IJCAT - International Journal of Computing and Technology, Volume 1, Issue 10, November 2014
ISSN : 2348 - 6090
www.IJCAT.org

Data Mining Techniques for Online Social Network Analysis

[1] Aditya Kumar Agrawal, [2] Shikha Kumari, [3] B.Giridhar, [4] Bhavani Shankar Panda

[1, 2] B.Tech Student,

[2, 3] Asst.prof in CSE dept

Abstract - In this paper we take into consideration the concepts of using algorithmic and data mining perspective of Online Social Networks (OSNs), with special emphasis on latest hot topics of research area. There are several factors which has made the study of OSNs gain enormous importance by researchers. Few such factors include the availability of huge amount of OSN data, the representation of OSN data as graphs, and so on. Analysis of data in OSNs also has a great prospective for researchers in a variety of disciplines. Hence this paper gives an idea about the key topics of using data mining in OSNs which will help the researchers to solve those challenges that still exist in mining OSNs.

Keywords **- Online Social Networks, Data Mining, Structure based Analysis, Content-based Analysis**

1. Introduction

With the advent of Online Social Networks (OSNs), a revolutionary change has occurred in the social interactions of people of this decade. Many popular OSNs such as Facebook, Orkut, Twitter, and LinkedIn have become increasingly popular. Nowadays, these OSNs allow many easy-to-learn online activities including chatting, online shopping, gaming, tweeting, etc. According to the site *thenextweb.com*, Indian citizens spend one in four minutes online using social networking sites, more than any other Internet activity [1]. In fact, social networking is considered to be the second-fastes growing activity, behind only entertainment.

However, social media sites provide data which are vast, noisy, distributed and dynamic. Hence, data mining techniques provide researchers the tools needed to analyze such large, complex, and frequently changing social media data. In this section, we introduce some representative research issues in mining social networking sites using data mining techniques as shown in Figure 1.

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. For example, influence propagation can help decide which movie to watch, which product to purchase, and so on. Thus, influence propagation has become an important mechanism for effective viral marketing, where companies try to promote their products and services through the word-of-mouth propagations in OSNs. This further motivates the research community to carry out extensive studies on various aspects of the influence propagation problem.

1.2 Community or Group Detection

In general, group detection in OSNs is based on analyzing the structure of the network and finding individuals that correlate more with each other than with other users. Clustering an individual in a particular way can help to further make an assessment about the individual such as what activities, goods, and services, an individual might be interested in.

1.3 Expert Finding

OSNs consist of several experts in a specific domain and other people who join the network to receive help from these experts. These OSNs can be used to search for such experts within a group of people. For example, a topic related expert can be searched based on the study of the link between authors and receivers of emails.

1.4 Link Prediction

The bulk amount of data available in OSNs can be mined to make predictions about *'who is a friend of whom*' as an individual might be only a few steps away from a desirable social friend but may not realize it. By gathering useful information about an individual, OSNs can infer new interactions among members of an OSN that are likely to occur in the near future.

1.5 Recommender Systems

Recommender systems (RS) provide recommendations to users about a set of articles or services they might be

interested in. This facility in OSNs has become very popular due to the easy access of information on the Internet. Few important applications of RS are its use in several websites for recommendation of items such as movies, books, gadgets, etc.

1.6 Predicting Trust and Distrust among Individuals

Due to the continuous expansion of communities in OSNs, the question of trust and distrust among individuals in a community has become a matter of great concern. Past assessments reveal that some users try to either disturb or take undue advantage of the normal atmosphere of such online communities. As a result, there arises a need of assessing each user of an OSN community to predict the level of trust or distrust that can be computed for them.

1.7 Behavior and Mood Analysis

Discovering human behavior or human interaction based on data mining techniques is also an interesting research field that is gaining huge attention in research. Here, human behavior may indicate any human-generated actions such as clicking on a specific advertisement, accepting a friend's request, joining a group or discussion forum, commenting on an image, music, etc, or dating with a person, etc.

1.8 Opinion Mining

OSNs have given rise to various review sites, blog repositories, online discussions, etc where people can express their ideas and opinions, exchange knowledge and beliefs, criticize products and ideas. Data mining of opinions on specific subjects allows the detection of user prospects and needs, and also feelings or reactions of people about certain beliefs, products, decisions or events.

Some other important data mining applications related to OSNs include information and activity diffusion, topic detection and monitoring, marketing research for businesses, data management and criminal detection. These applications are also gaining huge interest in the research community. Thus, it can be concluded that OSN data analysis has a great prospective for researchers in a diversity of disciplines.

2. Current Status of these Research Issues

As the demand and usage of OSNs are increasing on a daily basis, there arises the necessity to critically analyze and understand such networks in an efficient manner. There is a constant radical change creeping into the OSN research community in the way analysts are interpreting and characterizing OSNs. The current status of the abovementioned research issues related to OSNs is discussed next:

2.1 Influence Propagation

Domingos and Richardson [6] provided the first algorithmic treatment to deal with influence propagation problem. Then, Kempe et al. [4] studied influence propagation by focusing on two fundamental propagation models, named *Independent Cascade (IC) Model* and *Linear Threshold (LT) Model,* which led to the development of the *Greedy Algorithm* for influence maximization. However, their model is not scalable to large networks. Leskovec et al. [3] dealt with the influence propagation problem from a different perspective namely outbreak detection but their method too faced serious scalability problems [5]. Chen et al. proposed a new propagation model similar to the greedy algorithm but with a better efficient result [7, 8]. Saito et al. [9] were the first to study how to learn the probabilities for the IC model from a set of past propagations. Goyal et al. [10] also had made a study of the problem of learning influence probabilities using an instance of the *General Threshold Model*. Barbieri et al. [11] considered the study on influence propagation from a topic modeling perspective.

2.2 Community or Group Detection

A comparative analysis on various community detection algorithms can be found in [12]. Initial study on community or group detection was focused mainly on the link structure of OSNs while ignoring the content of social interactions, which is also crucial for precise and meaningful community extraction. It is only recently that few researchers have addressed the problem of discovering topically meaningful communities from an OSN. McCallum et al. [14] were the first to have presented the *Author-Recipient-Topic model*, a Bayesian network for social network analysis that discovers discussion topics conditioned on the sender-recipient relationships. Pathak et al. [17] have proposed a *Community-Author-Recipient- Topic (CART) model* which uses both link and content information for community detection. Liu et al. [16] also have built a model based on *Topic-Link Latent Dirichlet Allocation (LDA)* but which works only with document networks. Zhao et al. [13, 15] have addressed topic oriented community detection through social objects and link analysis in social networks. Sachan et al. [18] have proposed *Topic User Community Model (TUCM)* as well as *Topic User Recipient Community Model (TURCM)*which offer topic modeling capabilities but takes a significantly longer time for result.

2.3 Expert Finding

Study on expert ranking algorithm is usually based on either domain knowledge driven methods or domain

knowledge independent methods or both. The expert ranking problem is also researched on email communication relations [19]. Zhang et al. [20] have proposed propagation based approach as well as a mixed approach based on *Probabilistic Latent Semantic Analysis (PLSA)* [21] for expert finding in social networks. Authors in [22] have used the *RarestFirst* and *EnhancedSteiner* algorithms for expert finding while authors in [23] have modified the *RarestFirst* algorithm and found the *Simplified RareFirst* (SRareFirst) algorithm. Smirnova et al. [24] have proposed a user-oriented model for expert finding based on rational user behavior. Jin et al. [25] found the *ExpertRank* algorithm which is based on analyzing closeness and authority for ranking expert in social networks.

2.4 Link Prediction

Liben-Nowell and Kleinberg [26] have dealt with link prediction in social networks but which works with only a static snapshot of a network. Hasan et al. [27] have proposed several classification models for link prediction which provides a comparison of several features by showing their rank of importance as obtained by different algorithms. Fouss et al. [28] have presented a link prediction technique based on a *Markov-chain model* of random walk but which does not scale well for large databases. Zheleva et al. [29] have used a binary classification algorithm in which family relationships were used for link prediction. Tylenda et al. [30] have proposed time-aware and time-agnostic maximum entropy methods for link prediction but have tested data sets only from scientific collaboration networks. Chen et al. [31] have made a detailed study and comparison of four different algorithms for link prediction. Schifanella et al. [32] have proposed a sampling link-prediction algorithm which can help users find friends with similar topical interests. Papadimi-triou et al. [2] presented a paper on fast and accurate link prediction in social networking systems but which considers only friendship network and no other features for link prediction.

2.5 Recommender Systems

Recommender systems (RS) have developed in parallel with the web. A good survey on various RS can be found in [33]. They were initially based on demographic, content-based and collaborative filtering. Collaborative filtering is the most common technique used for RS [34, 35]. Linden et al. [36] presented their work on item-toitem collaborative filtering for amazon.com recommendations. However, the evolution of RS has shown the importance of hybrid techniques of RS, which merge different techniques in order to get the advantages of each of them. A survey focused on the hybrid RS has been presented in [37] but it does not deal with the role of social-filtering, a technique which has become more popular in the recent years through social networks.

Hybrid filtering techniques can use mixed collaborative and content-based filtering [38, 39, 40], or can use mixed collaborative and demographic filtering techniques [41].

2.6 Predicting Trust and Distrust among Individuals

A number of disciplines have looked at various issues related to trust. One of the first works on this task was the *EigenTrust* algorithm [42] that aims to reduce the number of inauthentic file downloads in a P2P network. Guha et al. [43] proposed methods of propagation of trust and distrust, each of which is appropriate in certain circumstances. *PowerTrust* [44] is a trust recommendation system that
aggregates the positive and negative opinions between the users into the local trust scores, similarly to *EigenTrust*. Other work that studies a social network with positive and negative opinions is presented in [45]. DuBois et al. [46] presented a paper for predicting trust and distrust based on path probability in random graphs. Kim et al. [47] have also proposed a method of predicting trust and distrust of users in online social media-sharing communities. Ortega et al. [48] proposed a novel system intended to propagate both positive and negative opinions of the users through a network, in such way that the opinions from each user about others influence their global trust score.

2.7 Behavior and Mood Analysis

Benevenuto et al. [49] measured the behavior of online social networks' users applying the proxy server-based measurement framework. Schneider et al. [50] also have conducted an in-depth analysis of user behavior based on network traces across several online social networks. Gyarmati et al. [53, 54] crawled the public part of users' profile pages, which contained online status information of the users. Simoes et al. [51] proposed distance, similarity, influence and adjustments-based methods for understanding and predicting human behavior for social communities. Zhang et al. [52] have developed a model called *socioscope* for predicting human-behavior in social networks. Yan et al. [55] also have presented a social network based human dynamics model to study the relationship between the social network attributes of microblog users and their behavior. However, because of the diversity and complexity of human social behavior, no one technique will detect every attributes that arises when humans engage in social behaviors.

2.8 Opinion Mining

Most of works in this research area focus on classifying texts according to their sentiment polarity, which can be positive, negative or neutral [56]. Authors in [57] provided an in-depth survey of opinion mining and sentiment analysis. In [58], the problem was studied

further using supervised learning by considering contextual sentiment influencers such as negation (e.g., not and never) and contrary (e.g., but and however). Wilson et al. [59] have studied several different learning algorithms such as boosting, rule learning, and support vector regression that can automatically distinguish between subjective and objective (neutral) language and also among weak, medium, and strong subjectivity. Zhang et al. [60] presented a novel model that unifies topic-relevance and opinion generation by a quadratic combination. Zafarani et al. [61] studied sentiment propagation in social network by making a case study of *LiveJournal* website. In [62], a method was proposed to deal with the problem of product aspects which are nouns and imply opinions using a large corpus. Authors in [63] have studied about several challenges in developing opinion mining tools for social media. Ortigosa et al. [64] developed a hybrid approach for performing sentiment analysis in Facebook with high accuracy.

3. Using Data Mining in OSNs:

Data available in OSNs are basically user-generated content which are vast, noisy, distributed, unstructured, and dynamic. These inadvertent characteristics pose challenges to data mining tasks to develop efficient algorithms and techniques. The following are some key points that need strong consideration while using data mining techniques for OSNA:

A. Structure-based Analysis versus Content-based Analysis:

As far as research in OSNA is concerned, a good amount of work has been done based on structure analysis of OSNs where the linkage structure is taken into consideration in order to gather interesting characteristics of the underlying network. However, some recent research has shown that content-based analysis of OSNs can yield valuable insights about the underlying social network. Hence researchers need to concentrate on using the concept of structure as well as content for making an analysis of OSNs.

B. Dynamic Analysis:

Past research on OSNs has mainly treated the network to be static. However, the fact remains that OSNs are dynamic and hence, to improve the quality of the results, a major amount of work yet remains to be done on dynamic analysis of OSNs which can evolve rapidly over time.

C. Adversarial Networks:

Study of adversarial networks (such as terrorist network) requires attention by researchers in which the relationship among the different adversaries may not be clearly known. Thus, such kind of complex heterogeneous OSNs require designing of new tools and techniques which can resourcefully analyze the network. In this regard, graphbased data mining can play a major role for such OSN analysis.

4. Conclusion

The analysis of OSN data though has its solid basis in graph theory, yet it is still in its infancy. Researchers still need to focus on these critical social network issues, taking into consideration an algorithmic and data mining perspective for attaining a better solution for the same. There are also strong motivations for efficiently propagating the right information to the right people via OSNs and which has become a research area of increasing importance. Thus, this survey paper is a step forward in the direction of evolving new data mining techniques to address the above mentioned critical online social networking issues.

References

[1] Josh Ong (August 2012), article retrieved from http://thenextweb.com/in/2012/08/20/social-networkingites- now-occupy-25-online-time-india/

[2] Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems", The Journal of Systems and Software 85, 2012, pp. 2119–2132

[3] Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'07), 2007, pp. 420–429

[4] Kempe, J. M. Kleinberg, and ´E. Tardos, "Maximizing the spread of influence through a social network," in Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03), 2003, pp.137-146

[5] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks", in Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09), 2009.

[6] M. Richardson and P. Domingos, "Mining knowledgesharing sites for viral marketing," in Proc. of the 8th ACM SIGKDD Int. Conf on Knowledge Discovery and Data Mining (KDD'02), 2002, pp. 61-70.

[7] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks", in Proc. Of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'10), 2010, pp. 1029-1038.

[8] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model", in Proc. of the 10th IEEE Int. Conf. on Data Mining (ICDM'10), 2010.

[9] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model", in Proc. of the 12th Int. Conf. on Knowledge-

Based Intelligent Information and Engineering Systems (KES'08), 2008.

[10] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in Proc. of the 3rd ACM Int. Conf. on Web Search and Data Mining (WSDM'10), 2010.

[11] N. Barbieri, F. Bonchi, and F. Bonchi, "Topic-aware Social Influence Propagation Models", in 2012 IEEE 12th
Int. Conf. on Data Mining, 2012, pp. 81-90.

[12] A. Lancichinetti, and S. Fortunato, "Community detection algorithms: a comparative analysis", arXiv:0908.1062v1 [physics.soc-ph], 2009.

[13] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic Models for Discovering ECommunities", In Proc. of the 15th Int. Conf. on World Wide Web, 2006.

[14] A. McCallum, A. Corrada-Emmanuel, and X. Wang "Topic and Role Discovery in Social Networks", in Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI), 2005, pp. 786-791.

[15] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams,
and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks", In Journal Knowledge-Based Systems 26, 2012, pp. 164–173.

[16] Y. Liu, and A. Niculescu-Mizil, "Topic-link LDA: joint models of topic and author community", in Proc. of the 26th Int. Conf. on Machine Learning, Montreal, Canada,
2009, pp. 665-672.

[17] N. Pathak, C. DeLong, and A. Banerjee, "Social Topic Models for Community Extraction", in the 2nd SNA-KDD
Workshop'08, 2008.

[18] M. Sachan, D. Contractor, T. A. Faruquie, and L. V.Subramaniam, "Using Content and Interactions for Discovering Communities in Social Networks", in Proc. Of the 21st Int. Conf. on World Wide Web, 2012 pp. 331-340.

[19] S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom, "Expertise Identification Using Email Communications", In Proc. of the 12th Int. Conf. on Information and Knowledge Management, 2003, pp.528-531.