ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

A Time Based Search Method for Online Social Networks

Aswathy.G.S[1], Sajni Nirmal[2]

M.Tech Student, Dept. of CSE, Marian Engineering College, Trivandrum, Kerala, India[1]

Assistant Professor, Dept. of CSE, Marian Engineering College, Trivandrum, Kerala, India[2]

ABSTRACT: Online social networking is booming in this modern era. The interest of users to explore the user network for information retrieval is increasing with the increasing usage of OSNs. So efficiency of the search methods becomes an important concern. Here proposing a search method based on sampling technique. By the usage of sampling, no need to visit all the nodes in the underlying user specified graph. The relevant information can be retrieved from the sampled nodes. Random walk sampling is a better method for collecting sample nodes from the network. Personalized searching also improves the searching. Log data can be used for personalized searching. To improve the efficiency of the algorithm, time can be considered as a factor to choose the samples. By using a time window the unwanted non-active samples can be eliminated and searching time efficiency can be improved.

KEY WORDS:- Online social networks,search methods, sampling, log data and time window.

I. INTRODUCTION

Nowadays the social networking emerges very quickly. Usage of the social networking becomes an inevitable part in this modern era. The large number of users in OSNs like Facebook, Twitter, Google plus etc. shows this. As the usage of OSNs increases, the interest of the users to collect information from the network structure also increases accordingly. So the efficiency of search method becomes an important concern. A research for an efficient search method has its own relevance.

The term social search describes the type of search which sort out the search results by considering the user related information. The social network structure can be viewed as a graph structure, with nodes representing the users and edges representing the relation among the users. Generally, the social search engines generate a ranked list of results when a user gives query to the search engine. Then the search engine retrieves information from the surrounding nodes of the user in the network structure. Using this information, search engine ranks the former list and a new ordered list is given to the user as the search result.

Many studies and researches carried out in the area of online social networking by considering its importance. Social searching has its own importance in the searching field. A work by A. Mislove [2] describes the benefits of social search. Many proposals for searching on OSNs are developed. But most of these methods are not efficient, because they returns large amount of results and many of these are unwanted. The social network structure can be viewed as a graph. Visiting all the nodes in the network to collect information is not a good practice. It is found that through sampling method efficient search can be performed. Personalized search also provides a good search result.
This paper proposes a sampling method which uses random walk sampling method integrated with time and log data to retrieve the information from the sample nodes.

II. LITERATURE SURVEY

The information collection from the nodes in the network is the important concern. The social network can be viewed as a graph structure. In this graphical consideration each user can be represented as the nodes and their relationship can be represented by edges in the graph. So for the collection of information from the network, it required

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

to traverse through the nodes. Different approaches have been proposed for this information collection from these nodes.
The traditional algorithms such as DFS and BFS are methods for collecting information from nodes. But these methods are not a good proposal because these algorithms visit all the nodes in the network. Hence the searching time will be increased and produces a large list of items as the result.

M. Makki and George Havas[3] proposed an improved DFS algorithm. In this method a depth first search tree is created for the communication network using distributed algorithms. So this method is called distributed DFS (DDFS). For an undirected communication graph G = (V, E), 2|V|-2 messages are used in the worst case to communicate by this method. The messages are sent to the nodes and receive back and a DFS tree is constructed based on this. In the worst case, the time and message complexity is |V| (1+r).

In [4], Knuth proposed random probing to estimate the size of the back tracking trees. This method performs random walks repeatedly from root to leaf nodes without backtracking. The next node is chosen in a way that at each node one of its successors chosen randomly with uniform probability. This method is a simple method but it has some limitations. Because this is an offline method. The size of the search tree size should be given to the algorithm prior to the search. So this method is not suitable for dynamic online networks.

Visiting all the nodes in the network is not an efficient way. It consumes much time. So an alternative approach is developed which is called sampling. In sampling method, not all the nodes in the network are visited. But the information is collected by visiting some sample nodes in the network. Different sampling methods are proposed for different cases.
For online applications Katzir proposed two methods [5]. The first method performs breadth first search on the network. This is not a good method. The second method collects samples from the network randomly. This method has a disadvantage that the samples must be uniform. But the online social network interfaces does not provide this functionality.

Arun S. Maiya and Tanya Y. Berger-Wolf [6] propose a novel method, to sample the communities in the network. They show that the sampling method can produce subgraphs from the original graphs. These subgraphs consist of members in the community network. They proposed two sampling methods called snowball sampling and MCMC sampling. These two methods are suitable to find communities in a network, not that much efficient for a single item searching.

In [7] and [8] the importance of personalized searching is mentioned. Personalized searching means, the searching methods consider the user specifying information such as their likes and dislikesfrom previous activities. By including this method, a user satisfied result can be produced.

III. EXISTING SYSTEM

G. Das, N. Koudas, M. Papagelis, and Nick Koudas [1] proposed a new sampling method called random walk sampling method. This method is integrated with personalized search to produce better search results. That is log data is used to collect information from the sample nodes.

The sample nodes are collected from the network by performing random walks on the graph. Each random walk is started from the user and crawls to the neighborhood of the user. The random walk selects a node as a sample if there is a self-loop occurs or if the maximum depth is reached. The number of the random walk depends on the number of samples required. By collecting all the random walks, a tree structure can be produced. This is the search tree in which the root node is the user and the leaf nodes are the sample nodes. The random walk sampling algorithm is given below:

1: procedure SAMPLING(u; n; d;C)
2: T = NULL, samples =0, Sample array of size n

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

3: while samples <= n do
4: if (v = randomWalk(u; d;C; T)) ! = 0 then
5: Sample[samples ++]= v
6: end if
7: end while
8: end procedure
9: procedure RANDOMWALK(u; d;C; T)
10: depth = 0, ps = 1
11: while depth < d do
12: pick v ∈children(u) U u with p_v = 1/degree(u)+1
13: if T U v has no cycle then
14: add v to T
15: $p_s = p_s - p_v$
16: if v = u then
17: accept with probability C/p_s
18: if accepted then
19: return v
20: else
21: return 0
22: end if
23: else
24: u = v, depth++
25: end if
26: end if
27: end while
28: return 0
29: end procedure

After collecting the sample nodes from the network personalized searching method is applied on the samples to retrieve information from the sample nodes and to rank the results. The method proposed in [1] is using the log data. Log data contains all the information regarding to the activities of each user such as sharing a file, liking a statement, visiting a profile etc. It assumed that the log data is accumulated over time for each user. Let $(x,count_x^v)$ be the form of the log at node *v*, where *x* is an item and $count_x^v$ is the number of times *x* has been endorsed by user *v*. Based on these count value, each item is ranked such that item having largest count value is given by first rank and so on.

The disadvantage with this method is that the sample set may contain non-active nodes in the network.

IV. PROPOSED SYSTEM

The time consumed part in the existing method is the calculation of count values for each item from each sample nodes. There are a huge number of users in online social networks. The network is dynamic in nature. Out of these users, there is a chance for occurring fake users and non-active members. These nodes are active in the network occasionally. So they do not contain much information required for the searching. And this will leads to unnecessary time consumption at the time of ranking. This is the main disadvantage of the existing method.

This paper proposes a solution to resolve this problem. That is, a time related concept is chosen to eliminate the non-active nodes from the sample set at the time of ranking purpose. After collecting the sample nodes through random walk sampling method, the sample set is given to the ranking algorithm. A time window [T, T+K] is used in the ranking algorithm to avoid the non-active nodes from these sample set. Those nodes whose login time that does not belongs with in this time window are not considered for the ranking purpose. By this method the unnecessary time taken for the information searching from non-active nodes can be saved.
The proposed ranking algorithm is given below:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

1: procedure RANKING(v;d;C;n;X;t)
2: S array of size n
3: Count array of size |X|
4: S = SAMPLING(v; n; d; C)
5: for all i ∈ S do
6: If login time of x lies with in [T, T+K]
7: for all x ∈ X do
8: Count[x] = Count[x] + $count_x^i$
9: end for
10: end for
11: return count
12: end procedure

At the end of this algorithm the count value for each item for each user is obtained. So aggregating these count values from each user from each node the total count value for each item can be obtained. This does not include the data from non-active nodes. Because they are eliminated before entering to the ranking process. Based on this total count value, the rank for each item is given. That is the item that associated with largest total count value is given by the first rank and so on. This ranked list is submitted to the user as the search result.

V. PERFORMANCE ANALYSIS

To calculate the performance of the proposed system and compare it with the existing system, a set of synthetic data is used. For this purpose, a synthetic network is created which contains hundred users and their relationships. In this synthetic network the users can establish relations with other users and can perform the activities like sharing a post, liking of images, commenting on posts etc. These activities are stored as log data and later this log data is used for ranking purpose. The existing searching method and proposed searching method is implemented on this synthetic network. Then the searching time, number of samples taken and the number of result produced are noticed. Using this information graphical representations are produced to compare and analyses the performance.
The obtained graphs are given below:

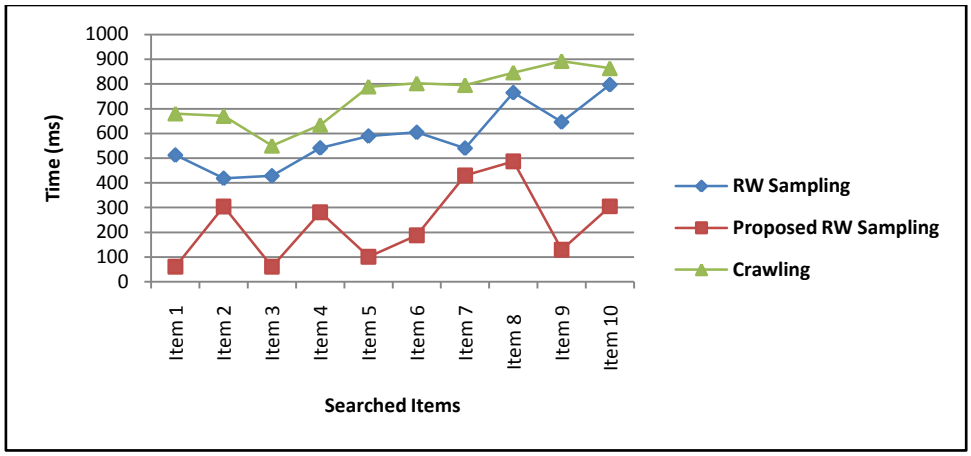

Fig1: Time taken by three methods to search different items

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

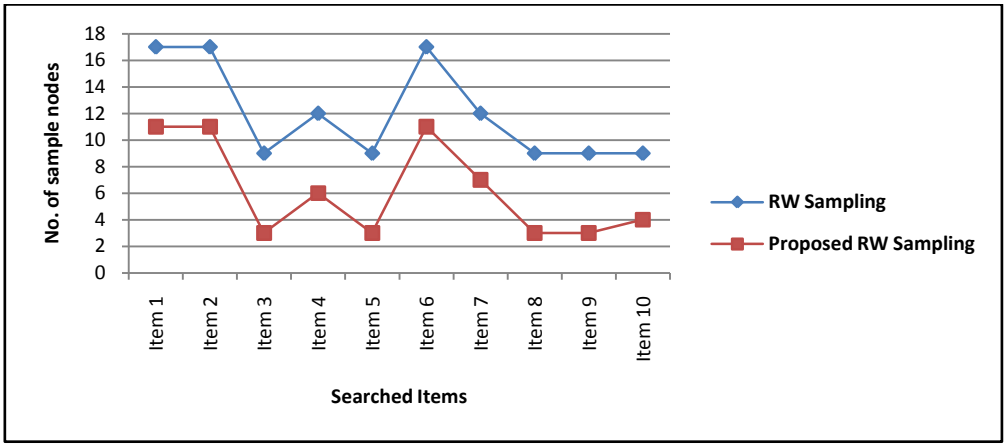

Fig2: No. of sample nodes taken by the existing and proposed RW sampling methods

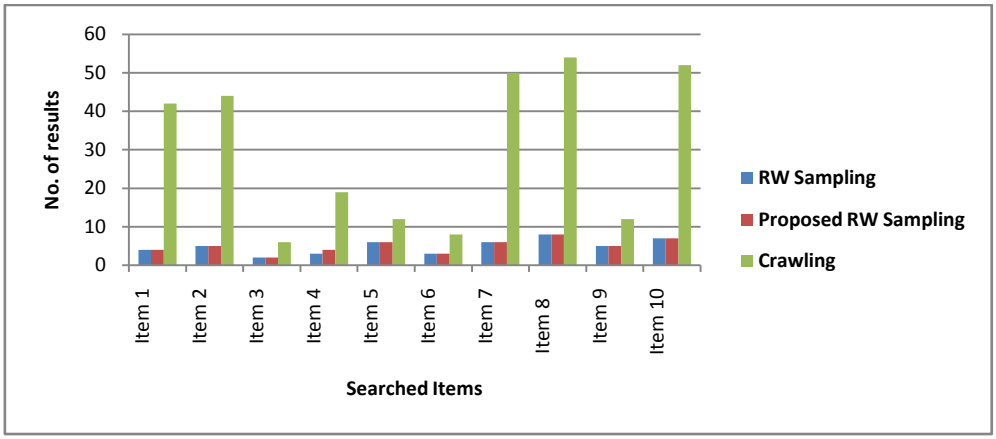

Fig 3: No. results obtained after searching using three methods

We get the following results from the graphs:

- Proposed randomwalk sampling method takes lesser time for searching, compare to existing randomwalk sampling method.
- Sampling algorithm reduces the time require for searching compare to the traditional crawling method.
- The no. of sample nodes taken by the proposed RW sampling method is lesser than the existing RW sampling method.
- The no. of results by the crawling method is very large as compared with other two methods. That means there are numerous unnecessary nodes presented in the search result.
- The no. of results of the two RW sampling methods are similar and it means that the proposed RW sampling method produced similar result to the existing RW sampling method but using less sample nodes and hence use lesser time. Hence the proposed RW sampling method is efficient than the existing RW sampling method.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

VI. CONCLUSION

The online social network contains numerous amounts of users. Searching methods have much relevance in the case of social networks. Visiting all the nodes for searching purpose is not a good practice because it takes long time and produce a large list as result in which most are unwanted. So an alternative method is proposed called sampling. The graphs shown that the sampling methods take lesser time for searching and produce lesser number of result compare to the crawling method. The existing random walk sampling method includes non-active nodes. So in the proposed method a time window is used to eliminate those nodes. The graphs shown that the number of samples taken by the proposed method is lesser compared with the existing method and produced similar amount of results within lesser time. So the proposed method is time efficient than the existing method.

REFERENCES

1.G. Das, N. Koudas, M. Papagelis, and Nick Koudas, "Sampling online socialnetworks," IEEE Transactions On Knowledge And Data Engineering, VOL. 25, NO. 3, March 2013.
2.A. Mislove, K.P. Gummadi, and P. Druschel, "Exploiting Social Networks for Internet Search," Proc. Fifth Workshop Hot Topics in Networks (HotNets), 2006.
3.S.A.M. Makki and G. Havas, "Distributed Algorithms for Depth- First Search," Information Processing Letters, vol. 60, no. 1, pp. 7-12, 1996.
4.D.E. Knuth, "Estimating the Efficiency of Backtrack Programs," Math. Of Computation, vol. 29, no. 129, pp. 121-136, 1975.
5.L. Katzir, E. Liberty, and O. Somekh, "Estimating Sizes of Social Networks via Biased Sampling," Proc. 20th Int'l Conf. World Wide Web (WWW), 2011
6.A.S. Maiya and T.Y. Berger-Wolf, "Sampling Community Structure," Proc. 19th Int'l Conf. World Wide Web (WWW), 2010.
7.J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2005.
8. Q. Wang and H. Jin, "Exploring Online Social Activities for Adaptive Search Personalization," Proc. 19th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2010.