KANGAROO: A DISTRIBUTED SYSTEM FOR SNA
Social Network Analysis in Huge-Scale Networks

Wu Bin, Dong Yuxiao, Qin Lei, Ke Qing and Wang Bai
School of Computer Science, Beijing University of Posts and Telecommunications, Tucheng 10, West Road, Beijing, China

Keywords: Social network analysis, Distributed system, MapReduce, Huge-scale network.

Abstract: Social network analysis is the mapping and measuring of relationships and flows between people, groups, computers and other information or knowledge entities. The continued exponential growth in the scale of social networks is giving birth to a new challenge to social network analysis. The scale of these graphs, in some cases, is millions of nodes and billions of edges. In this paper, we present a distributed system, KANGAROO, for huge scale social network based on two main computing models which are for finding common neighbour and maximal clique. KANGAROO is implemented on the top of the Hadoop platform, the open source version of MapReduce. This system implements most algorithms of social network analysis, including basic statistics, community detection, link prediction and network evolution etc. based on the MapReduce computing framework. More than anything else, KANGAROO is applied to a real-world huge scale social network. The application scenarios, including degree distribution, linear projection algorithm for community detection and community visualization of presentation layer, demonstrate KANGAROO is efficient, scalable and effective.

1 INTRODUCTION

Social networks have grown stronger as a form of organization of human activity. Social networks are nodes of individuals that tie in one or more types of independencies. It represents the human relationships by the links among the nodes. There are so many ubiquitous social networks around us, the co-authorship network, the mobile call network, the email network, the Facebook network etc. There are two kinds of forms to represent information of the social network: matrix and graph (Costa et al., 2005). In this paper, we use graphs to learn and understand the social networks with nodes representing actors and edges representing relations. The graph theory is the foundation of social network analysis.

With the expanding range of human activities, human relations are becoming increasingly complex and closely. This results in more and more large social networks. In some cases, the scale of the network has billions of nodes, trillions of edges. The continued exponential growth in scale of networks is giving birth to a new challenge to the social networks analysis. The traditional social network analysis based on graph theory mainly can not be able to play a significant role because of the restriction by Moore's Law.

Recently, Yahoo have come up with a peta-scale Graph Mining System, namely PEGASUS (Kang, 2009)and Google have dealt with this intractable problem by using PREGEL which is also a system for large-scale graph processing (Malewicz et al., 2010). The two great companies concentrate on the graph mining, and we put our focus on the distributed network analysis using graph mining in this paper. At the same time, Yahoo's PEGASUS is run on M45, one of the top 50 supercomputers in the world, however, our system, namely KANGAROO, is constructed on commodity machines without supercomputers. The powerful capability of MapReduce distributed computing framework provides a solution of efficient processing of graph mining. The move to distributed and parallel computing model is being driven not because it is an idea of dividing and conquering, but because it is an efficient way to utilize numerous surplus computing resource of large amount of commercial equipment (Yang et al., 2009).

The rest of the paper is structured as follows. Section 2 presents the related work. Section 3

Bin W., Yuxiao D., Lei Q., Qing K. and Bai W..
KANGAROO: A DISTRIBUTED SYSTEM FOR SNA - Social Network Analysis in Huge-Scale Networks.
DOI: 10.5220/0003387304040409
In *Proceedings of the 1st International Conference on Cloud Computing and Services Science* (CLOSER-2011), pages 404-409
ISBN: 978-989-8425-52-2

describes the model of computing. Section 4 presents the framework and Implementation of KANGAROO. Section 5 discusses several application scenarios. In the last Section, we discuss the conclusion and future work.

2 RELATED WORK

The related work forms two groups, social network analysis and Hadoop MapReduce framework.

2.1 Social Network Analysis

A social network is a social structure made of nodes that are tied by one or more specific types of interdependency (Costa et al., 2005). Its origin can be traced back to the pioneering works on percolation and random graphs by Flory (1941), Rapoport (1957) and Erdos and Renyi (1961), social network became a focus of attention for that it had characteristics which are not like the uniformly random complex network. Social network involves community structure, power law degree distributions and bubs, among other features (Barabasi and Albert, 1999).Two developments motivated many ongoing research: Watts and Strogatz's investigation of small-world networks (Watts and Strogatz, 1998) and Barabasi and Albert characterization of scale-free models (Barabasi and Albert, 1999). In this paper, we demonstrate the efficiency and effectiveness of our framework by focusing on presenting the insights of several large real social networks. In this paper, we put our focus on the basic statistics, link prediction, maximum clique and community detection.

2.2 MapReduce Framework

MapReduce is a programing framework (Dean and Ghemawat, 2004) for processing huge amounts of unstructured data in a massively parallel way, collectively referred to as a cluster. Computational processing can occur on data stored in a distributed file system. The framework is inspired by map and reduce functions commonly used in functional programming. In Map step, the master node takes the input, chops it up into smaller sub-problems, and distributes those to worker nodes. In Reduce step, it then takes the answers to all the sub-problems and combines them in a way to get the output. MapReduce framework has two may advantages: (a) the programmer is oblivious of the details of the data distribution, replication, load balancing etc. And furthermore (b) the programming concept is familiar, for example, the concept of functional programming (Wu et al., 2009).

Hadoop is an open source implementations of the MapReduce framework. It provides a distributed file system (HDFS) to store data and a means of running applications on large clusters built of commodity hardware which is near the data.

3 MODEL OF COMPUTING

The input to a KANGAROO computation is an undirected graph G. V(G) represents vertices and E(G) represents its edges. In this paper, each vertex is uniquely identified by a string and each edge is associated with the source vertex and destination vertex, and it consists of its weight, too. For a vertex v in G, let $\tau(v) = \{u \in V \,|(u, v) \in E\}$ represents the neighbors of v, and $k_v = |\tau(v)|$ represents the number of the neighbors, namely the degree of the node v.

For example in Figure 1, we lay out a typical KANGAROO computation. Figure 1.b represents the input data of the social network which has a simple graph structure composed of 7 vertices and 11 edges in Figure 1.a.

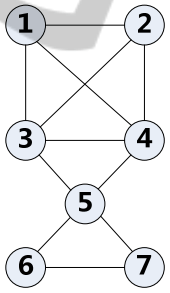

Figure 1.a.

1	2
1	3
1	4
2	3
2	4
2	5
3	4
3	5
5	6
5	7
6	7

Figure 1.b.

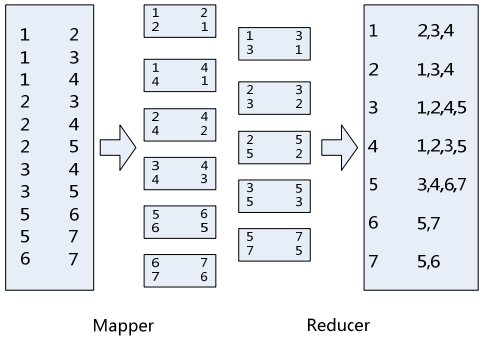

Figure 2: The Process for Adjacency List.

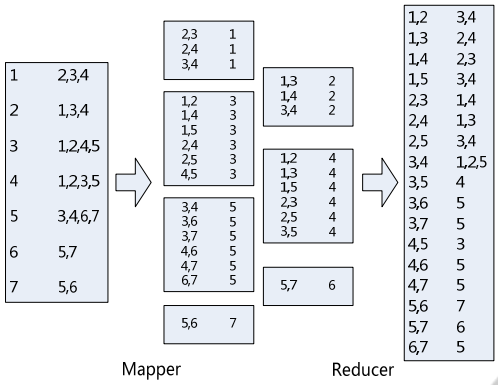

Figure 3: The Process for Common Neighbors.

3.1 MapReduce for Common Neighbors

In graph G, A node z which belongs to the common neighbors of nodes x and y, means that the neighbors of node x contain z, and the neighbors of node y also contain z. In Figure 1.a, the common neighbors of nodes 1 and 2 contain node 3 and 4.

Given a Graph, firstly, we generally transform it to an adjacency list from the form of Figure 1.b, which can be efficiently employed for further analyses. In Figure 1.a, we launched a MapReduce job to obtain the adjacency list of this graph. The flow of the data is represented in Figure 2.

With the adjacency list of the graph, we can get the common neighbors using MapReduce as described in Figure 3.

3.2 MapReduce for Cliques

In a graph G, a clique C is a complete subgraph of G in which any two nodes are adjacent. A k-clique defined as a clique containing k nodes (k>=3) (Yang et al., 2009). A k-clique is called maximal clique if its edge between any two nodes is not part of any other clique which has more than k nodes. In Figure 1, for example, the graph consists 3 maximal cliques, (1,2,3,4), (3,4,5) and (5,6,7).

The process for cliques has the same first step that is to get the adjacency list of every node with finding common neighbors. Adjacency list is the one-leap information of a node. However, to extract a maximal clique, we should obtain the two-leap information of every node. Figure 4 represents the process of getting maximal clique.

The models for common neighbors and cliques with MapReduce are the foundation of algorithms in KANGAROO, such as degree distribution, link prediction based on node similarity, community detection etc.

4 FRAMEWORK AND IMPLEMENTATION

In this chapter, we describe the overview of KANGAROO and the basic implementation of this system, without mentioning the specific technological detail.

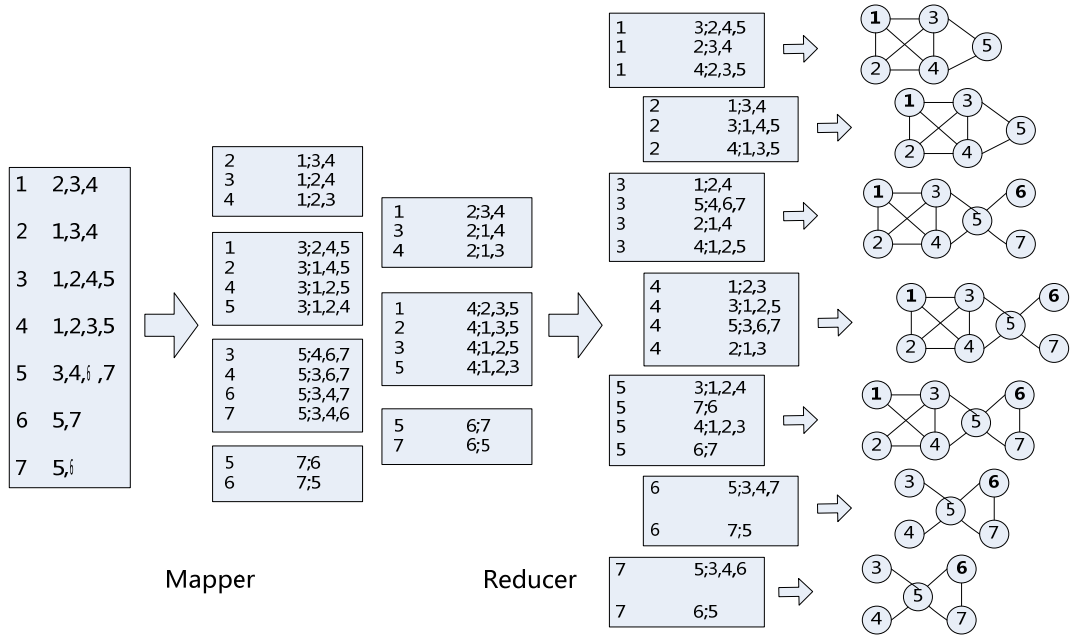

Figure 4: The Process for Maximal Clique.

4.1 Basic Architecture

The overview architecture of KANGAROO is presented in Figure 5. It has a hierarchical architecture with three main layers: cluster layer, core algorithm layer and high layer service.

The function and feature of each layer is described as followed:

- In cluster layer, there is cluster environment, distributed computing platform and layer of preprocess. The cluster environment consists of heterogeneous servers with which virtualization provides flexible and transparent perspective. Besides, Hadoop cluster supply the distributed file system and distributed computing ability.
- Core algorithm layer is constructed on the top of the cluster layer, intending to provide the core algorithms of social network analysis which consists of basic statistics, community detection, link prediction and network evolution etc.
- High layer services supply the presentation and solutions which are based on the basic algorithm layer. The result of the social network analysis is presented by the forms of the visualization, report, and text, also provides the data mining and business intelligence solutions.

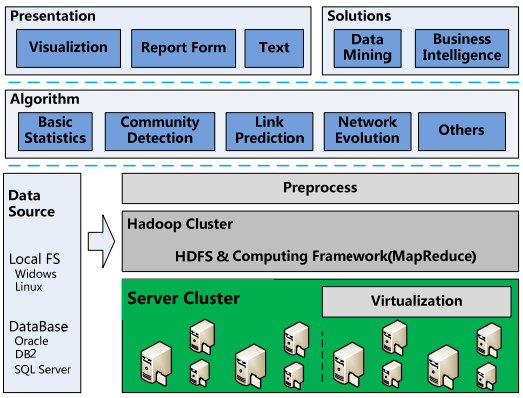

Figure 5: KANGAROO overview.

4.2 Hardware Construction

We set up a hadoop cluster environment, composed of one master and 32 computing nodes(Intel Xeon 2.50GHz×4, 8GB RAM, 250GB×4 SATA II disk, Linux RH4 OS). The cluster is interconnected through 1000Mbps Switch. And deployed Hadoop platform version is 0.20.0

4.3 Core Algorithms

Social network analysis is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information or knowledge entities (Kredbs 2004). To analyze the social network, an adequate acquaintance of network is necessary by basic statistics. After that, the process of social network analysis includes the community detection, link prediction, network evolution and other aspects. KANGAROO provides both a mathematical and visual analysis of human relationships by these basic algorithms as described below.

- Basic Statistics: In this section, KANGAROO supply the implementation for understanding the overview of the network, including the numbers of nodes and edges, degree distribution, weight distribution, betweenness centrality, cluster coefficience, closeness centrality, shortest path, diameter and page rank etc.
- Community Detection: KANGAROO implements the main community detection algorithm, including the clique percolation method and linear projection algorithm.
- Link Prediction: KANGAROO provides the nine local similarity-based link prediction algorithms, Common Neighbors, Salton Index, Jaccard Index, Sorensen Index, Hub Promoted Index, Hub Depressed Index, Leicht-Holme-Newman Index, Preferential Attachment Index and Adamic-Adar Index, and two global similarity-based algorithms including Katz Index and Leicht-Holme-Newman Index (Lv and Zhou, 2010).
- Network Evolution: Here, KANGAROO uses an effective algorithm TrackMapper (Yang et al., 2009)that can generate an evolving timeline throughout the network, because of its optimization for MapReduce model.

To the best of our knowledge, KANGAROO implements these algorithms by using MapReduce first, including betweenness centrality, cluster coefficience, closeness centrality, shortest path, network diameter, linear projection algorithm and all link prediction algorithms which are based on node similarity.

5 CASE STUDIES

In this chapter, we validate KANGAROO by an analysing procedure. The graph employed here is a

huge undirected social graph that has 4.8 million nodes and 43 million edges extracted from directed LiveJournal friendship online social network. The following experiments on this huge dataset are inspected from the scalability and the performance.

5.1 Degree Distribution

In the study of networks, the degree of a node in a network is the number of edges it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network (Barabasi and Albert, 1999).

In this section, KANGAROO run the MapReduce implementation of degree distribution job. We show the results of degree distribution in Figure 6 and the chart demonstrates the power low degree distribution. As represented in Figure 7, there is a good speedup ratio with the change of the computing node numbers. When the number of nodes comes 16 or 32, there is redundant computing resource for this job.

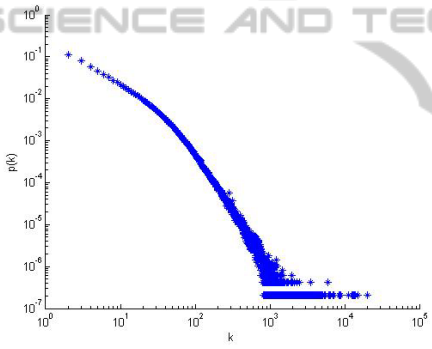

Figure 6: The degree distribution of LiveJournal.

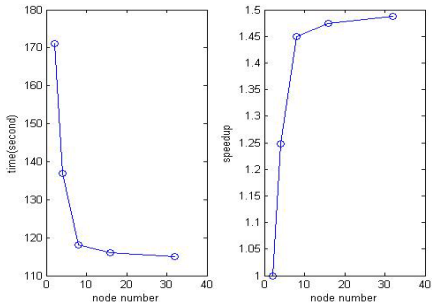

Figure 7: Running time and scalability of degree distribution.

5.2 Linear Projection Algorithm

Community is groups of nodes within which connections are dense, but between which connections are sparser. Next we intend to find meaningful communities by linear projection algorithm. Linear projection algorithm resolves community structure by transforming network community detection problem into a common clustering problem. It uses the factor that the main features of the community structure are actually captured by just the low-dimension vectors, which allows KANGAROO to reduce the computational cost (Liao, 2009).

The paralleled linear projection algorithm using MapReduce has a nearly linear speedup ratio before the number of computing nodes comes 8 in the environment of KANGAROO. As Figure 8 shows, that superfluous computing resource exists is the same with the process of degree distribution when there are 16, 32 or more nodes.

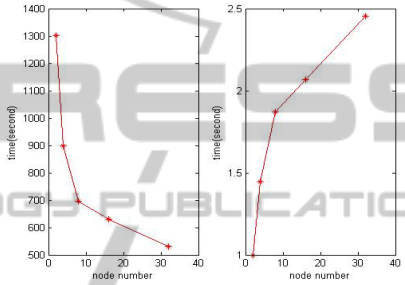

Figure 8: Running time and scalability of LPA.

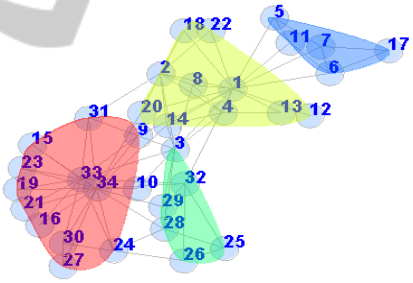

Figure 9: The community of Karate graph.

Figure 10: The community of American football games graph.

5.3 Community Visualization

The visualization is a significant part of KANGAROO, also a very difficult part. In this section, we show the community visualization which is most tough task for the presentation layer.

After community detection on Karate graph, the visualization of the communities is represented in Figure 9 which is clear and intelligible to distinguish the different nodes among all communities. In Figure 10, we represent a more beautiful graph which contains 12 communities in diverse colors.

6 CONCLUSIONS AND FUTURE WORK

Motivated by recently increasing request for the large scale social network analysis, in this paper, we introduce our solution, KANGAROO based on distributed system, and report how to construct it using the open-source Hadoop project, also the advantages of this system, including the ability to analyze the huge scale social network and the high performance with linear speedup ratio.

We employ a huge scale real-world network with ten million nodes and edges to analyze with KANGAROO, including basic statistics, community detection and community visualization. In these cases, we do not only show the result of our algorithms but also the speedup ratio of our system.

At present, KANGAROO is an on-going system, our future work will continue in the construction of this system, especially the computing modes and algorithms for large scale social network, scalability and performance. Fundamentally, we hope KANGAROO can serve as a practical social network analysis framework for huge scale real-world network.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation of China (Grant No.60905025, 90924029, 61074128), National High Technology Research and Development Program of China (No.2009AA04Z136), National Key Technology R&D Program of China (2006BAH03B05) and the Fundamental Research Funds for the Central Universities.

REFERENCES

U. Kang. Charalampos E. Tsourakakis, Christos Faloutsos. 2009. PEGASUS: A Peta-Scale Graph Ming System – Implementation and Observations. In *ICDM2009, Ninth IEEE International Conference on Data Mining.*

Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski. 2010. Pregel: A System for Large-Scale Graph Processing. In *SIGMOD2010, ACM SIGMOD International Conference on Management of Data.*

Shengqi Yang, Bai Wang, Haizhou Zhao and Bin Wu. 2009. Efficient Dense Structure Mining using MapReduce. In *ICDM2009, Ninth IEEE International Conference on Data Mining workshop on Large-scale Data Mining.*

A. L. Barabasi and R. Albert. 1999. Emergence of scaling in random networks. In *Science, 286(5439):509-512.*

D. J. Watts and S.H. Strogatz. 1998. Collective dynamics of small-world networks. In *Nature, 393(6684):440-442.*

Linyuan Lv, Tao Zhou. 2010. Link Prediction in Complex Networks: A Survey. In *arXiv:1010.0725v1 [Physics and Society (physics.soc-ph)] 4 Oct 2010.*

Shengqi Yang, Bai Wang, Haizhou Zhao, Yuan Gao, Bin Wu. 2009. DisTec: Towards a Distributed System for Telecom computing. In *International Conference on Cloud Computing 2009.*

Bin Wu, Shengqi Yang, Haizhou Zhao, Yuan Gao and Lijun Suo. 2009. CosDic: towards a Comprehensive System for Knowledge Discovery in Large-scale data. In *The 2009 IEEE/WIC/ACM International Conference on Web Intelligence 2009.*

J. Dean and S. Ghemawat. 2004. Mapreduce: Simplified data processing on large clusters. In *OSDI 2004*

L. da F. Costa, F. A. Rodrigues, G. Travieso, P. R. Villas Boas. 2005. Characterization of Complex Networks: A Survey of measurements. In *Condensed Matter/0505185*

P. J. Flory. 1941. Molecular size distribution in three-dimensional polymers. i. gelation. In *Journal of the American Chemical Society, 63:3083-3090*

A. Rapoport. 1953. Contribution to the theory of random and biased nets. In *Bulletin of Mathematical Biophysics, 19:257-277, 1957.*

P. Erdos and A.Renyi. 1961. On the strength of connectedness of a random graph. In *Acta Mathematica Scientia Hungary, 12:261-267, 1961.*

Valdis Kredbs 2004. Valdis Krebs' website for Inflow, a software-based SNA tool. In *http://www.orgnet.com/sna.html*

XiaoPing Liao, Wei Ren, Guiying Yan. 2009. A Linear Projection Approach for Resolving Community Structure. In *The Third International Symposium on Optimization and Systems Biology 2009.*