ISSN 2320 2610
Volume 2, No.4, July - August 2013
International Journal of Multidisciplinary in Cryptology and Information Security
Available Online at http://warse.org/pdfs/ijmcis02242013.pdf

Sampling Schemes for Exploring Correlations of Users in Social Networking

Jyothsna B[1], Dr.R.V.Krishnaiah[2], Sri Lavanya Sajja[3]

[1]Assistant Professor CSE, JNTU Hyderabad, DRKCET Hyderabad, Andhra Pradesh, India
jyothsna912@gmail.com
[2]Professor CSE, JNTU Hyderabad, NIT Warangal, Andhra Pradesh, India
r.v.krishnaiah@gmail.com
[3]Assistant Professor CSE, JNTU Hyderabad, DRKIST Hyderabad, Andhra Pradesh, India
sslavanya79@gmail.com

ABSTRACT
Recent past has witnessed tremendous increase in social networking that provides a virtual platform to people to share information, meet friends online and use the platform for various activities. This cased an increased interest among research communities to utilize the huge amount of data being available for their information needs. The peers that involve in social networking have its needs for information. Papagelis et al. introduced sampling based approaches to obtain and explore a user's social network. They introduced some sampling schemes that are used to find correlations across the samples in centralized and distributed environments. In this paper we have implemented those schemes in order to help ranking algorithms for ranking neighborhood of a user. We also built a prototype application that demonstrates the proof of concept. The experiments are performed using real and synthetic data sets. The results revealed that the proposed system is effective and can be used in the real world.

Keywords*:* Social networks, query processing, sampling algorithms

1. INTRODUCTION

The trends have been changing in the usage of web technologies over Internet. Now the trend is to have social connectivity, information sharing and also expressing once self which led to the popularity of social networking web applications such as MySpace, Twitter, Face book, YouTube and so on. The information needs of social networking peers are changing from centralized access to ad hoc access. They need user's neighborhood information without knowing underlying network structure. However, obtaining interactions beyond contact list of a given user is not possible at present. However, there is lot of interest in this area to explore the possibilities. Many application domains need such information. Thus social search mechanism came into existence. By exploring information collection explicitly and implicitly in social networking can improve the accuracy of social search results. In a typical scenario a social search is given by user. The search engine returns relevant results. Then the result are ranked and presented by a global ranking algorithm. Mislove et al. [1] explored such result and its accuracy based on the global ranking criteria which are based on how many times the users in the social environment involved in the communications. For online user search many researches were carried out. The aim of all the researches is to bestow the users with accurate search results instead of providing a bulk of information. There are many approaches to know underlying network structure which are not user specific but on all users. For some applications it is useful to focus on the network pertaining to a single user. In order to achieve this very efficient algorithms are required as the user's neighborhood information is not static.

Recently Papagelis et al. [2] proposed various sampling based algorithms that can analyze a user's neighborhood information and collect it without having much knowledge about underlying network structure. They also introduced some variants of those schemes that minimize the number of nodes in the network to be visited in order to collect required information. The remainder of this paper is structured into the following sections. Section II provides review of relevant literature. Section III provides the proposed schemes. Section IV presents the prototype implementation details. Section V provides experimental results while section VI concludes the paper.

2. PRIOR WORKS

In this paper we focused on the sampling algorithms that can help collect information about dynamically changing neighborhood of a social networking user. This work is related to other works explored in [3], [4], [5], [6], [7], and

[8]. All these researches have a common thread. They start from a node and use random walk approach to navigate to neighbors. Based on the degree in the graph the probability of association with each node is inversely related. The stationary correlations can be converted to a uniform distribution using algorithm such as Metropolis Hastings Algorithm proposed in [9]. It can also be done using an approach known as rejection sampling. Similar approaches with some differences were found in [10], [11]. Biased sampling method is proposed in [12] by Katzir et al. to explore social community structures.

In [2] the generic random walk algorithms are improved further by sampling from neighborhood. Over a traditional web search results, a lot of improvement is found on providing personalized information [13]. Further the literature reveals that statistical models for extracting behavior in social networking improve quality of ranking [14], [15]. Other approaches followed notion of relevance feedback for improving ranking quality [16], [17]. Social search started with general and gradually moved towards personalized search. Then the personalization has been extended to realize human filter on search results [18]. Some researches explored temporal correlations based on given user's web history [19], [20] and tag-based logs [21]. However, our work in this paper is closely related to that of [2]

3. PROPOSED SAMPLING APPROACH

The aim of this approach is to let social networking peers to collect information pertaining to neighborhood of given user in such a way that the neighbor users endorsed an item of given user. This has to be achieved without the knowledge of underlying network structure. The proposed algorithms provide probabilistic network structure that can reflect the neighborhood of the given user. The network assumed in the solution is dynamic in nature. It does mean that the neighbors of given user are not static. Instead, those changes as new users endorse an item of user under study. The algorithms implemented in this paper practically are provided here.

```
procedure SAMPLE (u, n, d, C)
 T = NULL, sample s = 0, Sample[ 1…..n]
          while sample<=n do
                    if (v = randomWalk (u.d.C.T))! = 0 then
Sample [sample++] = v

                    end if
end while
end procedure

procedure randomWalk(u,d,C,T)
depth = 0, ps = 1
while depth < d do
pick v ϵ children (u) U u with pv = (degree (u) +1) -1
          if T U v has no cycle then
          add v to T
          ps = ps.pv
     if v = u then
                    accept with probability c/ps
                    if accepted then
                              return v
                    else
                              return 0
                    end if
          else
                     u = v, depth + +
          end if
end if
end while
return 0
end procedure
```

Figure.1 : Algorithm for Sampling Dynamic Social Networks

The sampling algorithm for dynamic social networks which collects underling network structure and also information based on the user information provided. The algorithm searches in all he child nodes randomly to sample the probable neighbor nodes which are someway related to given user. Especially the neighbor's endorsement towards the items of given user is taken as criterion.

```
1: procedure EVALSlNGLE (v.d.C.n.X)
2: S may of size n
3: Count array of size |X|
4: for all x ϵ X do
5:          S = SAMPLE (v.n.d.C)
6:          for all i ϵ S do
7:                    Count[x] = Count[x] + countix
8:          end for
9:  end for
10:  return Count
11:  end procedure
```

Figure. 2 : Algorithm for count estimation

The count estimation algorithm takes given node and other details and returns the count based on the sampling of dynamic social networks. This algorithm is meant for counts estimation for separate samples.

```
1: procedure EVALBATCH (v,d,C,n,X)
2: S array of size n
3: Count array of size IXI
4: S = SAMPLE (v,n.d,C)
```

5: for all i ϵ S do
6: for all x ϵ X do
7: Count[x] = Count[x] + countix;
8: end for
9: end for
10: return Count
11: end procedure

Figure. 3: Algorithm for counts estimation (same sample)
The count estimation algorithm takes given node and other details and returns the count based on the sampling of dynamic social networks. This algorithm is meant for counts estimation for same samples.

4. EXPERIMENTAL RESULTS

We have built a prototype application to demonstrate the efficiency of the proposed algorithms. The application is built to perform experiments different sample sizes. The datasets used are synthetic and real world.

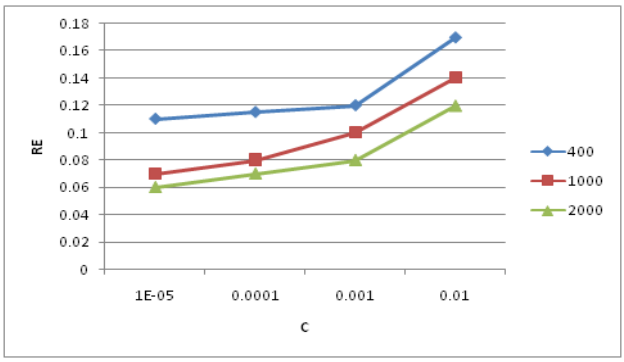

Figure. 4: Effect of C in sampling accuracy

As can be shown above figure 4 horizontal axis represents C while vertical axis represents RE values

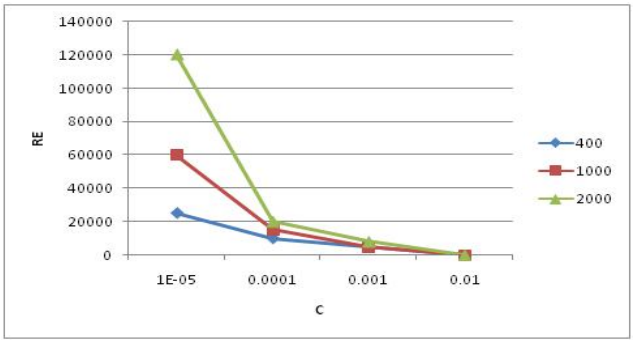

Figure.5: Effect of C in sampling cost.

As can be shown above figure 5 horizontal axis represents C while vertical axis represents RE values

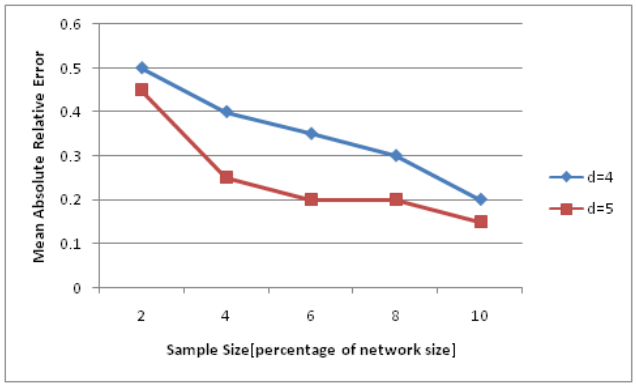

Figure.6: Network size estimation error

As can be shown above figure 6 horizontal axis represents sample size while vertical axis represents Mean absolute relative error.

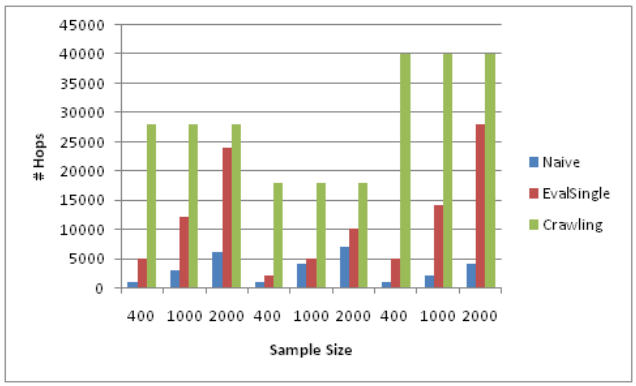

Figure.7: Sampling cost: Naive versus EvalSingle versus Crawling

As can be shown above figure 7 horizontal axis represents Sample size while vertical axis represents Hops.

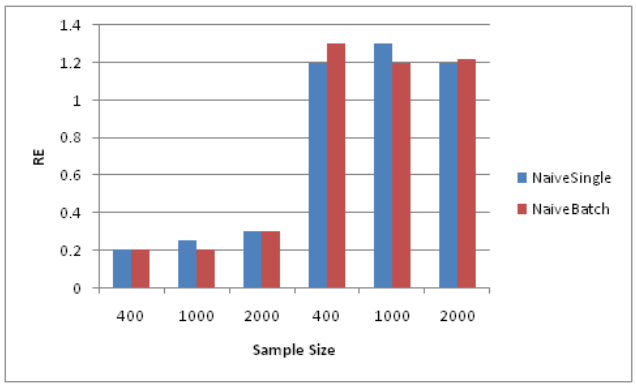

Figure.8: Batch sampling effect: naive sampling

As can be shown above figure 8 horizontal axis represents Sample size while vertical axis represents RE values.

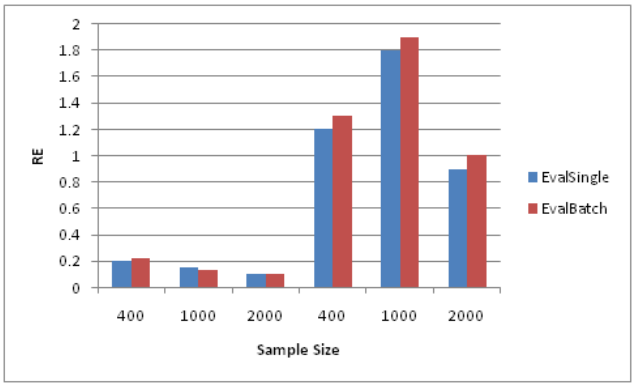

Figure.9: Batch sampling effect: our sampling.

As can be shown above figure 9 horizontal axis represents Sample size while vertical axis represents RE values

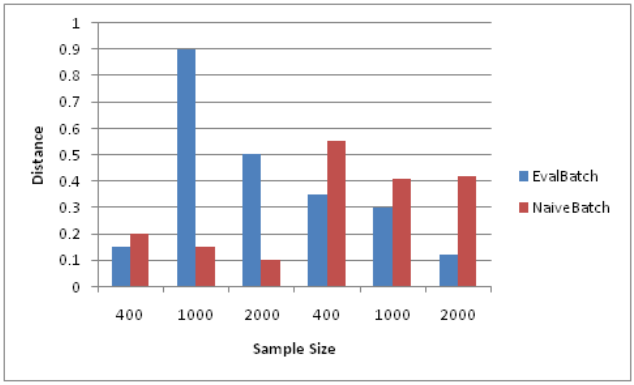

Figure.10: Ordering accuracy: distance between lists.

As can be shown above figure 10 horizontal axis represents Sample size while vertical axis represents Distance.

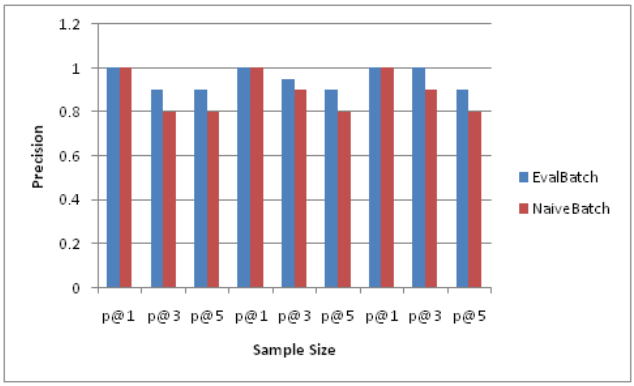

Figure.11: Ordering accuracy: precision at K for synthetic

As can be shown above figure 11 horizontal axis represents Sample size while vertical axis represents Precision

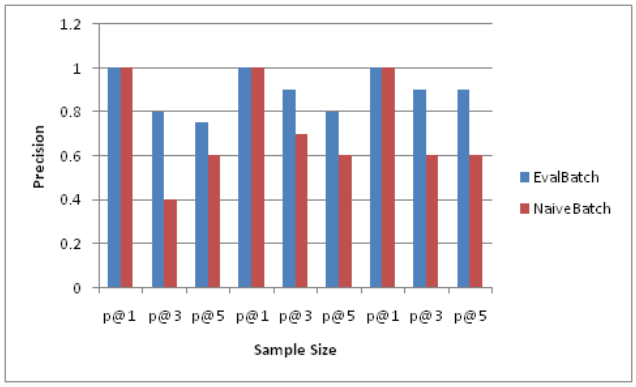

Figure.12: Ordering accuracy: precision at K for real network

As can be shown above figure 12 horizontal axis represents Sample size while vertical axis represents Precision.

5. CONCLUSION

Social networks have emerged as virtual platforms that serve users in various social aspects like online communication, information sharing, publishing and so on. Efficient information retrieval with respect to a user's social networking is required by many social networking peers. The proposed a set of sampling algorithms that can obtain a user's neighborhood information with having the knowledge of underlying structure of social network. In this paper we implemented the algorithms in order to provide efficient collection of user's information on the social networks like Facebook, Myspace, Twitter and so on. We built a prototype application for evaluating the concepts. The emperical results revealed that the proposed schemes are useful and support quick information collection of a user's neighborhood in social networking.

REFERENCES

[1] A. Mislove, K.P. Gummadi, and P. Druschel, "Exploiting SocialNetworks for Internet Search," Proc. Fifth Workshop Hot Topics inNetworks (HotNets), 2006.

[2] Manos Papagelis, Gautam Das and Nick Koudas, "Sampling Online Social Networks". IEEE Transactions on knowledge and data engineering, VOL. 25, NO. 3, MARCH 2013.

[3] C. Gkantsidis, M. Mihail, and A. Saberi, "Random Walks in Peerto-Peer Networks: Algorithms and Evaluation," PerformanceEvaluation, vol. 63, no. 3, pp. 241-263, 2006.

[4] M. Ajtai, J. Komlos, and E. Szemeredi, "Deterministic SimulationinLogspace," Proc. 19th Ann. ACM Symp. Theory of Computing(STOC), 1987.

[5] R. Impagliazzo and D. Zuckerman, "How to Recycle RandomBits," Proc. 30th Ann. Symp. Foundations of Computer Science(FOCS), 1989.

[6] D. Gillman, "A Chernoff Bound for Random Walks on ExpanderGraphs," SIAM J. Computing, vol. 27, no. 4, pp. 1203-1220, 1998.
[7] Z. Bar-Yossef and M. Gurevich, "Random Sampling from a SearchEngine's Index," Proc. 15th Int'l Conf. World Wide Web (WWW),2006.
[8] Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol, and D. Weitz,"Approximating Aggregate Queries About Web Pages viaRandom Walks," Proc. 26th Int'l Conf. Very Large Data Bases(VLDB), 2000.
[9] W. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," Biometrika, vol. 57, no. 1, pp. 97-109, 1970.
[10] G. Das, N. Koudas, M. Papagelis, and S. Puttaswamy, "EfficientSampling of Information in Social Networks," Proc. ACM WorkshopSearch in Social Media (SSM), 2008.
[11] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou, "Walkingin Facebook: A Case Study of Unbiased Sampling of Osns," Proc.INFOCOM, 2010.
[12] A.S. Maiya and T.Y. Berger-Wolf, "Sampling Community Structure,"Proc. 19th Int'l Conf. World Wide Web (WWW), 2010.
[13] J.-T. Sun, H.-J.Zeng, H. Liu, Y. Lu, and Z. Chen, "Cubesvd: ANovel Approach to Personalized Web Search," Proc. 14th Int'lConf. World Wide Web (WWW), 2005.
[14] J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search viaAutomated Analysis of Interests and Activities," Proc. 28th Ann.Int'l ACM SIGIR Conf. Research and Development in InformationRetrieval (SIGIR), 2005.
[15] E. Agichtein, E. Brill, and S. Dumais, "Improving Web SearchRanking by Incorporating User Behavior Information," Proc. 29thAnn.Int'l ACM SIGIR Conf. Research and Development in InformationRetrieval (SIGIR), 2006.
[16] Z. Dou, R. Song, and J.-R.Wen, "A Large-Scale Evaluation andAnalysis of Personalized Search Strategies," Proc. 16th Int'l Conf.World Wide Web (WWW), 2007.
[17] 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.
[18] D. Horowitz and S.D. Kamvar, "The Anatomy of a Large-ScaleSocial Search Engine," Proc. 19th Int'l Conf. World Wide Web(WWW), 2010.
[19] A. Papagelis, M. Papagelis, and C.D. Zaroliagis, "Iclone: TowardsOnline Social Navigation," Proc. ACM 19th Conf. Hypertext andHypermedia (HT), 2008.
[20] A. Papagelis, M. Papagelis, and C. Zaroliagis, "Enabling SocialNavigation on the Web," Proc. IEEE/WIC/ACM Int'l Conf. WebIntelligence and Intelligent Agent Technology (WI-IAT), 2008.
[21] R. Wetzker, C. Zimmermann, C. Bauckhage, and S. Albayrak, "ITag, You Tag: Translating Tags for Advanced User Models," Proc.ACM Third Int'l Conf. Web Search and Data Mining (WSDM), 2010.

JYOTHSNA BANDREDDI

Jyothsna Bandreddi is working as an Assistant Professor in DRK College of Engineering and Technology, JNTU Hyderabad University, Ranga Reddy, Andhra Pradesh, India.B.Tech from Nagarjuna University.Main Research interest includes Data Mining and Networking.

DR.R.V.KRISHNAIAH

Dr.R.V.Krishnaiah has received Ph.D from JNTU Anantapur, M.Tech(EIE) from NIT Warangal, M.Tech in CSE from JNTU Hyderabad, B.Tech from Nagarjuna University. He is a member of IETE, ISTE, IE and other Professional Organizations. He is Advisory Board Member for 12 Journals, Editorial Board Member for 4 journals and reviewer for many. He has published more than 50 papers in various Journals and Conferences. Currently he is working as Principal, DRK Institute of Science & Technology, Hyderabad, India.

SRI LAVANYA SAJJA

Sri LavanyaSajja is working as an Assistant Professor in DRK College of Engineering and Technology, Ranga Reddy, Andhra Pradesh, India. She has received M.Tech degree in Computer Science from JNTUH along with an M.Tech degree in IT. Her main research interest includes Data Mining and Networking.