A Survey of Online Social Network Data Mining

Teena Patidar[1], Prof. Avinash Sharma[2]

[1]*PG Scholar*, [2]*Associate Professor*

Department of CSE, MITS, Bhopal, India

Abstract - Online Social Network (OSN) mining has been a vast active area of research in the current years mainly due to the immense increase in the usage and popularity of such social networks. However, the huge amount of resources available in Online Social Networks (OSNs) poses a great challenge for researchers to analyze such networks. Also, the data generated from OSNs is dynamic and requires the need for intelligent mining of dynamic OSN data. In this paper we take into consideration three important current topics related to OSN mining with special emphasis on the current status of these research subareas. Hence this paper gives an idea about the hot topics related to OSN mining which will help the researchers to solve those challenges that still exist in mining OSNs.

Keywords: Online Social Networks, Data Mining, Influence Propagation, Community Detection, Link Prediction.

1. INTRODUCTION

In recent years the increasing tendency of people to use Online Social Networks (OSNs) such as Facebook, Twitter, and LinkedIn, have resulted in making different kinds of interactions and relationships which, in turn, have lead to the generation and availability of a huge amount of valuable data that has never been available before. Such huge valuable data can be used in some new, varied, eye-catching, and useful research areas to researchers. Although this research area has received a great deal of attention in the last few years, yet many problems related to mining OSNs is still in its infancy and needs more techniques to be developed in the future for further improvement.

However, since data generated from OSN is vast, noisy, distributed and dynamic, this requires appropriate data mining techniques to analyze such large, complex, and frequently changing social media data. Research is being carried out related to various issues in OSN mining such as, influence propagation, expert finding, recommender systems, link prediction, community detection, opinion mining, mood analysis, prediction of trust and distrust among individuals, etc which is depicted in Fig. 1 below. In this section, we introduce only three representative research issues in mining online social networking sites, namely, influence propagation, community detection and link prediction in OSNs, and give a detailed overview of the related work and current status of these issues.

1.1 Influence Propagation

Nowadays, as OSNs are attracting millions of people, the latter rely on making decisions based on the influence of such sites [1]. For example, influence propagation can help make a decision on which product to purchase, which audio/video to watch, which community to join, and so on. Thus, influence propagation has become vital for effective viral marketing, where companies can convince customers to buy products through the help of those active people in OSNs who can play a key role in influencing others. Hence requires the need for researchers to study influence propagation in OSNs which is currently a hot topic related to OSN mining functionalities.

1.2 Community or Group Detection

Community or group detection in OSNs is based on studying the OSN structure to cluster individuals into groups by finding which individuals correlate more with each other than with other users. Such detection of communities can help to further make an assessment about what products, services and activities an individual might be interested in. Hence, discovering of such OSN community structures is of prime importance which attracts the research community in data mining to make an in depth study of community detection in OSNs.

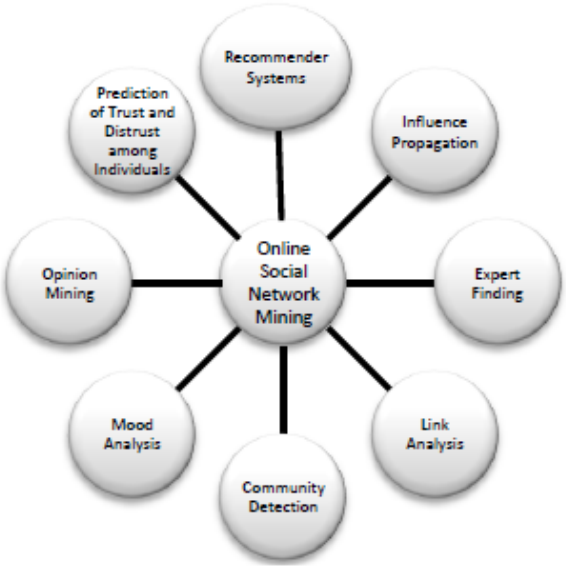

Fig 1 Some Key Research Issues in Online Social Network Mining

An illustration of three OSN communities is depicted in Fig. 2. Two nodes can be connected by an edge if and only if a relationship exists between them according to social definitions such as their participation or role in the OSN. Connections of nodes in this case are binary. Such representation of communities can be best explained with the help of graphs. In fact, one of the most applicable features of graphs for representing real systems is community structure.

1.3 Link Prediction

Link prediction in OSNs is all about predicting the possibility of a future association between two nodes, knowing that there is at present no association between these nodes. This is possible by mining the bulk amount of data available in OSNs to find out „who is a friend of whom" or „which products have an association with other products". By analyzing and gathering useful information about an individual or product, OSNs can infer new interactions among members or nodes of an OSN that are likely to occur in the near future. Thus, link prediction is a key research subarea for OSN mining.

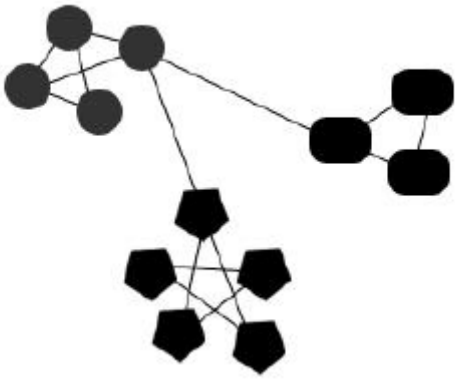

Fig 2 An Online Social Network Depicting Three Communities

2. BACKGROUND AND RELATED WORK

The previous works on OSN analysis reveals that intelligent data mining techniques can help effectively deal with the several challenges related to OSN data. Such challenges pose critical challenge to handle OSN data. First, OSN data sets are very large. For example, if we consider the millions Facebook users, analyzing such a network requires complex graph mining techniques. Second, OSN data sets can be noisy. For example, the unwanted or insignificant tweets in Twitter that need not be taken into consideration during OSN mining. Third, data from OSNs is dynamic, i.e., changes and updates occur frequently in a flicker of a second which poses a great challenge for researchers in dealing with such OSN data. Considering these challenges being faced in research related to OSN mining, this section describes the related work covered in the field of three research sub-areas of OSN mining, namely influence propagation, community detection and link prediction in OSNs.

2.1 Influence Propagation

The basic computational problem in influence propagation is that of selecting a set of initial users that are more likely to influence the other users of an OSN. The first algorithmic treatment of this problem was provided by Domingos and Richardson in [8] and [13] by giving a heuristic solution to the problem. Kempe et al. [18] have further proposed two elementary influence propagation models namely, the Independent Cascade (IC) model and the Linear Threshold (LT) model. These models consider a node to be either active or inactive for a given timestamp. An active node can be a customer who already purchased a product or an adopter of the innovation, and so on. Also, each node"s tendency to become active increases monotonically as more of its neighbors becomes active. Leskovec et al. [2] developed an efficient algorithm called Cost-Effective Lazy Forward selection (CELF) based on a "lazy-forward" optimization, which though is much faster than a simple greedy algorithm, yet takes long time to find even 50 most influential nodes in a network. In [16], Saito et al. proposed an Expectation Maximization (EM) algorithm that uses simple shortest path technique to propagate influence in a network. However, their approach did not deal with topic-wise influence propagation of OSNs. Chen et al. [19] proposed a new heuristic algorithm to estimate influence spread under the IC model. More recently, Chen et al. [17] proposed a scalable heuristic called Local Directed Acyclic Graph (LDAG) for the LT model but their model did not deal with various social media sites for influence propagation. Goyal et al. [15] also studied the problem of learning influence probabilities. Few papers such as [23] and [25] have studied influence propagation in social networks from the topic modeling perspective, i.e., analyzing the heterogeneous link information as well as the textual content associated with each node in the OSN to mine topic-level direct influence.

2.2 Community or Group Detection

Most of the OSNs exhibit strong community structure. A key research agenda thus is to categorize communities of interest and study their behavior over time. However, community detection in OSNs is a challenging job for traditional community detection algorithms, such as hierarchical clustering, spectral clustering and partitional clustering algorithms, due to the extraordinary large scale of the network. These algorithms only consider links and do not take into account other node properties and user interactions in the social graph. Many of these algorithms also do not allow users to have membership in multiple communities. Some other previous approaches in detecting

community structures also use only the social structure among people to discover communities. Such approaches can be classified into optimization based methods and heuristic methods. The heuristic methods are quite popular and often use a graph clustering algorithm for community detection. Few examples of heuristic-based algorithms include Maximum Flow Community (MFC) algorithm [12], Hyperlink Induced topic search (HITS) algorithm [10], Clique Percolation Method (CPM) algorithm [7], and Finding and Extracting Communities (FEC) algorithm [9]. However recent work suggests that community detection can be dealt with based on topic modeling approach which involves analyzing the content of social objects. Few examples of such topic models include Probabilistic Latent Semantic Analysis (pLSI) [26], Latent Dirichlet Allocation (LDA) [14], and Author Topic (AT) [27] models. Recently, the Community-Author-Recipient-Topic (CART) model [28] was designed and this model was one of the first attempts to combine link-based community detection with content based community detection. However, CART takes into account only recipients of every post which may not be feasible for discovering communities in OSNs such as Twitter which allow broadcasts. In [24], the Topic-link LDA model was developed which can draw latent topical and community distributions for every node in document networks. In [20], Sachan et al. have developed Topic-User-Community-Models (TUCM) which also use topics, social graph topology and nature of user interactions to discover latent communities in social graphs. However, the models designed by Sachan et al. take significantly longer time for an expected output.

2.3 Link Prediction

The research for link prediction in OSNs tries to infer new interactions among members of a social network that are likely to occur in the near future. The applications of link prediction are varied and include friends suggestion, terrorist network monitoring and building of recommender systems. In [4], one of the earliest link predictions models were proposed that works explicitly on a social network. Here, two main approaches have been developed to handle the link prediction problem. The first one is based on local features of a network, focusing mainly on the local nodes‟ structure. Examples of few local-based approaches of link prediction include Common Neighbors Index or else known as Friend Of A Friend (FOAF) algorithm [3], Jaccard Coefficient [4], Adamic/Adar Index [21] and FriendLink algorithm [22]. The second approach on link prediction is based on global features, which detects the overall path structure in a network.

Examples of such global-based link prediction algorithms include Katz [11] Status Index, Random Walk with Restart (RWR) algorithm [5] and SimRank algorithm [6]. However, the techniques for link prediction can be varied such as feature-based classification and kernel-based to matrix factorization and probabilistic graphical models. These techniques differ from each other with respect to scalability, generalization ability, model complexity and prediction performance. The link prediction problem can also be studied in the context of relational data where explicit graph representations are not used. Hence, graphical models, probabilistic relational models, stochastic relational models and different variants of these are the main modeling paradigm used in link prediction problem.

3. CURRENT STATUS OF THESE RESEARCH ISSUES

As explained earlier, OSN data are largely user-generated content on social media sites which are vast, noisy, distributed, unstructured, and dynamic. These characteristics pose challenges to data mining tasks to invent new efficient techniques, algorithms and frameworks. It can be understood from the previous discussions that OSN mining is a huge area, and it is not possible to study every aspect of it in limited time period. The related work of the research issues mentioned in the previous section reveal that novel algorithms and frameworks for these OSN mining functionalities need to be developed to generate better and faster results. A brief discussion on the current status of these research subareas is done below:

3.1 Influence Propagation

Most existing works on influence propagation in OSNs have focused on verifying the existence of social influence. Few works systematically investigate how to mine the strength of direct and indirect influence between nodes in heterogeneous networks [8]. Very few papers such as Tang et al. [29], Liu et al. [23], Weng et al. [30], and Lin et al. [31], have considered influence propagation from the topics perspective. But in none of these papers a topic-wise influence model has been designed for generating influential users in OSNs. However, in [25], Barbieri et al. have discussed about topic-wise influence propagation models namely, Topic-aware Independent Cascade (TIC) model, Topic-aware Linear Threshold (TLT) model and Authoritativeness-Interest-Relevance (AIR) model. Thus, for influence propagation, new models and techniques need to be developed to deal with topic-aware social influence propagation that can provide more accurate predictions and can extend the focus on influence propagation to other application domains such as recommender systems. The key factors to take into consideration while designing a new propagation model is that the model should not rely on very large number of parameters and should be able to deal with the problem of scalability of large networks.

3.2 Community or Group Detection

Most of the previous works in community detection of OSNs have dealt with either link-based community

detection methods or topic-based community detection methods. The main drawback in case of link-based community detection methods is that these methods can reflect the strength of connections but cannot reflect the semantics such as being tagged together in several photos, liking similar movies, or interesting topics shared by people. Again, the main drawback in case of topic-based community detection methods is that these methods can generate communities that are topically similar but cannot reflect the strength of connections available in an OSN. Thus, the previous works on this research subarea reveal that the community detection methods for OSNs is either topic blind or not fit for OSNs. Few methods have been developed to deal with community detection both from link and topic perspective. However, majority of them are not fit for OSNs. This demands the data mining research community to build novel hybrid topic-link community or group detection models which can ease in finding which people are attracted to which community, and by what topic. The model should be such that it can be applied to many kinds of OSNs which contain social objects.

3.3 Link Prediction

In recent years, research works on link prediction in OSNs have evolved over various aspects. But still, a critical concern is the scalability of the proposed solutions for link prediction. Also, most existing works on link prediction have focused on building a model based on studying only the network structure of OSNs. There are few link prediction techniques that have considered building a time-aware link prediction model which is also feature-based and studies the categorical attributes of nodes. However, novel algorithms or techniques can still be developed that can provide more accurate and faster link predictions by considering not only the friendship network but also other implicit social networks which forms due to users‟ daily interactions such as co-rating similar products, commenting on an institution where both individuals have studied, and so on. Thus, as the usage of OSNs is constantly increasing on a daily basis, there arises the necessity to critically analyze and understand such networks in an efficient manner.

4. CONCLUSIONS

Never has the social nature of human beings been more evident than today with the world-wide ubiquitous use of social media. OSN mining can reveal insights into the social nature of users of OSNs and is poised to help us change the way we see society as a whole and as individuals. As the number of social media users continues to grow, we will likely continue to see significant changes in the way we communicate and share information with each other. In this regard, the research community needs to continue to look to data mining approaches to provide us with the empowering ability to look deeper into these large data sets generated from OSNs in a more meaningful way. Also, researchers still need to continuously focus on several critical issues revolving around OSN mining for attaining a better solution for the same. This paper has contributed a step forward in the direction of thought on OSN mining to address several vital issues that still revolve around in mining OSNs.

REFERENCES

[1] Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip S. Yu and Ming-Syan Chen, “A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining”, IEEE Transactions on Knowledge and Data Engineering, 2017.

[2] G. Nandi and A. Das, “A Survey on Using Data Mining Techniques for Online Social Network Analysis”, in the International Journal of Computer Science Issues, November 2013, vol. 10, issue 6, pp. 162–167.

[3] SJ. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Briesen and N. S. Glance, “Cost-effective outbreak detection in networks”, in Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 420–429, 2007.

[4] J. Chen, W. Geyer, C. Dugan, M. Muller and I. Guy, “Make new friends, but keep the old: recommending people on social networking sites”, in Proceedings of the 27th International Conference on Human factors in Computing Systems, pp. 201–210, 2009.

[5] D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks”, in Proceedings of the 12th International Conference on Information and Knowledge Management, pp. 556–559, 2003.

[6] J. Pan, H. Yang, C. Faloutsos and P. Duygulu, “Automatic multimedia cross-modal correlation discovery”, in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 653–658, 2004.

[7] G. Jeh and J. Widom, “SimRank: a measure of structural-context similarity”, in Proceedings of the 8thACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543, 2002.

[8] G. Palla, I. Derényi, I. Farkas and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society”, Nature, http://dx.doi.org/10.1038/nature 03607, vol. 435, no. 7043, pp. 814-818, 2005.

[9] P. Domingos and M. Richardson, “Mining the network value of customers”, in Proceedings of the 7th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 57–66, 2001.

[10] B. Yang, W. Cheung and J. Liu, “Community mining from signed social networks”, In IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 10, pp. 1333–1348, 2007.

[11] J. Kleinberg, "Authoritative sources in a hyperlinked environment", In Journal of the ACM, vol. 46, no. 5, pp. 604–632, 1999.

[12] L. Katz, "A new status index derived from sociometric analysis", In the Journal of Psychometrika, vol. 18, no. 1, pp. 39–43, 1953.

[13] G. Flake, S. Lawrence, C. Giles and F. Coetzee, "Self-organization of the web and identification of communities", In IEEE Computer, vol. 35, no. 3, pp. 66–71, 2002.

[14] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 61-70, 2002.

[15] D. Blei, A. Ng and M. Jordan, "Latent dirichlet allocation", in the Journal of Machine Learning Research, pp. 993–1022, 2003.

[16] A. Goyal, F. Bonchi and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," in Proceedings of the VLDB Endowment, vol. 5, no. 1, pp. 73-84, 2011.

[17] K. Saito, R. Nakano and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in Proceedings of the 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES), pp. 67-75, 2008.

[18] W. Chen et al, "Scalable influence maximization in social networks under the linear threshold model", in Proceedings of the 2010 IEEE International Conference on Data mining (ICDM), pp. 88–97, 2010.

[19] D. Kempe, J. M. Kleinberg and ´E. Tardos, "Maximizing the spread of influence through a social network", in Proceedings of the 9thACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD¨03), pp. 137-146, 2003.

[20] W. Chen, C. Wang and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD), pp. 1029-1038, 2010.

[21] M. Sachan, D. Contractor, T. A. Faruquie and L. V. Subramaniam, "Using Content and Interactions for Discovering Communities in Social Networks", in Proceedings of the 21st International Conference on World Wide Web, pp. 331-341, 2012.

[22] L. Adamic and E. Adar, "How to search a social network", In the Journal of Social Networks, vol. 27, no. 3, pp. 187–203, 2005.

[23] A. Papadimitriou, P. Symeonidis and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems", In the Journal of Systems and Software, vol. 85, pp. 2119-2132, 2012.

[24] L. Liu, J. Tang, J. Han, M. Jiang and S. Yang, "Mining topic-level influence in heterogeneous networks," in Proceedings of the 19thACM International Conference on Information and knowledge management (CIKM), pp. 199-208, 2010.

[25] Y. Liu, A. Niculescu-Mizil and W. Gryc, "Topic-link lda: joint models of topic and author community", in Proceedings of the 26thAnnual International Conference on Machine Learning, pp. 665-672, 2009.

[26] N. Barbieri, F. Bonchi and G.Manco, "Topic-aware Social Influence Propagation Models", in Proceedings of the 2012 IEEE 12thInternational Conference on Data Mining (ICDM), pp. 81-90, 2012.

[27] T. Hofmann, "Probabilistic latent semantic indexing", in Proceedings of the22nd Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR), pp. 50–57, 1999.

[28] M. Steyvers, P. Smyth, M. Rosen-Zvi and T. Griffiths, "Probabilistic author-topic models for information discovery", in Proceedings of the 10thACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 22–25, 2004.

[29] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson, "Social topics models for community extraction", in Proceedings of the 2nd International Workshop on Advances in Social Network Mining and Analysis (SNAKDD), pp, 77-96, 2008.

[30] Jie Tang , Jimeng Sun , Chi Wang , and Zi Yang, "Social influence analysis in large-scale networks", in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009.

[31] Jianshu Weng, Ee-Peng Lim, Jing Jiang, Qi He, "TwitterRank: finding topic-sensitive influential twitterers", in Proceedings of the 3rd ACM international conference on Web search and data mining, 2010.

[32] Cindy Xide Lin , Qiaozhu Mei , Jiawei Han , Yunliang Jiang , and Marina Danilevsky, "The Joint Inference of Topic Diffusion and Evolution in Social Communities", in Proceedings of the 2011 IEEE 11th International Conference on Data Mining", pp.378-387, 2011.