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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Survey on Social Media Data Mining Techniques

Sonali More, P. P. Joshi

M.E. Student, Department of Computer Engineering, Pune Institute of Computer Technology, SPPU, Pune, India

Assistant Professor, Department of Computer Engineering, Pune Institute of Computer Technology, SPPU, Pune, India

ABSTRACT: Social network has gained remarkable attention in the last decade. Social media provide a platform with tools to share information, to debate health care policy and practice issues, to promote health behaviours, to engage with the public and to educate and interact with patients.There are multiple social sites like Facebook, Twitter, User Forums where users share their experiences.The heavy reliance on social network sites causes them to generate massive data characterised by three computational issues namely; size, noise and dynamics. Analyse such data manually is very complex task which resulting in the pertinent use of computational means of analysing them. There are multiple range of data mining techniques for detecting useful knowledge from massive datasets like trends, patterns and rules. This survey discusses different data mining techniques used in mining social media.

KEYWORDS: Social Network, Social Network Analysis, Data Mining Techniques

I. INTRODUCTION

Social media is providing limitless opportunities for patients to discuss their experiences with drugs anddevices, and for companies to receive feedback on their products and services. Social media allows a virtual networking environment.Modeling social media the usage of available communitymodeling and computational gear is one manner of extracting, understanding and traits from the facts cloud: a social community is a structure made of nodes and edges that connectnodes in various relationships. Graphical representation is the maximum commonplace technique to visually represent the statistics. Community modeling can also be used for readingthe simulation of community houses and its inner dynamics. Manually analyses social media data is very complex difficult. Traditional method includes taking survey from consumer in order to collect data, resulting in small data size per study, where social media provides access to large amount of data which readily available specifically combined with internet-crawling and scraping software program that might allow actual-time tracking of modifications within the network. This is allowing real time monitoring of network. Text mining is a burgeoning new area that tries to glean significant statistics from natural language text. It could be loosely characterized as the system of analyzing text to extract statistics this is useful for unique purposes. This study shows multiple techniques for data mining of social media data and analyze social network.

II. RELATED WORK

Social network analysis has gained prominence due to its use in different applications - from product marketing (e.g. viral marketing) to search engines and organizational dynamics (e.g. management). Now a days interest regarding social network analysis has been a rapidly increased. The main aim is that to extract knowledge from huge amounts of data collected, also analysing a social behaviour of users in online environments.

A number of research issues and challenges facing the realisation of utilising data mining techniques in social network analysis could be identified as follows:

In order to mine data of social media we are going through network modelling. Degree of node, network density and other important large-scale parameters can derive information about the importance of certain entities within the network. The nature of social networks makes data collection difficult. Several methods have been employed.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

L. Getoor and C. Diehl (2005) have used link mining that is focused on finding patterns in data by exploiting andexplicitly modeling the links among the data instances [5]. Q.Lu. and L. Getoor (2003) proposed simple framework to modeling link distribution using link structure which helps to improve classification accuracy [7]. A. Ng, A. Zheng, and M. Jordan(2001) uses page rank and HITS algorithm to findout the influential articles and also go through the stability of both algorithm [8]. This focused on web graph setting with multiple connected components. Further Katherine Faust(2007) have proposed triadic structure in order to patterns of triads is accounted for by lower-order properties pertaining to nodes and dyad [9]. Huda Alhazmi, Swapna S. Gokhale, Derek Doran(2013) have studied triadic analysis that helps unveil social influences that can explain users [10]. KitYan Chan,C.K. Kwong and T.C.Wong developed models for relating customer satisfaction and design attributes using genetic programming [11].

Most of Previous studies used technical solutions to extract user sentiments. S. R. Das and M. Y. Chen(2007)gone through Naive Classifier, Vector Distance Classifier,Discriminant-Based Classifier for to extract user sentiment.These algorithms coupled by a voting scheme are evaluated using a range of metrics[12]. This study on influenza by Courtney D. Corley, Diane J. Cook, Armin R. Mikler, Karan P. Singh(2010) .They used Subdues discovery algorithm that requires high processing time [13]. Altug Akay, Andrei Dragomir, Bj orn-Erik Erlandsson have used self orgnizing map for user sentiment [1].X. Feng, A. Cai, K. Dong, W. Chaing, M. Feng, N.S. Bhutada, J. Inciardi, and T. Woldemariam have usedtechnical solutions to extract user sentiments on government health monitoring[14].

Link Mining:

Link mining can be seen as the task of applying data mining techniques on networks, while explicitly considering and emphasizing on links between social network actors.Link mining is an emerging area within data mining that is focused on finding patterns in data by exploiting and explicitly modelling the links among the data instances[4]. Link mining tasks are broadly categorized into following tasks:

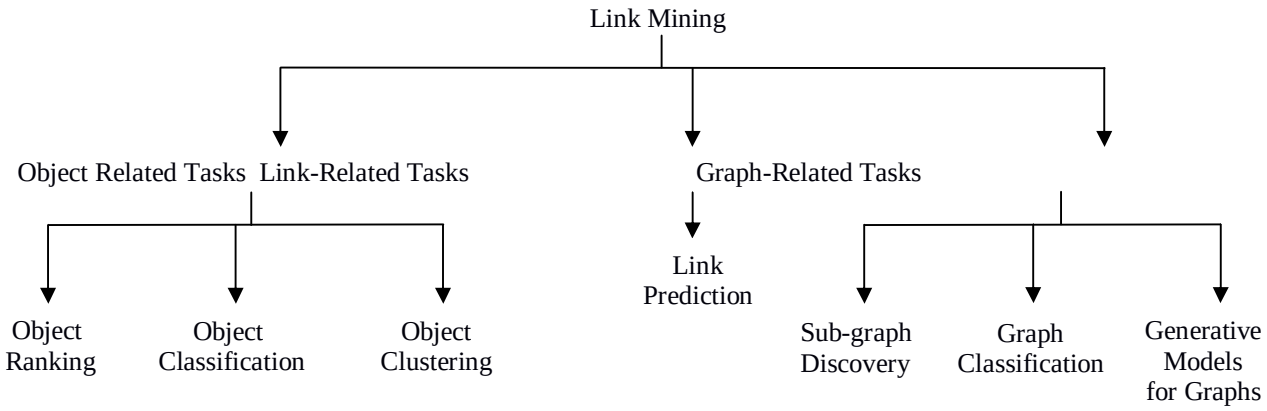

Fig 1: Link mining Task classification

Graph Theory:

In social network analysis graph theory is probably the main method .The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links (for example influencers and the followers). Influential users have been identified as users that have impact on the activities or opinion of other users by way of followership or influence on decision made by other users on the network. Graph theory has proved to be very effective on large-scale datasets (such as social network data). This is because it is capable of bye-passing the building of an actual visual representation of the data to run directly on data matrices. Burt, R S was used centrality measure[17] was used to inspect the representation of power and influence that forms clusters and cohesiveness [18] on social network. Thus, centrality is shown to be intimately connected with the cohesive subgroup structure of a network. The Ghosh, R., Lerman, Kemployed parameterized centrality metric approach [19] to study the network structure and

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

to rank nodes connectivity. Their work formed an extension of a-centrality approach which measures the number of alleviated paths that exist among nodes.

Community Detection Using Hierarchical Clustering:
A community is a smaller compressed group that present a larger network. Community formation is known to be one of the important characteristics of social network. Users who have similar interest form the communities which part of social network which displaying strong structure. Communities on social networks, like any other communities in the real world, are very complex in nature and difficult to detect. To detecting and understanding the behaviour of network communities by applying appropriate tool is crucial. Different authors have applied diverse clustering techniques to detect communities on social network with hierarchical clustering being mostly used [15]. This technique is a combination of many techniques that find outthe strength of individual groups which is then used to divide the network into communities. Vertex clustering belongs to hierarchical clustering methods, graph vertices can be resolved by adding it in a vector space so that pairwise length between vertices can be measured. Structural equivalence measures of hierarchical clustering emphasis on number of common network connections shared by two nodes. Peoples who have several mutual friends on social network with are more likely to be closer than People with less mutual friends on the network. Users in the same social network community are closely connected and often recommend items and services to one another based on the experience on the items or services involved.

Recommender System in Social Network Community:
On the basis of mutuality between nodes in social network groups, collaborative filtering (CF) technique, which forms one of the three classes of the recommender system (RS), can be used to exploit the relation and connection among users [14]. Items can be recommended to a user based on the basis of his similar interest connection rating. CF's main downside is that of data sparsity, content-based (which is another RS method) explores the structures of the data to produce recommendations. However, the hybrid approaches usually suggest recommendations by combining CF and content-based recommendations. The experiment in proposed a hybrid approach named EntreeC, a system that pools knowledge-based RS and CF to recommend restaurants. By using a greedy implementation of hierarchical agglomerative clustering is used improved on CF algorithm which suggests forthcoming conferences.It increases recommendation effectiveness by incorporating social network information into CF.

Opinion Definition and Opinion Summarization:
For recognizing opinion of users opinion definition and opinion summarization are essential techniques. Opinion definition can be located in a text, sentence or topic in a document; it can also reside in the entire document.Opinion summarization sums up different opinions aired on piece of writing by analysing the polarities of sentiments, degree and the related occurrences.
Opinion summarization helps in improving policies and products respectively**,** so it is always useful to businesses and government.In opinion extraction, if multiple people that give their opinion on a any particularsubject,the more important that portion might be worth extracting. Opinion can aim at a particular article while on the other hand can compare two or more articles. The formal is a regular opinion while the latter is comparative [16]. Opinion extraction identifies subjective sentences with sentimental classification of either positive or negative.

1. **SVM**

Sentiment analysis is a classification task as it classifies the orientation of a text into either positive or negative. Many experimental results that applied Support Vector Machine (SVM) on benchmark datasets to train a sentiment classifier. N-grams and different weighting scheme were used to extract the unique features of data. It also explores Chi-Square weight features to select informative features for proper the classification. Chi-Square feature selection for selection that may improve the classification accuracy.
For content-based classifications M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer have used linear support vector machines (SVMs) to discriminate between users in the two classes as 'left' and 'right' class. In the simple case of binary classification, an SVM works by embedding data in a highdimensional space and attempting to find the hyperplane that best separates the two classes [20]. Support vector machines are widely used for document

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

classification because they are well-suited to classification tasks based on sparse, highdimensional data, such as those commonly associated with text corpora [21].

2. Naïve Bayes

Bayesian network classifiers are a popular supervised classification paradigm. Bayesian network classifier is the Naïve Bayes' classifier is a probabilistic classifier which works on the Bayes' theorem, considering Naïve (Strong) independence assumption. It was introduced under a different name into the text retrieval community and remains a popular (baseline) method for text categorizing, the problem of judging documents as belonging to one category or the other with word frequencies as the feature. An advantage ofNaïve Bayes' is that it only requires a small amount of training data to estimate the parameters necessary for classification. Abstractly, Naïve Bayes' is a conditional probability model. Despite its simplicity and strong assumptions, the naïve Bayes' classifier has been proven to work satisfactorily in many domains. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined.

3. SOM

Self-Organizing Maps (SOMs) are neural networks that produce low-dimensional representation of high-dimensional data.Within this network, a layer represents output space with each neuron assigned a specific weight. The weight values reflect on the cluster content. The SOM displays the data to the network, bringing together similar data weights to similar neurons [1]. When new data is fed into the network, the closest weights matching the data change to reflect the new data. The neurons farther from the new data rarely change. This process continues until data is no longer fed, resulting in a two-dimensional map.

4. LVQ

Learning Vector Quantization (LVQ) is a supervised version of vector quantization. LVQ can be used when we have labelled input data. This learning technique uses the class information to reposition the class vectors slightly, so as to improve the quality of the classifier decision regions. The LVQ algorithm is based on neural competitive learning, which enables defining a group of class labels on the input data space using reinforced learning. Like SOM,LVQ also transforms high dimensional input data into a two-dimensional map, but without taking topological order of input data. For extract the user sentiment, LVQ utilizes pre assigned class labels to documents, thus minimizing the average expected misclassification probabilityand might be improve accuracy[22]. Hence, unlike the SOM, where clusters are generated by unsupervised manner based on feature-vector similarities, the LVQ categories are predefined.

III. Conclusion and Future Work

Different data mining techniques have been used in social network analysis as covered in this survey. There are multiple techniques range from unsupervised to semi-supervised and supervised learning methods.So far different levels of successes have being achieved either with solitary or combined techniques. The outcome of the experiments conducted on social network analysis is believed to have more focus on the different structure and activities of social network.

References

1. A. Akay, A. Dragomir, and B. E. Erlandsson, "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care", IEEE journal of biomedical and health informatics, vol. 19, no. 1, january 2015
2. W. Cornell and W. Cornell. (2013). "How Data Mining Drives Pharma: Information as a Raw Material and Product". [Online]. Available: http://acswebinars.org/big-data.
3. L. Dunbrack, Pharma 2.0 "social media and pharmaceutical sales and marketing", in Proc. Health Ind. Insights, p. 7, 2010.
4. L. Getoor and C. Diehl, "Link mining: a survey", SIGKDD Explor. Newsl. vol. 7, pp. 312, Dec. 2005.
5. Q. Lu. And and L. Getoor, "Link-based classification", in Proc. 20th Int.Conf. Mach. Learning, Washington, D.C., USA, pp. 496503.
6. 2003
7. A. Ng, A. Zheng, and M. Jordan, "Stable algorithms for link analysis", in Proc. SIGIR Conf. Inform. Retrieval., New Orleans, LouisianaLO, USA, pp. 258266,2001.
8. K. Faust, Very local structure in social networks, Sociological Methodology, vol. 37,pp. 209256, Nov. 2007.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

9. Huda Alhazmi, Swapna S. Gokhale, Derek Doran, "Understanding Social Effects in Online Networks", Kno.e.sis Publications, v2, pp 1048, 2015.
10. Kit Yan Chan*, C.K. Kwong and T.C. Wong, "Modelling customer satisfaction for product development using genetic programming", Journal of Engineering Design Vol. 22, No. 1, pp. 5568, January 2011.
11. S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web", Manag. Sci., vol. 53, pp. 13751388, Sep.2007.
12. C. Corley, D. Cook, A. Mikler, and K. Singh, "Text and structural data mining of influenza mentions in web and social media", Int. J. Environ.Res. Public Health, vol. 7, pp. 596615, Feb. 2010.
13. A. Akay, A. Dragomir, and B. E. Erlandsson, "A novel data-mining approach leveraging social media to monitor consumer opinion of sitagliptin", J. Biomed Health Inform. Vol: PP, Issue: 99.
14. X. Feng, A. Cai, K. Dong, W. Chaing, M. Feng, N.S. Bhutada, J.Inciardi, and T. Woldemariam, "Assessing pancreatic cancer risk associated with dipeptidyl peptidase 4 inhibitors": data mining of FDA adverse event reporting system (FAERS),Pharmacovigilance, vol. 1, Jul. 2013.
15. Liu, F., Lee, H. J.: "Use of social network information to enhance collaborative filtering performance". Expert Systems with Applications, 37, 4772-4778, 2010.
16. Newman, M.: "Networks: An introduction". Oxford University Press, 2010.
17. Jindal, N., Liu. B. "Mining Comparative Sentences and Relations". In: Proceedings of National Conf. on Artificial Intelligence (AAAI- 2006), 2006.
18. Burt, R S. Brokerage and closure: "An introduction to social capital". Oxford University Press, 2005.
19. Borgatti, S. P., Everett, M. G.: "A graph-theoretic perspective on centrality. Social networks 28", 466-484, 4, 2006.
20. Ghosh, R., Lerman, K.: Parameterized centrality metric for network analysis. Physical Review E, 83(6), 066118, 2011.
21. Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: "Predicting the political alignment of twitter users. In Privacy, security, risk and trust (passat)", 2011 IEEE third international conference on and 2011 ieee third international conference on social computing (socialcom) (pp. 192-199). IEEE, 2011.
22. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. of the 10th European Conference on Machine Learning (ECML), pp. 137–142, 1998.
23. Anuj Sharma1 , "Shubhamoy Dey, Using Self-Organizing Maps for Sentiment Analysis" **[v1]** Mon, 16 Sep 2013.

BIOGRAPHY

Sonali More is a student of M.E. in the Computer Engineering Department, Pune Institute of Computer Technology. Her research interests are Data mining, Neural Network.