Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 5 (2013) 170 – 176

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Parallel and Distributed Computing and Systems

Parallel Micro Blog Crawler Construction for Effective Opinion Leader Approximation

Xiaolong Deng[a], Yuxiao Li[b], Shu Lin[a,b,*]

[a, b] *Beijing University of Posts and Telecommunications, Beijing, 100876,China*

Abstract

It is important to research and find out opinion leader in time of increased huge number of Micro Blog users around our world and daily life. Our governments in different countries around the world have found that it is a vital fact to analysis and control message spreading in Micro Blog and other social network. For the reason that opinion leaders always play vital role in information spreading in Micro Blog and it is necessary to detect opinion leaders precisely. In this article, we has constructed a SINA Micro Blog APIs based Micro Blog crawling and analysis tool. Furthermore, a novel MapReduce based Micro blog crawler and node betweenness approximation computation method was proposed better accuracy and less running time to detect core opinion leaders in Micro Blog graphs.

"Keywords:Opinion leader detection, Micro blog, Betweenness, Approximation Modeling;"

1. Introduction

In recent years, our earth has experienced some important changes such as culture globalization, Internet popularization and so on. In the last decade, there are more crisis accidents in social public which occurred much more frequently than before especially in some serious natural environment cases. All the government public management around the world cannot avoid the challenges which has become particularly vital in recent years. These public crisis events began to frequently occurred in China also. There are some examples. SARS broke out event in 2003 of China, big discussion over in Blogs for WenChuan earthquake in 2008,Sichuan Lushan earthquake in 2013, April 20 and so on. From mentioned above events, we can infer

doi:10.1016/j.aasri.2013.10.074

that China has entered a time of public crisis happens much more frequently than before and Government has been more concentrated on information spread in our society in these social crisis.

In this article, It had done some research on the different centralities of information dissemination mechanism such as degree centrality and Betweenness centrality statistics from the related outstanding foundlings of node centrality and network information dissemination efficiency in complex systems. According to our experiment results on SINA Micro Blog dataset of social network and some simulated datasets from scale-free network [1], betweenness centrality of the node , which always acts as a very efficient role for detecting opinion leaders to research message spread mechanism in simulated and real social network dataset, and it tell us to control some vital node of opinion leader in public crisis management can promote the efficiency of public crisis managing system.

The rest of this article will be organized as follows: section II gives a detailed introduction on the related research work on complex network for Betweenness computing, node centrality and parallel computing, including introduction of three important centralities for social network analysis in important opinion leader detection. Section III will presents detail information of our SINA APIs based crawler. In Section IV will we will present the parallel Design details of parallel SINA Micro blog crawler .And In Section V, we developed an effective Betweenness approximation algorithm for large scale graph and evaluated it in experiments on real SINA Micro Blog User social network dataset. Section VI concluded the paper.

2. Related Work

2.1. Social Network Model Definition

Social Network is often abstracted into graph model for further research and experiment. Social network's network models discussed in this article are defined as networks whit unweighted and undirected nodes and edges which can often be abstracted as $G = (V, \varepsilon)$. And in the definition, ε is the set of edges with edge number $L = |\varepsilon|$ while V is the set of nodes with node number $N = |V|$. In the model of social network in this article, the node has its meaning of real person or registered network ID of Micro Blog and so on.

In this article, social network of Micro Blog is expressed as undirected graph G and unweighted graph G .And here follows the definition: G is a unweighted graph and the edge weight between different nodes will not be considered and its adjacency matrix is $\mathcal{A}$ which can be presented in $N \times N$ to express N nodes in G .If there exists an edge in graph G between node v_i and v_j or not, the matrix element $\mathcal{A}_{ij}=1$ and else $\mathcal{A}_{ij}=0$ will demonstrate the status of this edge. While in weighted graph G ,when the weight of edge e_{ij} is w_{ij} ,the matrix element $\mathcal{A}_{ij}$ will be filled with $\mathcal{A}_{ij}=w_{ij}$.

2.2. Node Centrality Statistics

Social network researchers can discover the dependent relation among different nodes in social network structure by analysis of node statistic characters in node status and influences [1]. And the different role of node can be test and evaluated by quantitative investigation.

In past three decade, society researchers have done some statistics research which focus on role and power of social network node by using the centrality characters of node. Among these researchers , Scott [2], and Han [3] has done some shining result on point centralities or the graph centralities which composed two separated different research objects in social network. And when we do actual analysis of social network structure, we often chose some existing definitions to test node's importance with node centralities which

including the most common node centrality statistics are Betweenness centrality[4],Degree centrality[5] and Closeness centrality[5] .

- Firstly, we will introduce Betweenness centrality. Node's Betweenness centrality shows how many shortest paths across the node or the edge. Node's Betweenness centrality can be defined as[4]: $B_u = \sum_{i,j} \frac{\sigma(i,u,j)}{\sigma(i,j)}$, Where $\sigma(i,u,j)$ is the number of shortest paths between vertices i and j that pass through vertex or edge u , $\sigma(i,j)$ is the total number of shortest paths between i and j ,and the sum is added over all pairs i and j of different and distinct vertices.
- Secondly, we will introduce Degree centrality. Node's Degree centrality shows how many ties to other node this node holds. In graph theory a node's degree is adjacent edges number which the node has. While $deg(v_i)$ is degree value of node v_i , the sum degree of all nodes in graph G is $\sum_{v_i \in V} deg(v_i)$, and degree centrality of node v_i is $D(v_i)$ .Definition of $D(v_i)$ is $D(v_i) = \frac{deg(v_i)}{\sum_{v_i \in V} deg(v_i)}$.
- Thirdly, we will introduce Closeness centrality. Node's Closeness centrality shows how closely the node ties to other nodes. If the geodesic path length between v_i and v_j is $d(v_i, v_j)$, the closeness centrality of node v_i is $C(v_i) = \frac{1}{\sum_{j=1}^{g} d(v_i, v_j)}$,in which $\sum_{j=1}^{g} d(v_i, v_j)$ stands for the total all geodesic path length of node v_i to all the nodes v_i could connect to.
- For the Node's centralities above, Betweenness centrality is the most frequently used centrality statistics to detect opinion leaders in social network such as in Micro Blog relationship networks.

2.3. Node Betweenness Computation and Parallel Programming

For the reason that node's Betweenness centrality reflects the vital role of node in information transmission and node's important role social of network structure,it is necessary to develop some fast algorithm to computation node's Betweenness centrality statistic with lower computation cost and better precision .There are some traditional computation methods such as such as GN [6],Fast GN[6] and Floyd Warshall algorithm [6].Among all the existing Betweenness comuputation algorithm ,it is necessary to compute all node's Betweenness in the whole network.And because its heavy computation cost in time complexity , GN algorithm and Floyd Warshall algorithm often needs $o(m^2 n)$ in network with high density and $o(n^3)$ in some sparse network in the network with m edges and n nodes.In recent years ,there some researchers proposed some faster Betweenness computation algorithm for whole network such as Brandes [7] developed the faster node Betweenness computation algorithm on whole network based on graph backtracking algorithm.In 2001,Brandes's algorithm has the time complexity of $o(mn)$ in undirected and unweighted graph,while time complexity of $o(mn + n^2 \log n)$ in weighted graph.Other researcher also proposed some fast node Betweenness computation algorithm for whole network such as David's algorithm in 2008 [8] proved the bottom edge of time complexity is $o(mn)$ in the above Betweenness algorithm which are based on graph path compare method with m edges and n nodes.

With the development of various computation in and industries and services, both in the fields of physics, chemistry, biology and so on. There for in the manufacturing industry, military industry, telecommunications service industry field needs the powerful computing ability and data analysis ability. But in actual computation , all node's Betweenness centrality computation in social network always needs very heavy time complexity so that it is necessary to develop some parallel Betweenness centrality computing algorithms.In many cases, the calculation and data processing requirements can not be able to complete the computing work

for results which have strict requirements and limitations in computing speed, needing time and accuracy. The traditional uni-processor and serial program has resulted many problems in some large-scale data processing and calculation.And in these cases, enterprises and scientific researchers always demands calculation for the data (Data-Oriented Computing) as the most important thing.In order to support these large-scale computing tasks, related hardware support system have emerged one after another, such as Intel Paragon, IBM SP2, Intel TFLOPS in 1990s, and Lenevo's dawning-1000 distributed memory computer. They are all present to carry on large scale cluster computation as Amazon cluster system, Google and Yahoo (typically containing tens of thousands of computing nodes) cluster ,and the new IBM super computer HPC Roadrunner, which have reached the computing capability of billion times.

On the other hand, researchers used MPI[9] and Hadoop[10] to develop parallel program to promote algorithm running efficiency. Programming method and theory of MPI is often difficult for program developers and existing methods and algorithms does not construct node Betweenness approximation algorithm for the whole social network especially in large scale graph in MapReduce APIs.It is necessary to construct that in MapReduce API to cut down computation complexity.

3. SINA Micro Blog Crawler Construction

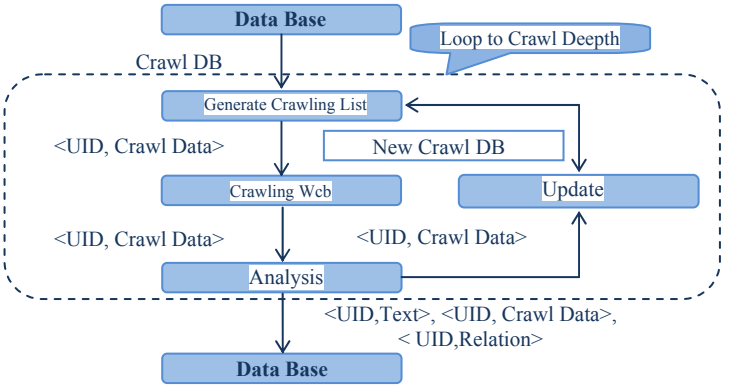

Figure 1. Working Process of SINA APIs Based Crawler

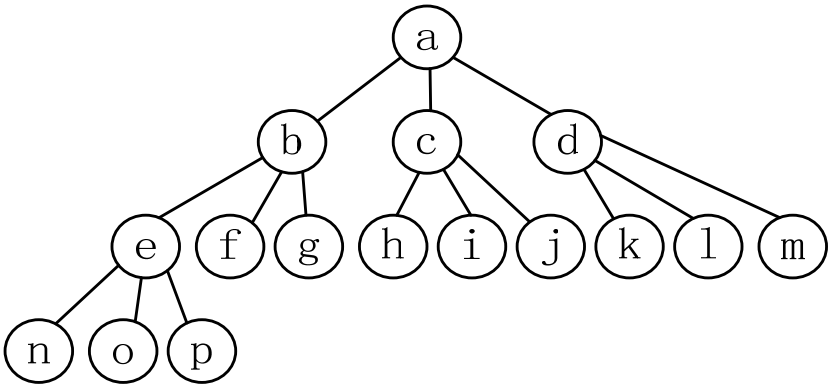

Figure 2. User's Fans Relationship

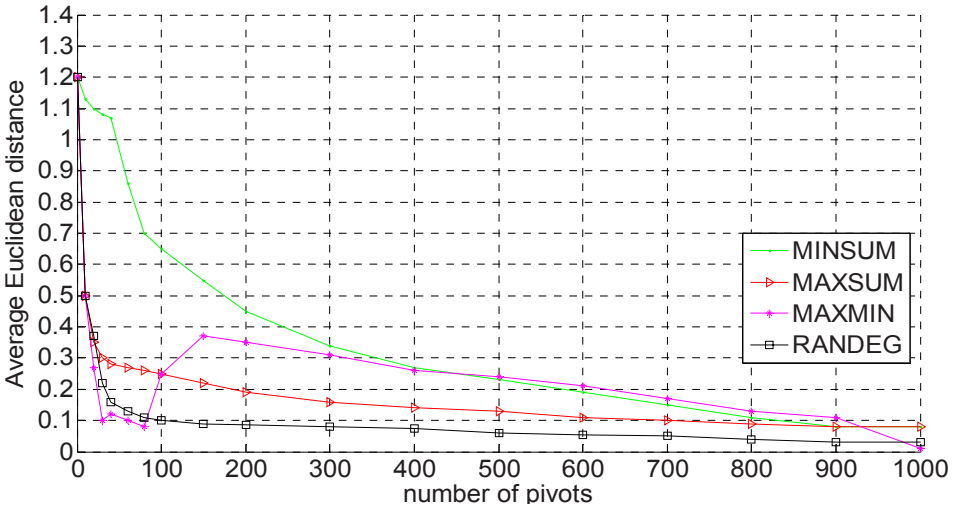

Figure 3. Betweenness approximation results in 2012.July.7 Dataset

We designed and implemented SINA Micro Blog APIs based crawler which can be found the crawling process in Figure 1.When Micro Blog information crawling began, Crawling DB thread get a random user ID from database and retrieve its Micro Blog fans user into crawling queue. For an example which can be founded in Fig 2,the Crawler thread will search and get the Micro Blog home page of the initial user a and push its fans b,c,d into the crawling queue. After process of a, the crawling thread will handle b,c,d and so on.

With the help of our SINA APIs based Micro Blog page crawling software, we has collect amount of some SINA Micro Blog datasets of famous emergency cases such as the rainstorm in Beijing City in July 21,2012.And in the following prcess ,we will extract the user's fans relationship in Figure 1 for very large graph for user node Betweenness Centrality computation for detecting opinion leaders. And here follows some java code of the SINA Micro Blog crawler:

```
import weibo4j.*;
public class MyWeibo {
    static private String myUid;
            MyWeibo() throws WeiboException, JSONException {
            String access_token = MyAccessToken.value;Weibo weibo = new Weibo();
            weibo.setToken(access_token);Account am = new Account();myUid = am.getUid().getString("uid");
    }
    /*      To get basic information of  SINA Micro Blog User,return : User */
    public User user() {
            Users um = new Users();                    //User information for origional log in
            try {    User us = um.showUserById(myUid);
            } catch(WeiboException e) {e.printStackTrace();
            } return us;
    }
    /* To Get Followers list of the User,return : UserWapper*/
    public UserWapper followers() {
            Friendships fm = new Friendships();UserWapper uw = null;
            try {     UserWapper users = fm.getFollowersById(myUid);uw = users;
            } catch(WeiboException e) {e.printStackTrace();
            }return uw;
    }
    /* To Get the interested person of User,return : JSONArray*/
     public JSONArray suggestionHot() {
            Suggestion sg = new Suggestion();
```

```
            JSONArray sh = null;
            try {
            JSONArray hot = sg.suggestionsUsersMayInterested();sh = hot;
            } catch(WeiboException e) {e.printStackTrace();
            }return sh;
    }
……}
```

4. Implementation Detail of Parallelled SINA Micro Blog Crawler

To promote crawling efficiency of SINA Micro Blog message, a Hadoop(using hadoop-0.19.0-core.jar) based parallel crawler is constructed to carry this job. Backend data of the parallel crawler is storaged on Hbase(using hbase-0.19.6.jar) and Zookeeper(using zookeeper-3.3.3.jar) to coordinate distributed system storage and parallel computing. And input of every iteration from the parallel crawler is original SINA Uids and accumulated added new SINA UIDs in last crawling iteration, while result some real-time SINA Micro Blog massages' Chinese word segmentations for further human Micro Blog sentiment process.

5. Betweenness Computaion of Opinion Leader Detection

5.1. Betweenness Approximation Algorithm

In implementation of our analysis tool, parallelized node Betweenness computation algorithm is developed by graph Forward and Backward steps basing on Brandes's algorithm steps [7] which using the Node Path Dependence thesis. From Nanavati's research founding[11] that Network Diameter of some typical social network such as call graph and SMS graph is around 12. In ours Betweenness calculation algorithm sub graph's hop number n around any node is $6 \le n \le 12$ according to Nanavati's and Strogatz's result which network diameter in people small-world network is around 6.While hop is $6 \le n \le 12$, algorithm proposed by this article decreased much computation time while keep a fine approximation accuracy result.

In our algorithm implementation,our software developed different four Pivot node selection strategies and they are RANDEG，MAXMIN，MAXSUM and MINSUM strategy and each selected 1000 Pivot nodes while using formula of $X(v)=\frac{1}{k}(X_1(v)+X_2(v)+...X_k(v))$ and $X_i(v)=\frac{n}{n-1}\delta(p_i|v)$ to calculate node's Betweenness centrality of a subgraph in 12 hops according to Brandes's founding. [7]And $\delta(p_i|v)$ is Pivot p_i 's Node Path Dependence to node v .

5.2. Computation Result

From Fig.3,the Betweenness approximation result can be found in a dataset which comes from SINA's Micro Blog user fans relationship graph in case of July 21 rain storm case of Beijing in 2012,and there are 1000 nodes was randomly selected to make Betweenness estimation. In Fig.3,we can find the selected node number in x-axis and the Average Euclidean distance among estimated Betweenness value and real Betweenness value of selected 1000 nodes mentioned above in y-axis.From Fig.3. MAXMIN strategy was found to be the best one and deviation of MAXMIN strategy reaches still below 0.1 while Pivot number reaches 80.

6. Summary

From this article, a more efficient parallel SINA Micro Blog crawler and parallel betweenness

approximation computation method was proposed in large scale social graph for some emergency case in our daily life such as rainstorm. And it is proved that our algorithm gained better accuracy in less cost of commutating time. It is useful to detect some important opinion leaders in social network in large scale graph and in parallel data mining.

Acknowledgements

Support by Youth Innovation Project of BUPT(G470638,G470704),NSFC Project(91124002), China Culture Support Project (2013BAH43F01)

References

[1] Xiaolong Deng, Lei Zhang, Bai Wang, Bin Wu. Implementation and Research of Social CRM Tool in Mobile BOSS Based on Complex Network[C]. Proceedings of ICISE, 2009, pp.870-873.
[2] Scott, Social Network Analysis: A Handbook. Newberry Park, CA: Sage.
[3] J.W. Han, M. Kamber. Data Mining: Concepts and Technique [M]. The Morgan Kaufmann Series in Data Management Systems 2nd Edition, Morgan Kaufmann Publishers ,2006
[4] Jure Leskovec, Kevin J. Lang, Michael Mahoney.Empirical comparison of algorithms for network community detection[C].In WWW '10: Proceedings of the 19th international conference on World wide web (2010), pp. 631-640.
[5] B.Bollobas. Modern Graph Theory [M]. New York: Springer, 1998.
[6] M E J.Newman. Fast Algorithm for Detecting Community Structure in Networks. [J] Phys. Rev. E, 2004, 69: 066133.
[7] Brandes. A Faster Algorithm for Betweenness Centrality [J]. Journal of Mathematical Sociology, 2001, 25(2):163-177
[8] David.A. Betweenness Centrality: Algorithms and Lower Bounds [J]. Journal of Computing Research Repository. abs.0809.1906: (2008)
[9] Du Zhihui.High performance MPI Programming[M]. Tsinghua University press,2001.
[10] HDFS, http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html
[11] Nanavati. On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications[C]. In CIKM '06. pp.435-444, in New York, USA.
[12] Michael Baur.Group-Level Analysis and Visualization of Social Networks [J]. Algorithmics of Large and Complex Networks.Lecture Notes in Computer Science, 2009, Volume 5515/2009, 330-358