*Advances in Computer Science Research, volume 78*

3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018)

# A Survey of the Application of Social Media Mining

**Yehong Han**

Institute of Information Science and Engineering, Qilu Normal University, Jinan, China

email: sdzzhyh@163.com

**Keywords:** Social media; Community detection; Behavior analysis; Emotion prediction; Web crawler.

**Abstract:** Based on the rapid integration of online social networks and physical social networks, social networks are permeating all aspects of national security, economic development and social life. The application of social networks and data mining is widely used everywhere which include the generation of big data, the processing of data based on group intelligence and the consumption of information. From the technical point of view of the applications for social media mining, this article outlines the development trends in the four technical areas which inclde community detection, behavior analysis, emotion prediction and web crawler.

## 1. Introduction

Social networking can be seen as a medium.Because in this network platform, countless pieces of information are filtered by the nodes in the network and spread.Valuable information can be quickly circulated around the world, and worthless information can be forgotten or only spread on a small scale.Social media which is created by people using highly accessible and scalable publishing technologies. is the tool and platform for people to share opinions, insights, experiences and perspectives with each other. At present, social media mainly include social networking sites, wechat, blogs, forums, podcasts and more. Social media is thriving on the fertile ground of the internet which create a dazzling amount of energy. There are many interesting research topics in social network analysis. For example, the identification of community circles in social networks, the calculation of the influence of characters in social networks, the propagation model of information on social networks, the identification of false information and robot accounts, the prediction of stock markets, general election and infectious diseases based on social networks. From the perspective of the application of social network data mining, this article outlines the development trend of the following topics which inclde community detection, behavior analysis, emotion prediction and web crawler.

## 2. Community detection

The community reflects the local characteristics of the individual behaviors in the network and their relationship with each other. Through studying communities in research networks,we can understand the structure and function of the entire network and help us analyze and predict interactions across elements of the network.The research on community discovery is a basic research in network science and has great theoretical and practical value. In the meantime, the community has found some changes in its research focus and focus. In view of some characteristics of the network topology and community structure in the social network environment such as the online social network driven by the current Internet technology, the research on community discovery faces the following challenges. Traditional research on community discovery is generally based on the assumption that "each node is uniquely assigned to a certain community." In the real social network, people often belong to different communities at the same time. Such people who belong to multiple communities at the same time are the key to social interaction.Many traditional community discovery algorithms are based on global information such as edgemedium in GN algorithm, similarity between

any two points in random walk algorithm, module degree in modularity-based algorithm. The above algorithms must be considered in the premise of the entire network structure can be drawn. With the continuous improvement of informatization, the size of social network is becoming larger and larger, and it becomes very difficult to obtain the global information of the network.And these communities find algorithms to be inefficient under massive social network data. Waht's more ,Social networks are often sparse. The vast majority of individuals have direct contact with the outside world is limited. However, many researches and applications only concern the local structure of some nodes.

Santo Fortunato[1] pointed out identifying communities is an ill-defined problem. There are no universal protocols on the fundamental ingredients, like the definition of community itself, nor on other crucial issues, like the validation of algorithms and the comparison of their performances. The authour gives a guided tour through the main aspects of the problem and find weaknesses of popular methods.M. E. J. Newman[2] gives an exact equivalence between two widely used methods of community detection in networks, the method of modularity maximization in its generalized form which incorporates a resolution parameter controlling the size of the communities discovered, and the method of maximum likelihood applied to the special case of the stochastic block model known as the planted partition model.In the stochastic block model ,Jess Banks and Cristopher Moore [3] computer the upper bounds and the lower bounds on the information-theoretic threshold which define a symmetric stochastic block model.In orer to solve the problem that no suitable algorithm can computer which nodes belong to which communities with success any better than a random guess. where the sizes or average degrees differ is disgussed In the paper [4]. This asymmetry can assign nodes to communities with better-than-random success by examining their local neighborhoods. In the paper [5], node interests and interconnections are As the key element of community discovery which is different from traditional detection algorithm basde on network structure and ignore node interests and interconnections. A new termed interest-based clustering is difined. Above clustering includes structure, interaction, and node interest along with nodes friends' interests.

## 3. Behavior analysis

Individuals exhibit different behaviors in social media, and individual behavior is part of a wider range of group behaviors. Group behavior is the behavior of a group of individuals after planning or uncoordinated. Grasping the behaviors, characteristics and rules of information dissemination of users in social networks can not only help enterprises to provide better services and products based on user behavior characteristics, conduct more effective network marketing and promotion, but also provide the relevant departments with a reasonable monitoring and intervention provide a theoretical basis.In the social media environment, after a high-frequency interaction, friends who originally had weak connections (unknown or unfamiliar in the real world) have developed into strong connections (close contacts). Social objects in the new media environment are readily available, and new social media platforms can choose their own friends to add or not add through the address book. They can also form friendships based on common interests and concerns.The rapid development of new media and its complexity have caused a great impact on social life especially for people's values, thinking and behaviort.The development of social media has changed the social media system in which people interact. New media use a new platform to integrate the various types of traditional mass media to achieve a composite and borderless communication. Due to the diversity of virtual communities in the new media environment and the openness and inclusiveness of the new media environment, all aspects of social information are disseminated in all aspects from different perspectives, including the interactive dissemination of positive and negative information.

Based on parallel back propagation neural network and k-nearest neighbor algorithms, Xu G [6] designs a user behavior prediction model which improve the prediction accuracy. A parallel k-nearest neighbor algorithm is developed for user decision-making rule selection. In order mine social hotspots ,Yunpeng Xiao[7] designs a user participation behavior prediction model which can predict user participation behavior and topic development trends through ananlysizing user behavior and relationship data. Behavioral analysis considers three factors which include individual, peer and

community influence.In the paper[8], behavior prediction with explanations in social networks is studied . This paper designs an ontology-based deep learning model in order to predict behavior based on bottom-up algorithm ,the purpose of which is find the user representation.

## 4. Emotion prediction

In recent years, emotion analysis has become one of the hot issues in the field of social media processing. With the continuous emergence of heterogeneous data on the internet, the research on emotion analysis of multi-modal information has gradually become a new research hotspot in the field of social media processing. Research interests include textual sentiment classification, emotion calculation in the cross-media field and public opinion monitoring based on social media.At present, emotion analysis can be divided into two kinds. One is the classification of subjective and objective texts, the other is the subjective textual sentiment analysis which is the automatic analysis of emotional texts and the prediction of their emotional polarity.Emotional analysis based on different applications are divided into two areas: evaluation and emotional analysis. The former focuses on product performance evaluation which focuses on the psychological feelings of people. In social networks, the subjective emotion of user nodes is an important factor that affects the spread of information in the network. The complexity of the traditional user emotion prediction model is high which affects the efficiency of the emotion prediction in the social network.Emotional classification is the use of computer text sentiment classification. The use of emotional classification can predict the public opinion in the network, the advantages and disadvantages of commercial products can be analyzed. Even in the online user behavior mining can also add emotion classification content.

Soujanya Poria [9 ]uses all the data which include audio, visual and textual modalities as a source of data for emotional analysis in order to predict sentiments from web vedios.Both feature and decision fusion methods can be used to merge affective information extracted from multiple modalities. In the paper[10], the words in twitter are gave a semantic sentiment representation which can analyze semantic context based on their contextual semantics in order to predict perform entity- and tweet-level level sentiment analysis. Soujanya Poria does some good analysis in multimodal sentiment prediction [11][12]. A set of algorithm which is called recurrent model can be used for capturing contextual information among utterances by improving both context learning and dynamic feature fusion.What’s more, based on deep convolutional neural networks, features from visual and textual modalities can be extracted with high efficiency

## 5. Web Crawler

Different areas, different backgrounds of users often have different search purposes and needs. The results returned by the common search engine include a large number of pages that the user does not care about. The goal of a universal search engine is to maximize network coverage. The conflict between limited search engine server resources and unlimited network data resources will be further deepened. Common search engines mostly provide keyword-based searches that make it difficult to support queries based on semantic information.In order to solve above problems, focused crawler crawling related web resources came into being. Focused crawler is a program that automatically downloads web pages and selectively accesses social network-related links to get the information they need, based on their intended crawling goals. Unlike general purpose web crawlers, focused crawlers do not seek large coverage. The goal is to crawl pages related to a particular topic data resources for topic-oriented user queries. Focus crawler crawling strategy to achieve the key is to evaluate the importance of page content and links. For social media information mining, reptiles can be divided into four types which include crawling strategy based on content evaluation, crawling strategy based on link structure evaluation, crawling strategy based on enhanced learning and crawling strategy based on context map. Most social networking platforms offer open APIs for users and developers to obtain platform-related data, but there are usually a limited number of API calls. This makes the need to obtain large amounts of data is extremely inconvenient. Currently most

of the social networking platform using AJAX technology in order to provide rich functionality and good user experience. The process of crawling social networking platforms is also a process of AJAX page parsing.The amount of data generated by social networking platforms is huge. The data is usually some sparse unstructured format. So using the traditional relational database for storage is not convenient.There are three problems in the process of obtaining social network information.First, the use of traditional reptiles can not effectively obtain social network information.Social networking includes content that is diverse, such as videos, images, and text.This feature enhances the difficulty of using traditional reptiles to retrieve social network information.Second, the relational database storage is hard .In the social networking platform, in addition to the text content, a variety of information such as user information, information comment content are included. The existence of these information reduces the relevance of different content, and enhance the difficulty of information storage.In this case, it is difficult to use relational databases to store information on social networking platforms.Third, information filtering is difficult.The openness of social networking platforms has attracted the attention and use of more users.The increase in the number of users makes the social networking platform contains some noise content in the information.The existence of noise content enhances the difficulty of information crawling.Therefore, social network information needs to be filtered when crawling social network information.

In general, the crawler can not gain all of the information based on collecting data in a non-login status because the crawler is not provided with the capabilities to log into the system. To enable information sharing among users, the sign microblog open platform offers a number APIs in which certain limitations such as those on the number of user requests of the microblog server, still exist. S Zhang and J Zhou [13] design a method which can simulate user login to collect microblog data related to hot topics on the SINA platform.For data acquisition on deep web, PA Madhusudan[14] did some useful work. The deep web refers to non-surface network content on the internet that cannot be indexed by a standard search engine. The focused semantic web crawler is designed to gain topical deep web contents based on retrieving the relevant sites through deep search by in-site exploring.In the paper[15], a set of keywords relevant to the topic of interest of the user is used to shoot queries on search interface which can be found on webpage of the website corresponding to seed URL. in order to get most relevant links from the domain without actually going in depth of that domain.

## 6. Conclusion

With the rapid development of internet technology, social media services play an increasingly important role in the real life of users. The same user may have accounts of multiple social media websites at the same time, which respectively correspond to different online community identities. Based on these community identities, users can simultaneously participate in multiple social media platforms and enjoy the application services provided. Meanwhile, with the migration of social media service platforms to mobile clients, an intelligent mobile communication tool often binds a variety of application services. This has further strengthened this trend of multi-community identity.Social network analysis also gradually from the macro network topology analysis, development of medium-sized community discovery and more micro-social relations, influence and user behavior modeling, etc..There are also many essential challenges, including user interaction, the basic theory of social information theory, the key technologies of social data mining.

## References

[1] Fortunato S, Hric D. Community detection in networks: A user guide[J]. Physics Reports, 2016, 659: 1-44.

[2] Newman M E J. Community detection in networks: Modularity optimization and maximum likelihood are equivalent[J]. arXiv preprint arXiv:1606.02319, 2016.

[3] Banks J, Moore C, Neeman J, et al. Information-theoretic thresholds for community detection in sparse networks[C]//Conference on Learning Theory. 2016: 383-416.

[4] Zhang P, Moore C, Newman M E J. Community detection in networks with unequal groups[J]. Physical review E, 2016, 93(1): 012303..

[5] AlSuwaidan L, Ykhlef M. Interest-Based Clustering Approach for Social Networks[J]. Arabian Journal for Science and Engineering, 2018, 43(2): 935-947.

[6] Xu G, Shen C, Liu M, et al. A user behavior prediction model based on parallel neural network and k-nearest neighbor algorithms[J]. Cluster Computing, 2017, 20(2): 1703-1715.

[7] Xiao Y, Lai J, Liu Y. A user participation behavior prediction model of social hotspots based on influence and Markov random field[J]. China Communications, 2017, 14(5): 145-159.

[8] Phan N, Dou D, Wang H, et al. Ontology-based deep learning for human behavior prediction with explanations in health social networks[J]. Information sciences, 2017, 384: 298-313.

[9] Poria S, Cambria E, Howard N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content[J]. Neurocomputing, 2016, 174: 50-59.

[10] Saif H, He Y, Fernandez M, et al. Contextual semantics for sentiment analysis of Twitter[J]. Information Processing & Management, 2016, 52(1): 5-19.

[11] Poria S, Chaturvedi I, Cambria E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE, 2016: 439-448.

[12] Poriaa S, Cambriab E, Hazarikac D, et al. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis[C]//2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017: 1033-1038.

[13] Zhang S, Zhou J, Shi M, et al. A simulated login-based SINA microblog data collection method and its data analysis[C]//Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE, 2016: 1-4.

[14] Madhusudan P A, Lambhate Poonam D. Deep Web Crawling Efficiently using Dynamic Focused Web Crawler[J]. 2017.

[15] Kumar M, Bindal A, Gautam R, et al. Keyword query based focused Web crawler[J]. Procedia Computer Science, 2018, 125: 584-590.