An unbiased crawling strategy for directed social networks

Xuehua Yang[1,2], HongbinLi[2*]

[1]*School of Software, Shenyang Normal University, Shenyang 110034, Liaoning, China*

[2]*Shenyang Institute of Computing Technology Chinese Academy of Science, Shenyang 110168, Liaoning, China*

Received 1 June 2014, www.cmnt.lv

Abstract

Online Social Networks (OSNs) is a hot research topic and data crawling or collection is an important and based task for OSN analysis and mining. Due to the large amount of data, not open and other factors, the acquisition of social networking is different from the ordinary crawling technology. The quality of the data determines the effect of the majority of social network data mining analysis, data crawling technology is essential. Micro-blog is different from social network such as Facebook, the need for better crawling strategies to obtain the data set is huge. Improving Random Walking (RW) algorithm, an unbiased crawling strategy is proposed to crawling directed social networks. By contrast with the uniform sampling method, the strategy has been proved to ensure data crawling with all similar data at the same time to ensure the unbiasedness of the sampling data.

Keywords: social networks, sample collection, unbiased directed graph, crawling strategies

1 Introduction

With the prosperity of Online Social Networks (OSNs) and more prominent of its application value, the related researches are also flourishing. Most of the research concentrated on Social Network analysis, complex network, large-scale data mining and so on, including small-word rule, idempotent law and other network features, community discovery, friends recommendation and other applications, advertisement, market analysis, user behaviour analysis and other application researches. It’s easy to find a great research and practical value in mining and analyzing social network. However, data collection, a basic and important direction, has not yet been studied enough.

The effectiveness of research can be maintained only by sample data of high quality and practically effective, which can be achieved by data collection. Data collection of social network is, however, different from general crawling technology, which mainly caused by two reasons: (1) social network data cannot be collected totally; (2) sample data collected partly must meet characteristics of real social network, including small-word rule and so on. To (1), currently there are no data sets of total amount on the internet so that researchers have to collect good experimental data sets by themselves. This is mainly because: (a) social network data is too large. According to recent statistics, the amount of active users on Facebook per month has exceeded 1 billion and domestic micro-blog users have exceeded 500 million. It’s a colossal cost for researchers to crawl, store and calculate such a huge data. (b) Data of social network belongs to enterprises and commonly is kept secret. It’s hard to obtain due to problems such as privacy protecting. (c) Data crawling usually is highly limited by time, amount and network because of the huge swell of traffic to social network and service requirements. To (2), it’s a challenge and difficulty for collecting technology to determine whether the partial data collected can represents the complete real data or the distributions between the two are similar, whether the network graph is isomorphic, whether the sample data is bias to nodes of some kinds, etc.

The most popular social network sample crawling algorithms are BFS (Breath First Search) and Random Walking (RW) [7]. However, study [2] has implied that the data collected by these two methods deviates to nodes with high degree, meaning that these methods are biased. Metropolis-Hasting Random Walking (MHRW) algorithm from paper [3] proposes sample collection from undirected network graph, mainly aiming at social network such as Facebook. This algorithm can maintain unbiasedness during sampling, thus holding all the statistical properties of undirected social networks. It’s coming to nothing, however, when crawling algorithm like this is applied to social network like micro-blog for one can follow anybody but is followed by nobody. In other words, micro-blog is directed social network, thus in which the former method cannot be used to sample collecting. Due to the difference between inland and outland, it is more valuable to study different kinds of users’ proportion and activeness in social network like micro-blog and it’s a direction in need of development in domestic to study micro-blog data collecting technology.

In this paper we have improved the selection strategy of Random Walking algorithm and proposed an unbiased sampling and crawling algorithm in directed social network graphs. For the characteristics of directed graph, this crawling method is unbiased as well as holding the

[*] ***Corresponding author’s*** e-mail: yangxuehua_sict@aliyun.com

basic data quality. This method, compared with the marked real data (global uniform sampling) in experiment, can be used to collect unbiased data sample in directed social network.

2 Social networks and sample collection

Social network data contains content data and user data. It's the network graph formed by users that is useful for network analysis. This paper will mainly concentrate on the collection of user data.

2.1 SOCIAL NETWORK'S GRAPHICAL MODEL

The constituent parts of social network include user profiles, relationship, activity, contents and so on. Relationship between users is Friend ship in social networks like Facebook but Follow ship in micro-blog, in both of which users make up nodes and relationships make up edges, thus a network graph is constituted. The former is bidirectional which can be simplified as an undirected graph while the latter is unidirectional which can be simplified as a common directional graph.

Let graph G=(V,E) be a social network, where v∈V represents a user and e∈E represents relationship either a friend ship or a follow ship between users. div represents in-degree of vertex v and dov for out-degree in directional graph while dv represents the degree of vertex v in unidirectional graph.

Paper [5] thinks it necessary to judge the value of sampling from two aspects, where the sample should statically be similar to original graph (Scale-down) and should be corresponding with network evolution (Back-in-time). Given an original graph G, of which the scalar is n, and a graph S sampled from G, of which the scalar is n', and n'<<n, the principle of similarity says that the statistical properties of S must be similar to those of G. This paper has also counted nine kinds of statistical properties including distribution of in-degree, distribution of out-degree, distribution of connected subgraph, etc. The above properties of S and G are D-statistically tested. The corresponding principle of evolution requires the similarity of statistical properties of network snapshots at some point. Sampling based on Random Walking and methods based on Forest-Fire model are compared with. To structural similarity principle the former is optimal as experimental result indicates.

2.2 SOCIAL NETWORK CRAWLING TECHNOLOGY

Crawling strategy is a certain strategy by which the next crawling goal is chosen, thus deciding the collecting path, which has a decisive impact on the data collected by crawling. To search engine the crawling order is often determined by website's importance judged by PageRank and so on. To social network, however, these methods are not applicable because the network's global topological structure and inner features should be known first. Breath first search (BFS) strategy and walking strategy are mainly crawling methods used in social network graph. Walking strategy can be based on vertex or edge or both in network graph. It can be random walking or walking with certain probability or walking by path graded and selected by some algorithm, or like Forest-Fire algorithm, which randomly selects a seed node then sets it on fire and with some probability burns a certain out-edge, whose end point is fired according to certain probability.

There are crawling technologies commonly used in social networks. The first one to be considered is breath first search (BFS), so far the most widely used sample collecting strategy of OSNs. BFS is well-known as a biased strategy on importing nodes with high degree. Besides, this bias has no rules to follow. Let's secondly consider sampling by Random Walking (RW), which also will be biased to nodes with high degree. However, its bias at least can be quantized by Markov Chain analysis, thus be corrected through a proper Re-Weighted Random Walking (RWRW) algorithm. Thirdly, Metropolis-Hastings Random Walking (MHRW) algorithm will achieve directly goals about uniform and stable nodes (users) distribution. This technology recently is repeatedly applied to OSNs sampling [1]. Finally, as benchmark data, UNI strategy [6] is used to get sample of practically marked real data (UNI). This method is that carry a uniform sampling on users from a real social network website, which are selected from system's 32-bits id space by a sample-denying program. Real data like this cannot be accessed usually. With the existing dataset we can take it as a benchmark of the deviation.

3 Unbiased crawling strategies

In directed network like micro-blog, we may get a probability when we visit a node with 0 out-degrees, meaning that this node, from which, in general, walking to other nodes is infeasible, has not yet followed other user nodes. This signifies that once we've walked to this kind of node, we can never walk to other nodes with purely random. Besides, without proper strategy, nodes with high degree might be crawled all the time, which is beneath the representativeness of sample data. A proper crawling strategy is necessary when typical user nodes need to be collected as well as sample data's unbiasedness needs to be considered.

3.1 RANDOM WALKING ALGORITHM

Metropolis-Hasting Random Walking (MHRW) algorithm is a Markov Chain Monte Carlo (MCMC) algorithm. This is an algorithm getting sample from probability distribution due to the difficulty in directly sampling. It's improved from RW algorithm, whose basic idea is to choose the next crawling goal by a certain selection function.

In RW algorithm the transition probability from vertex u to v is:

$$P_{u,v}=\begin{cases}0 & \text{if v is not u's neighbor}\\ \dfrac{1}{d_u} & otherwise\end{cases}\quad . \tag{1}$$

This strategy is useful and, however, biased at the same time. The sample is more biased to nodes with high degree. It's a Markov processing because the next node rely on current visiting node. The probability of each edge is 1/|E|, thus the visiting probability of each vertex is ku/(2*|E|), that is to say that vertexes with high degree are easy to be collected with greater probability, which caused biasedness in sampling data set.

MHRW strategy aims at diminishing this weakness by improving selection strategy and therefore changing the transition probability function at the same time. It will generate a number α uniformly distributed between 0~1 and compare it with ratio of degrees to determine whether or not to transform:

$$Q_{u,v}=\begin{cases}1 & \text{if } \alpha<\dfrac{d_u}{d_v}\\ 0 & otherwise\end{cases}$$

Transition is chosen if $Q_{u,v}=1$. Then we'll get a new transition probability function:

$$P_{u,v}=\begin{cases}\min(\dfrac{1}{d_u},\dfrac{1}{d_v}) & \text{if v is u's neighbor}\\ 1-\sum\limits_{v\neq u}\min(\dfrac{1}{d_u},\dfrac{1}{d_v}) & else \text{ if v = u}\\ 0 & otherwise\end{cases} \tag{2}$$

It's found that transition probability of nodes with high degree has been reduced. At last, every node can be accessed with uniform probability. This algorithm is unbiased.

3.2 UNBIASED DIRECTED GRAPH CRAWLING STRATEGY

MHRW isn't suitable for directed graph because in this graph once it has walked to a node with 0 degree, it will never walk to another node with purely random walking. An intuitive solution is to randomly choose an adjacent node, which has a nonzero in-degree, as the next walking node. But this is biased to nodes with high degree. The Markov chain of every originating node is not sufficient to converge to target probability distribution. Therefore we've proposed unbiased crawling (UC) algorithm to take a full advantage of social network's properties and solve this problem by transforming directed graph to undirected graph through changing follow ship between users. It can, under this circumstance, access all the other connected nodes from an original node with connection. The problem of getting stuck into a certain node with 0 out-degrees doesn't exist any longer.

UC algorithm also relies on random walking while the transition probability function should be redefined. In undirected graph whether to transform depends on the comparison between ratio of nodes' degrees and the generated uniform distribution number while in directed graph this kind of degree doesn't exist. Every vertex in directed graph has in-degree and out-degree, which cannot be used directly in probability function. Noting that, in social network like micro-blog, the Following and Fans of users are actually in-edge and out-edge of vertices. The obtained Following and Fans can be treated as user's friends when crawling user's information and then it can be treated as an undirected graph to crawl. An unbiased crawling algorithm in directed graph can be obtained by unbiased MHRW algorithm in undirected graph.

The process of UC algorithm is as follows. First, the crawler crawls to vertex v as current status. Hereafter the transition probability function is $P_{u,v}$. Merge Following nodes and Fans nodes of node v to a set, that is, take unidirectional edges as bidirectional edges, both of which are considered as connected nodes of node v. Then, take u from connecting nodes of node v as a next sample. Next, generate a number α from uniform distribution U(0,1), take u as the next sample when $\alpha<d_u/d_v$. Otherwise, v still is sample. Therefore, the sample data tends to be uniform distribution after enough crawling.

MHRW ensures unbiased sample from an undirected social network graph. UC algorithm has improved this algorithm and can maintain data's unbiasedness and get sample data akin to complete data space distribution at the same time.

3.3 UNBIASED DIRECTED GRAPH CRAWLING ALGORITHM

The crawling strategy mentioned above can be represented as an algorithm with pseudo code, namely unbiased directed graph crawling algorithm (UC algorithm for short).

Algorithm 1: Unbiased Crawling Algorithm

```
Input: seed node v
Output: nodes data collected
get all the followings of node v, which is all the vertices corresponding with out-edge
get all the fans of node v, which is all the vertices corresponding with in-edge
nodes above duplicate removal, and put them into a set
for node u in set do
    get all the related nodes of node u
    compute the degree of node u
    generate α from 0-1 uniform distribution
    if α < d (u)/d(v)
        make u as the next collecting node
        put u in set
    else
        continue selecting next node from current node
end for
```

For every user node choose next collecting node, walk in social network and crawl nodes and relationship to get experimental data set.

The algorithm above is simple and useful, costs little to compute and makes full advantage of social network's features. It can avoid getting stuck into some local nodes, meanwhile can ensure the final data unbiased. It's, in data collecting from social network, like a common crawling process, which starts with some seed nodes, chooses path to traverse this network by a kind of crawling strategy until some end condition is satisfied and required data is obtained. This algorithm can be used as a crawling strategy to achieve collector commodiously.

4 Experimental analyses

With micro-blog as data source in experiment, it is found that in collecting the micro-blog is strict with network and too little data can be collected with provided API. Only crawling can be used to fetch and parse webpages. However, due to social network's kinds of anti-crawler technologies, it will be slow to crawl in micro-blog along with situations like interruptions and temporary validations. So we've chosen various kinds of social network corpus provided openly by Stanford [5]. Akin to crawler, we begin with original node, use UC crawling strategy, take advantage of directed graph relationship of corpus, and crawl in current corpus and finally collect sample data. Compared with given data, the evaluation is real and effective.

4.1 EXPERIMENTATION

The process is totally similar with crawler and the collecting result is consistent with the real crawling results when experiment on real data set. The data set taken by us can be found in paper [5]. This lab has collected a large number of network data from social network including undirected social network graph based on Friendship like Facebook and twitter which can be abstracted as directed graph. We use data set from directed social network including Twitter with unidirectional follow ship and so on. There are 81306 nodes and 1768149 vertices in Twitter data set, 77360 nodes and 905468 vertices in Slashdot0811 data set and 75879 nodes and 508837 vertices in Epinions1 data set.

For every data set, we calculate the degrees and connecting nodes of every node. The directed graphs are transformed to undirected graphs. Nodes with 0 degree are ignored for those nodes cannot be used as originating nodes, by which the result won't be affected because once we start random walking we'll never walk to those nodes. Then UC algorithm is used to crawl sample data. And finally the mean degrees of sample data and complete data are calculated and then the degree's cumulative distribution.

Testing under above environment, we calculate social network's mean degree and degree's cumulative distribution. Nodes with 0 degree have been eliminated for they are useless to general social network analyses. The distribution of complete nodes is our benchmark. This data set is obtained by uniform sampling methods so the benchmark is a data set of UNI method.

4.2 EVALUATION AND ANALYSIS

BFS and walking algorithm mainly used in collection of social network are primarily aiming at undirected graph like Facebook, and have considered not the goal of unbiasedness, because of which they cannot be contrasted with directed graph crawling algorithm proposed by this paper. The UC algorithm used in this experiment is to compare with UNI data, which is the distribution of original data, to determine whether or not the algorithm has generated experimental data akin to original data.

For user of directed social networks the out-degree is the number of user nodes followed by it and the in-degree is the number of its fans or nodes following it, both of which are close by mean value in a huge number of users. Therein, with evaluation result of UNI data set as benchmark, the data crawled by UC strategy proposed by this paper is very close to distribution of benchmark. Take deviation formula to evaluate this effect. Define formula of deviation σ as follows:

$$\sigma = \left| \frac{UC\text{-UNI}}{\text{UNI}} \right| \cdot 100\%$$

The in-degree and out-degree of different data set are counted according to our experimental purpose, between which the derivation with UNI data is calculated. It is found that in-degree and out-degree in directed social network are close and the difference between the data set generated by UC algorithm with distribution of UNI data is less than 6%. The result is stated in Table 1.

TABLE 1 Experimental result caused by UC algorithm with different data set

Data set		UNI	UC strategy	Deviation
Twitter	In-degree	21.747	21.515	1.07%
	Out-degree	22.361	21.796	2.53%
Slashdot 0811	In-degree	11.705	11.421	2.40%
	Out-degree	11.566	12.130	5.41%
Epinions1	In-degree	6.705	6.998	4.36%
	Out-degree	6.944	7.326	5.50%

The result indicates that the data collected have a similar structure in general with original data in which the number of in-edge is close with that of out-edge. It's not enough by only mean degree to explain why our strategy can get sample whose structure is akin to original data. For a more precise evaluation about whether distributions of the two are approximate, The Cumulative Distribution Function (CDF) has been introduced to compare UNI data with user nodes' degree distribution in data crawled by UC strategy. CDF can completely demonstrate the distribution of a real random variable X. It's the integral of probability density function.

We've mainly considered nodes' degree distributions and chosen in-degree and out-degree of sample nodes to calculate distribution function. To compare UC algorithm's result intuitively, we calculate all degrees' numbers of UNI data and sample data separately and then draw a cumulative distribution graph taking out-degree as variable as follow.

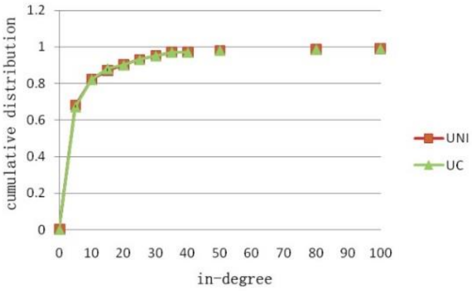


FIGURE 1 Crawling Strategy's In-degree Cumulative Distribution Graph

From these three different data sets, we have similar results. The number of user nodes will decrease with the increasing of degree. Because this is a sample data, in which the degree of collected data is relatively small and the mean degree is about 10. But this data compared to the complete real data is structurally similar, thus can be treated as a diffused space of original data. It's a UNI data. The real effect can be stated by comparison between data collected by UC strategy with UNI data.

Due to space limitation and the similarity in distribution diagrams of several data sets, only the in-degree cumulative distribution graph of Slashdot0811 has been displayed. As is expected from our experience, degrees mainly centralized at the beginning and also there're some nodes whose in-degrees or out-degrees have exceeded 1000. Compared to Twitter, Slashdot is a social media website where the activity of users is relatively low and the mean degree is relatively small. Only CDFs of nodes with in-degrees or out-degrees less than 100 have been drawn in above graph. From the graph we can see that the sample taken by our method is almost the same with UNI data set through our method. This demonstrates that UC strategy can get unbiased sample from directed social network.

5 Conclusions

In this paper we have presented sample collecting technologies of directed social networks, stated the formalized model and features of directed social networks and proposed and verified an unbiased sample crawling strategy, which can collect unbiased social network data set with features like distribution and can provide social network's mining and analyzing with a high-quality data foundation.

Because of the complexity and kinds of limitation of social network, the method proposed by this paper can be used to collect sample data set which is structurally similar with complete data in directed network like micro-blog. But the dimension of time evolution is not included. Big data technology is a new but not yet mature direction which contains many aspects like data modeling, data sampling and collecting, data mining, big data processing, big data storing, etc., in which data modeling and collecting is the most important foundation and needs to improve constantly.

Acknowledgments

This work was supported by Research of higher education Intelligence Platform at the age of big data, Education Scientific Research Planning in Liaoning province (JG14DB385).

References

[1] Leskovec J, Faloutsos C 2006 Sampling from large graphs *Proceedings of the 12th ACM international conference on knowledge and data mining* 631-6

[2] Catanese S A, De Meo P, Provetti A 2011 Crawling Facebook for social network analysis purposes *Proceedings of the International Conference on Web Intelligence Mining and Semantics* 52:1-52:8

[3] Gjoka M, Kurant M, Butts C T, Markopoulou A 2010 Walking in Facebook: A Case Study of Unbiased Sampling of OSNs *Proc. of IEEE Infocom* 1-9

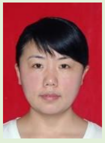
[4] Kwak H, Lee C, Park H, Moon S 2010 What is Twitter, a Social Network or a News Media *Proc of WWW* 335-45

[5] Stanford large network dataset collection: http://snap.stanford.edu/data/index.html

[6] Gjoka M, Kurant M, Butts C, Markopoulou A 2009 A walk in facebook: Uniform sampling of users in online social networks *Arxiv preprint arXiv*: 0906.0060

[7] Lu J, Li D Sampling online social networks by random walk. *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research ACM* 33-40

Authors



Xuehua Yang, 24.10.1978, China.

Current position, grades: lecturer at Shenyang Normal University, China.
University studies: pursued Ph.D. degree in computer science and technology from Shenyang institute of computing technology university of Chinese academy of science, China.
Scientific interest: big data and cloud computer.

Hongbin Li, 20.01.1973, China.

Current position, grades: researcher at computer science and technology from Computing Technology Chinese Academy of Science, China.
University studies: PhD degree in computer science and technology from Shenyang institute of computing technology Chinese academy of science, China in 2012.
Scientific interest: information security and big data.