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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

A Survey on Data Mining Techniques in Social Network Data

T.Sureshkumar[1], G.Vadivel[2]

M.Phil. Scholar, Department of Computer Science, SNMV College of Arts and Science, Coimbatore, Tamilnadu, India[1]

Assistant Professor, Department of Computer Science, SNMV College of Arts and Science, Coimbatore, Tamilnadu, India[2]

ABSTRACT: Analysing social media data, in particular Twitter feeds for sentiment analysis and privacy analysis, has become a major research and business activity due to the availability of web-based application programming interfaces provided by Twitter, Facebook and News services. Social Media sites such as Facebook, Twitter, LinkedIn and Google+ contain large volume of unprocessed raw data. By analysing this data new knowledge can be gained. Since this data is dynamic and unstructured traditional data mining techniques will not be appropriate. In this paper we discuss about data mining, social media data, data mining techniques applied in the social media. In this paper a survey of the works done in the field of social network data mining analysis and techniques followed to perform the data mining on the social network data. Results of this survey can serve as the baselines for future data mining research.

KEYWORDS: data mining, social media, social network data analysis

I. INTRODUCTION

Data mining is a powerful tool that can help to find patterns and relationships within our data. Data mining discovers hidden information from large databases.[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Social networks can be used in many business activities like increasing word-of-mouth marketing, marketing research, General marketing, Idea generation & new product development, Co-innovation, Customer service, Public relations, Employee communications and in Reputation management.

Supervised and unsupervised algorithms are used to identify the hidden patterns in data. Supervised approaches depend on some a-priori knowledge of the data (e.g. class labels). Unsupervised algorithms are used to characterize data without any prior instruction as to what kinds of patterns will be discovered by the algorithm. The variety of work accomplished to date pertaining to data mining of online social media data is accomplished with some version of either supervised or unsupervised learning algorithms. Determining whether a supervised or an unsupervised approach would be best depends on the data set and the particular question being investigated. Data sets can be generalized into three types: data with labels, data without labels, and data with only a small portion of labels.

Classification is a common supervised approach and is appropriate when the data set has labels or a small portion of the data has labels. Classification algorithms begin with a set of training data which includes class labels for each data element. The algorithm learns from the training data and builds a model that will automatically categorize new data elements into one of the distinct classes provided with the training data. Classification rules and decision trees are examples of supervised classification techniques.

Clustering is a common unsupervised data mining technique that is useful when confronting data sets without labels. Unlike classification algorithms, clustering algorithms do not depend on labelled training data to develop a model. Instead, clustering algorithms determine which elements in the data set are similar to each other based on the similarity of the data elements. Similarity can be defined as Euclidian distance for some numerical data sets but often in data associated with social media, cluster techniques must be able to deal with text. In this case, clustering techniques use keywords that are represented as a vector (to represent a document) and the cosine similarity measure is used to distinguish how similar one vector (data element) is to another.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

Several data mining techniques have been developed by scientists in order to overcome the problems such as size, noise and dynamic nature of the social media data. Due to the large volume of data in the social media, an automatic data processing is needed in order to analyse it within a given time span. The dynamism in the social media data leads to the rapid evolution of the data sets over time; such dynamic data can be easily handled by various data mining techniques (Adedoyin-Olowe, Gaber, & Stahl, 2013) [2].

II. BACK GROUND

Social media is an internet based communication tool that empowers people to share information. To understand better the term social media, social indicates associating with people and spending time in order to develop their relationships whereas media indicates tool for communication such as internet, TV, radio, newspaper so on, here our focus is internet. Social media is stated as an electronic platform for socializing people. Some example of social media sites are Facebook, Twitter, Youtube, LinkedIn, Digg etc. Initially people involved in social media to associate with their friends and lost friends,gradually they improved to the status of updating and consuming any information on social media, these led to vast generation of user data which could be further processed for future development.

In the social media data analysis, data mining classification techniques fall under three types of learning methods:

- In Supervised learning where the network is trained by providing it with input and matching output patterns.
- Unsupervised learning where the output is trained to react to clusters of pattern within the input. There is no a priori set of categories into which the patterns are to be classified.
- Semi supervised learning where the test nodes need to be predicted are known. Reinforcement learning is the intermediate form between supervised and unsupervised learning.

Unsupervised classification: unsupervised learning tries to the find hidden structure in unlabelled data. Since the examples given to the learner are unlabelled then there is no error or reward signal to evaluate a potential solution. Sentiment lexicon, Opinion definition and summarization, sentiment orientation and opinion extraction are performed.

Semi supervised classification: Since the social network data usually come in huge sizes, in addition there are usually, a huge number of unlabelled instances. In such cases, it could be possible to use the information other than labels that exists in the unlabelled data, which leads to use of semi supervised learning algorithms. When the test nodes whose class will need to be predicted are known. In semi supervised learning, the unbalanced instances can be used to monitor the variance of the produces classifiers, to maximize the margin and hence to minimize the complexity.

Supervised classification: While clustering techniques are used where basis of data [3] is established but data pattern is unknown, classification techniques are supervised learning techniques used where the data organisation is already identified. It is mentioning that understanding the problem to be solved and opting for the right data mining tool is very essential when using data mining techniques to solve social network issues. SVM, Naive bayes, Neural Networks, K-nearest neighbour and text mining techniques are used.

The various data mining techniques involved in the social media data analysis such as

a. **Characterization:** Characterization is used to generalize, summarize and possibly different data characteristics.
b. **Classification:** Data classification is a process in which the given data is classified in to different classes according to a classification model.
c. **Regression:** This process is similar to classification the major difference is that the object to be predicted is continuous rather than discrete.
d. **Association:** In this process the association between the objects is found. It discovers the association between several data bases and the association between the attributes of single database.
e. **Clustering:** Clustering involves grouping of data into several new classes such that it describes the data. It breaks large data set into smaller groups to make the designing and implementation process to be simple. The task of clustering is to maximize the similarity between the objects of classes and to reduce the similarity between the classes.
f. **Change Detection:** This method identifies the significant changes in the data from the previously measured values.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

g. **Deviation Detection:** it focuses on the major deviations between the actual values of the objects and its expected values. This method find out the deviation according to the time as well the deviation among different subsets of data.
h. **Link Analysis:** It traces the connections between the objects to develop models based on the patterns in the relationships by applying graph theory techniques.
i. **Sequential Pattern Mining:** This method involves the discovery of the frequently occurring patterns in the data.[1] Social networks are important sources of online interactions and contents sharing, subjectivity, assessments, approaches, evaluation, influences, observations, feelings, opinions and sentiments expressions borne out in text, reviews, blogs, discussions, news, remarks, reactions, or some other documents.[4]

III. LITERATURE REVIEW

In the literature survey it has been clearly noticed that many researches work how these interests of the users can be analyzed for future conclusions and usages. There were many different methods proposed. Few are listed here.

Bogdan Batrinca et al [5] very systematically and strategically explains the techniques, tools and platforms for social media analytics, a detailed study on data retrieval techniques are presented in their paper. They also discuss important terms such as social media, sentiment analysis, scraping, opinion mining, behavior economics, NLP and various toolkits and software platforms. The recent research challenges such as scraping, data cleaning, holistic data sources, data protection, data analytics, analytics dashboards and data visualization is been discussed. They also critique the companies' policies of restricting the data access to gain monetary benefits.

Karthikeyan & Vyas[6], The text mining or text data mining is the data mining technique in which one derives high-quality information from texts. Large media companies such as the tribune etc. By using the text mining techniques in order to make the information clear and to provide better search experiences to the readers which consequently increases the 'stickiness' of the site and helps the site to generate larger revenue.

Adedoyin-Olowe, Gaber, & Stahl [2], Sentiment lexicon can be regarded as a dictionary of the emotional words which are frequently employed by the reviewers in their communication. It comprises a list of ordinary words that helps in the improvement of the data mining techniques when they are used for mining a sentiment in the certain document. Depending upon the diversity in subject matters, various collections of sentiment lexicon can be generated. The sentiment words employed in the sports, for example, are unlike those employed in the politics. We can focus more on topic-specific occurrence by expanding the occurrence of sentiment lexicon combined with the use of high man power.

J. Bonneau, J. Anderson, and G. Danezis [7] , A social networking site like Facebook or LinkedIn consists of connected users with unique profiles. User can link to friends and colleagues and can share news, photos, videos, favorite links etc. Users customize their profiles depending on individual preferences but some common information might include relationship status, birthday, an e-mail address, and hometown. Users have options to decide how much information they include in their profile and who has access to it. The amount of information available via a social networking site has raised privacy concerns and is a related societal issue.

It is important to protect personal privacy when working with social network data. Recent publications highlight the need to protect privacy as it has been shown that even anonymizing this type of data can still reveal personal information when advanced data analysis techniques are used [7,8]. Privacy settings also can limit the ability of data mining applications to consider every piece of information in a social network. However, some nefarious techniques can be employed to usurp privacy settings [7].

Ritu Mewari[9], presented an opinion mining provides a clear platform to catch public's mood by filtering the noise data. It also provides computational techniques used to extract and consolidate individual's opinion from unstructured and noisy text data. Opinion mining is a burning field of web mining. There exist a lot of benefits of opinion mining at customer and business level. A bulk of data is daily posted on web sites like face book and twitter. User post their sentiments in the form of comments, reviews and feedback daily. An opinion mining process gives us the way to extract pearl knowledge from it.

A study by Al-Daihani, S. M., [10] applied the text mining approach on a large dataset of tweets. The complete Twitter timelines of 10 academic libraries were used to collect the dataset for this research. Nearly 23,707 tweets formed the total dataset, where there were 7625 hashtags, 17,848 mentions, and 5974 retweets. The significance of

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

data and text-mining approaches are reported within the study and their purpose is to gain an insight with the aggregate social data of academic libraries so that the process of decision-making and strategic planning could become facilitated for marketing of services and patron outreach.

Kermanidis,K.L.,[11] stated that sentiment analysis through social media usage has witnessed a huge interest from scholars in the last few years. In that, the authors discussed the influence of tweets' sentiment on elections and the impact of the elections' results on web sentiment.

Pang, B., Lee, L [13], Millions of people access social media sites such as Twitter, Facebook, LinkedIn, YouTube and MySpace to search out for information, breaking news and news updates. Most of the updates are often posted by unfamiliar people they have never and may never have contact with. Consequently information gathered on social media is sometimes used to make valuable decisions. While some groups are retrieving information from social media sites, others are posting information for the use of other internet users.

IV. CONCLUSION

This paper provides a more current evaluation and update of social network analysis research available. Literatures have been reviewed based on different aspects of social network analysis.

This paper studies the application of the techniques and concept of data mining for social networks analysis, and reviews the related literature about text mining and social networks. Social networks investigation carried out through the techniques of Web mining is an interesting field of research. However, there are many challenges in this research field to be resolve with improvement.

REFERENCES

1. M. Vedanayaki, "A Study of Data Mining and Social Network Analysis", Indian Journal of Science and Technology, Vol 7(S7), 185–187, November 2014.
2. Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617.
3. Batista, G., & Monard, M.C., (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, vol. 17, pp.519-533.
4. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber and Frederic Stahl, "A Survey of Data Mining Techniques for Social Network Analysis", Journal of Data Mining & Digital Humanities, 2014.
5. Bogdan Batrinca and Philip C Treleaven, "Social media analytics: a survey of techniques, tools and platforms," AI \& SOCIETY, vol. 30, no. 1, pp. 89-116, 2015.
6. Karthikeyan, M., & Vyas, R. (2014).Cloud Computing Infrastructure Development for Chemoinformatics.In Practical Chemoinformatics (pp. 501-528).Springer India.
7. J. Bonneau, J. Anderson, and G. Danezis. Prying data out of a social network. pages 249 –254, july 2009.
8. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Computational social science. Science, 323:721–723, 2009.
9. Ritu,Ajit Singh,Akash Srivastava, "Opinion Mining Techniques on Social Media Data", International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 6, May 2015.
10. Al-Daihani, S. M., & Abrahams, A. (2016). A Text Mining Analysis of Academic Libraries' Tweets. The Journal of Academic Librarianship, 42(2), 135-143.
11. Kermanidis,K.L., & Maragoudakis, M. (2013). Political sentiment analysis of tweets before and after the Greek elections of May 2012. International Journal of Social Network Mining, 1(3-4), 298-317.
12. https://data-flair.training/blogs/text-mining/
13. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis; Foundations and Trends in Information Retrieval; Vol. 2, Nos. 1–2, 1–135, 2008