where entities are the nodes, and edges connect two nodes if the nodes are related by the relationship that characterizes the network. Most vertex centric graph processing approaches like Pregel[10],Grace[20], GraphLab[18] ,GPS[25]. Within these entire approaches user write vertex level program and that program is executed by the approaches in the synchronous or asynchronous manner. These vertex-centric approaches give the result in high communication expenses and slow iterative convergence. Instead a novel approach called NSCALE[1] is suggested which uses sub-graph-centric framework and allows distributed execution of analysis task. NSCALE approach allows users to write programs at neighborhood or sub-graph level then that program is executed on subgraph in iterative manner. The graph extraction and loading phase[1] of NSCALE extracts only the relevant data from the graph then applies the cost based optimizer for replicating data to minimize the number of machines. NSCALE also uses the distributed execution engine to reduce the memory. The distributed execution engine executes the user program on the particular subgraph. The optimization of data replication helps to improve the performance and scalability of the system.

II. LITERATURE SURVEY

In 2003 Arvind Arasu presents an EXALG algorithm for extracting the structured data from a collection of web pages generated from a common template[4]. But this algorithm cannot develop the techniques for indexing and providing querying support for the structured pages in the web. Then N. Kashtan proposes a efficient sampling algorithm that allows the evaluation of sub graph concentrations[5] and detecting network motifs at a runtime that is independent of the network size. The communities in social network rises over time so the L. Backstrom addresses the questions of which communities will grow rapidly and how do the overlaps among the pairs of communities change over time?[6] .By using two large sources of data he addresses that questions. The sampling algorithm [4] is used for detecting the network motifs but there is limitation for that motifs only 6 to 8 nodes are detected so [7] presents a novel algorithm for discovering large network motifs based on a novel symmetry breaking technique and this algorithm also increase the speed of previous methods. In 2010 predicting the positive and negative links in social network are identified by the J. Leskovec. For this identification there is sign prediction mechanism for determining the sign of links in the social network [9]. The Pregel is a system for large scale graph processing framework. In this vertex centric approach is used.

Programs are expressed as sequence of iterations and the vertex receives the messages sent by the previous iteration and then send messages to other vertices and modify its state[10].

For declarative analysis of large scale information networks W. E. Moustafa presents the architecture of data management system that system is called as GRDB[11]. The key of this approach is to decouple the operations required for traversingthe graph and do the modification and updation of that graph. The supervised random walk is a new learning algorithm for link prediction and link commendation. This algorithm combines the information from the network formation with node and edge level attribute [12]. The J. Kong develops a robust and formal approach to recovering interface semantics using graph grammars [14]. But there is a disadvantage of this approach that an error in image recognition algorithm can affect the interpretation of a web interface. To process large amount of data in parallel is called Map Reduce presented in [15]. This also described hadoop an open source implementation of map reduce to process large scale datasets. The Y. Dong proposes a ranking factor graph model for predicting links in social networks [16]. After that how to automatically reformulate users initial keyword query into similar or related ones for better query understanding or recommendation is the shortcoming of the [17]. But an automatic keyword query reformulation is proposed to overcome this shortcoming by exploiting structural semantics in structured data.

Then a graphlab structure [18] is proposed for graph parallel computation this also designed a distributed data graph for efficient load balancing. The number of triangle is a computationally graph statistics which is frequently used in complex network analysis so for counting this triangle an efficient triangle counting approximation algorithm[19]is used.

This algorithm combines the sampling algorithm of [4] and partitioning the set of vertices into a high degree. By combining the simple programmability of Bulk Synchronous Model with the high performance of asynchronous execution GRACE [20] presents a new parallel graph processing framework. GRACE provides a synchronous program execution with the asynchronous execution. To effectively partition a large graph to process complex graph operations efficiently the VB partitioner [21] method is used. D. M. Thomas propose data structure called the augmented extremum graph and use it to design a novel symmetry detection method [23] based on robust estimation of distances. But there is the shortcoming of this method that is the symmetry detection critically depends on the selection of a meaningful set of seeds. A bad selection of seeds may lead to incorrect formation of super seeds which in turn affect the quality of the detected symmetries.

GPS (Graph Processing System)is based on the BSP[20] and used to execute the large graphs. MPI standard interface is used for building a broad range of message passing algorithm [25]. In the previous method of vertex centric graph model the execution time of graph processing or iterative execution takes very long time to overcome this problem think like a graph paradigm [26] is proposed. For discover social circles in ego network J. McAuley develop a model for detecting the circles that combines network structure as well as profile information[27]. The Sequential Pattern Mining algorithm is designed to efficiently identify the most common sequences occurring in a large collection of event series. To provide the scalability to thousands of computers, cloud computing is the emerging technology for large scale data analysis for this MapReduce programming framework is used[30].

III. Methodology

A. Existing System

In the existing system most vertex centric approaches are used for the analysis of large graph. That are GPS, Kineograph, Grace, Pregel, GraphLab. In all these frame work there is a huge communication and high memory requirements because user write a vertex level program and that program is executed by the framework. So the vetex centric approaches failed to analyze the large scale graph data.

B. Proposed System

In proposed system, NSCALE framework will load the graph on to the distributed memory then MapReduce framework generates the key value pairs. Hadoop supports the distributed execution of computation so this framework loads graph on few machines as possible to reduce the memory requirements and minimize the communication cost. In this system user specifies the set of subgraph of interest and then write a user specified program .This system also increase the overall performance and reduce the communication overhead.

C. Working of MapReduce Phase

The MapReduce phase generates the key value pairs sort the same key value pairs as shown in fig. 1.

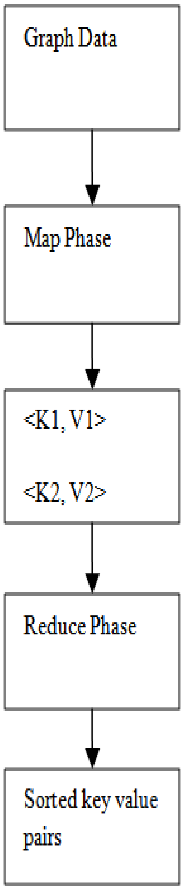


Fig. 1 Working of MapReduce.

IV. IMPLEMENTATION DETAILS

A widely-used class of cloud computing programming model is Map-Reduce , in which map functions read in the input data, and emits multiple intermediate <key, value> pairs. Reduce functions then merge the emitted <key, value> pair such that all values associated with the same key are paired together. Programs written using map and reduce functions are automatically parallelized and scheduled on a large cluster of commodity machines. In this process, scheduling, balancing maintenance, and error detection and recovery are automatically managed by the cloud computing platform such as Hadoop. Users only need to concern with the implementation detail of the algorithm. This allows users without parallel programming experience to utilize the computing resource.

From social network data, where user and his social network is represented, social circles will be identified and recommended to users. Social circle automatic detection can be treated as a clustering problem in the users ego-network which can find other densely connected user set. To find social circle of users, vertex centric oriented computation with map reduce generates huge communication, memory overheads and multiple iterations. Subgraph centric computation of community detection will be implemented using map reduce framework on Hadoop.

Initially social network dataset will be pre-processed to generate key value pairs which will be representing edge of users social graph. First map reduce job will be configured to emit this key value pairs. Further subgraphs will be computed with second map reduce job. In this, map phase will consume results of first job and replicate sufficient data to reduce communication between map reduce job instances. After this subgraph computation, next map reduce job will be configured and launched for community detection on precomputed sub graphs.

Social network data from facebook, twitter, gplus hosted on http://snap.stanford.edu/ will be used for analysis. Dataset description is given in Table 1. These datasets will be pre-processed and till converted to key value pairs of graph edges. Edge data of graph will be further analyzed to create social circles by clustering key-value pair of edges. Few clustering iterations will help to associate person/vertex to appropriate circle. Result of clustering of edges will find social circles. These social circles will be used to create similarity matrix between circles. From similarity matrix, recommendation of communities will be computed. Every analysis used map-reduce approach for computation. cost-based optimizer for data replication and placement. In doing this it aims to minimize the number of machines needed and balances load amongst them.

Table 1

Name	# Nodes	# Edges
EU Email Conn Network	265214	840090
Note Dame Web Graph	325729	2,994,268
Google Web Graph	875713	10,210,078
Wikipedia Talk Network	2,394,385	10,042,820

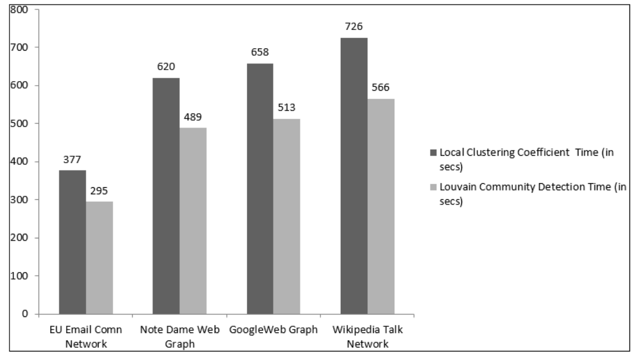
V. PERFORMANCE ANALYSIS

In this system the result is taken on to the hadoop clusters to reduce the time and communication in detecting the communities. So it requires 1 Master node and 6 slave nodes to run the large scale graph structure data. The every node having the configuration as intel i3 processor ,2GB RAM and 500 GB hard disk. The table 2 shows the result of our system,time is reduces as compared to the existing system.

Table 2

Algorithm /Dataset	EU Email Conn Network	Note Dame Web Graph	Google Web Graph	Wikipedia Talk Network
Local Clustering Coefficient Time (in secs)	377	620	658	726
Louvain Community Detection Time (in secs)	295	489	513	566

The following fig .2 shows the performance of our system.



This graph shows the execution time in seconds.

fig. 2 graph denotes the execution time

VI. CONCLUSION

There is growing interest in executing graph analytics on the large scale graph structure data. But the previous vertex-centric are limited in their ability to express and efficiently execute complex and rich graph analytics tasks. NSCALE proposes a sub-graph-centric framework where the users can write computations against entire sub-graphs or multi-hop neighborhoods in the graph and provide ease-of-use and efficiency. Also the graph extraction and loading phase saves total execution time for small where when Apache Giraph fails.

NSCALE approach increase the performance of system and minimize the communication overhead.

ACKNOWLEDGMENT

The authors would like to acknowledge Computer Engineering department, SITRC and all the people who provided with the facilities being required and conductive conditions for completion of the paper.

REFERENCES

[1] Abdul Quamar University of Maryland Amol Deshpande University of Maryland Jimmy Lin University of Maryland ” NScale: Neighborhood centric Large-Scale Graph Analytics in the Cloud”, cs.DB,7 may 014.

[2] Mark S. Granovetter,””The strength of weak ties”, American Journal of Sociology, 1973.

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U.Alon, ”Network motifs: Simple building blocks of complex networks”,Science, 2002.

[4] A. Arasu and H. Garcia-Molina, ”Extracting Structured Data From Web Pages”,SIGMOD 2003 June 9-12 ,2003.

[5] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon,”Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs”, Bioinformatics, 2004.

[6] L. Backstrom , D. Huttenlocher, J. Kleinberg and X. Lan , ”Group Formation in Large Social Networks: Membership ,Growth and Evolution”,KDD’ 06 August 20-23,2006.

[7]J. A. Grochow and M. Kellis, ”Network Motif Discovery Using Subgraph Enumeration and Symmetry Breaking”,RECOMB 2007,LNBI 4453,pp. 92-106,2007.

[8] Ronald S Burt, ”Structural holes: The social structure of competition”, Harvard university press, 2009.

[9] J. Leskovec, D. Huttenlocher and J. Kleinberg, ”Predicting Positive and Negative Links in Online Social Networks”, April 26- 30,2010.

[10] G. Malewicz, M. H. Austern, A. J.C Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, ”Pregel: a system for large-scale graph processing”,In SIGMOD, 2010.

[11] W. E. Moustafa, G. Namata, A. Deshpande, and L. Getoor , ”Declarative analysis of noisy information networks”, In ICDE Work- shops,2011.

[12] L Backstrom and J Leskovec, ”Supervised random walks: Predicting and recommending links in social networks”, In WSDM, 2011.

[13] Jiewen Huang., D. J. Abadi, and Kun Ren, ”Scalable SPARQLQuerying of Large RDF Graphs”, In PVLDB, 2011.

[14] J. Kong, O. Barkol, R. Bergman, A. Pnueli, S. Schein, K. Zhang and C. Zhao , ”Web Interface Interpretation Using Graph Grammars”, Grammars”,IEEE Transaction on Systems VOL,42,NO. 4,JULY,2012.

[15] S. Daneshyar and M. Razmjoo, ”Large Scale Data Preprocessing Using Mapreduce In Cloud Computing Environment”,ijwsc.2012.

[16] Y. Dong , J. Tang, S. wu, J. Tian, N. V. Chawla, J, Rao, H. Cao,”Link Prediction and Recommendation across Heterogeneous Social Networks”,IEEE 12th International Conference on data mining, 2012.

[17] J. Yao , B. Cui , L. Hua and Y. Huang, ”Keyword Query Reformulation on srtuctured Data”,IEEE 28th International Conference on Data Engineering, 2012.

[18] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein,”Distributed GraphLab: A Framework for Machine Learning in the Cloud”, PVLDB, 2012.”

[19] M. N. Kolountzakis, G. L. Miller, R. Peng, and C. E. Tsourakakis, ”Efficient triangle counting in large graphs via degree-based vertex partitioning”,Internet Mathematics, 2012.

[20] G. Wang, W. Xie, A. J. Demers, and J. Gehrke, ”Asynchronous Large-Scale Graph Processing Made Easy”,In CIDR, 2013.

[21] K. Lee , L. Liu, ”Efficient Data Partitioning Model for Heterogeneous Graphs in the Cloud”,SC,13 November 17-21,2013.

[22] J. Seo, S. Guo, and M. S. Lam, ”Socialite: Datalog extensions for efficient social network analysis”,In ICDE, 2013.

[23] D. M. Thomos and V.Natarajan,Member IEEE,”Detecting Symmetry in Scalar Fields Using Augmented Extremum Graphs”,IEEE Transaction on Visualization and Computer Graphics, vol.19,no. 12, December 2013.

[24] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, ”Wtf: The who to follow service at twitter”,In WWW, 2013.

[25] S. Salihoglu and J. Widom,”GPS: a graph processing system” ,In SSDBM, 2013.

[26] Y. Tian, A. Balmin, S. A. Corsten, S. Tatikonda, and J.McPherson, ”From Think Like a Vertex to Think Like a Graph”PVLDB, 2014.

[27] J. McAuley and J. Leskovec, ”Learning to Discover Social Circles in Ego Networks”,In NIPS, 2014.

[28] C.D.stopler,A. peter and David Gotz, Progressive Visual Analytics :User Driven Visual Exploration of In-Progress Analytics”, IEEE Transaction on Visualization and Computer Graphics,2014.

[29] Shiming Zhang, Yin Yang, Wei Fan and Marianne Winslett, ”Design and Implementation of a Real-Time Interactive Analytics System for Large Spatio-Temporal Data”,VLDB Endowment, Vol. 7, No. 13,2014.

[30] R. Baraglia, C. Lucchese and G. De Francisci Morales, ”Large Scal Data Analysis on the Cloud”.

AUTHOR’S PROFILE

	Author’s Name : **Harsha J. Kolhe** B.E. Computer(North Maharashtra University) PG Student, Savitribai Phule Pune University, SITRC College, Nashik-422213 E-mail id: harsha.kolhe123@gmail.com

	Author’s Name: **Asst. Prof. Amitkumar Manekar** Work Experience 1.6 Yrs in Company and 5 Yrs in teaching Ph.D Pursuing in CSE as specialization in Big Data Analytic and Cloud Computing Domain . Manekar A.S. Malti Nalge, et al.” Sifting Of A Potent Convex Hull Algorithm For Scattered Point Set Using Parallel Programming”, 2013 5th International Conference on Computational Intelligence and Communication Networks,IEEE/DOI 10.1109/CICN.2013.121.