ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

A Study on Social Network Analysis through Data Mining Techniques – A detailed Survey

Annie Syrien[1], M. Hanumanthappa[2]

Research Scholar, Department of Computer Science and Applications, Bangalore University, India[1]

Professor, Department of Computer Science and Applications, Bangalore University, India[2]

Abstract: The paper aims to have a detailed study on data collection, data preprocessing and various methods used in developing a useful algorithms or methodologies on social network analysis in social media. The recent trends and advancements in the big data have led many researchers to focus on social media. Web enabled devices is an another important reason for this advancement, electronic media such as tablets, mobile phones, desktops, laptops and notepads enable the users to actively participate in different social networking systems. Many research has also significantly shows the advantages and challenges that social media has posed to the research world. The principal objective of this paper is to provide an overview of social media research carried out in recent years.

Keywords: data mining, social media, big data.

I. INTRODUCTION

Data mining is an important technique in social media in recent years as it is used to help in extraction of lot of useful information, whereas this information turns to be an important asset to both academia and industries. Social media with big data is a very hot topic in the recent years, the companies are willing to share this data in order to absorb large market share [1]. On 30 January 2015 Facebook updated its terms and policies, accordingly if the user wants to continue using the service he should abide the new advertising rules. This helped the Facebook to be more specific on online advertising which impacted the increase in its revenue. According to Irena Pletikosa Cvijikj et al [2] advancement in social media platforms has changed the traditional marketing communication through placing the control to consumers who started dictating the nature of marketing contents.

The structure of this paper follows as. Discussion of data mining, social media and big data terms, literature survey on the terms presented, summarizing of the literature survey, conclusion.

II. DATA MINING

Data mining is technique of examining data from different aspects and recapitulate into useful information – information can be used to enhance quality, increase revenue and reduce costs so on.

According to Usama Fayyad et al [3]**,** Data mining is KDD (knowledge discovery in databases) step which comprises of data analysis to produce a particular number of patterns on the data. Earlier days to transform the data as valuable information required lot of human efforts through manual scrutiny and analysis. In any industry data mining technique based reports become the base line for making decisions and formulating the decisions. This report becomes the groundwork for future decision making and planning for any organizations [3].

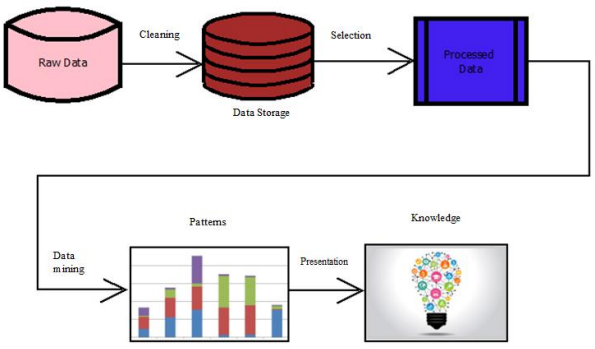

Figure 1: Data Mining Process

III. SOCIAL MEDIA

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

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

consuming any information on social media, these led to vast generation of user data which could be further processed for future development.

IV. BIG DATA

Data is raw facts and figures which is produced and available in the real world when day to day transaction takes place. While this day to day transaction data was computerized when it was understood such data is very useful for future reference. Hence it gave birth to data storage. In recent years the development of smart phones contributed to easy access of social networking sites like anytime and anywhere. So the number of people and frequency of usage increased very significantly. Whereas the users' day to day activities were stored in the data storage units of social networking sites which produced magnanimous volume of data, such data is called as Big data. Due to it's characteristics of beyond processing capacity it welcomed many researchers in recent years.

V. LITERATURE SURVEY

The social media analysis has led to an important type of promotions such as advertising exactly what the user is looking for, years ago the advertising was common on most of the websites, but due to advancements in the big data and data mining, the organizations specifically advertise depending on the user requirement, which was searched earlier, this leads to an smart business dealings and also increases the sales and profit.

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.

Haishuai Want et al [4] presented the research work SocialAnalysis in which he designed real time systems in order to discover and summarize emergent social events from data streams which obtained from social media. The important aspects the system includes are follows:

1. Data denoising methods
2. Abnormal events detection
3. Geographical position recognition

Different tools are used in order to achieve these steps, some of which are Sina Weibo for data collection and detection methods, event peak identification algorithm for abnormal time points identification. Different location weights are used to discover the genuine locations.

The socialAnalysis system consists of four components

- Data collection
- Noisy text filtering
- Anomalous event detection
- Geographical location detection

The data collection was achieved through web page crawler and Sina API functions; the noisy text filtering was performed through event correlation analysis, in which the irrelevant texts were filtered. Anomalous event detection was carried out through statistical analysis and finally the geographical location detection was through correlated weibo contents, extracting and analyzing the locations of the events.

The event correlation analysis aimed to remove unwanted data which took place in four steps

1. Pre-processing: removing insignificant characters and URLs were performed at this phase
2. Correlation computation: the correlation is calculated through, VSM, String matching, Levenshtein distance and SimHash algorithms
3. Data filtering: This is achieved when correlation is less than 0.2 or text length no more than 8 characters
4. Ranking: A comparative variable L is designed to integrate the correlation and number of message forwarding,

$$L = \beta * \frac{\text{Count}}{\text{Maxcount}} + (1 - \beta) * \frac{\text{sim}}{\text{Maxsim}}$$

Where Maxcount is the maximal number of forward messages, count is the number of message forwarding, sim is the correlation between a weibo text and the query words, and Maxsim is the maximal correlation result in the computation result.

Real-time Anomalous events detection

The real-time anomalous events detection takes place in the following steps

- Through visualizing the number of weibo messages related to the events
- Discovering the abnormal time points by a peak-finding algorithm

The above research paper had a very significant demonstration of work on SocialAnalysis, which expressed the solution for data collection, abnormal event detection and location identification. The SocialAnalysis system could detect events, estimate locations of the events and also give early warnings.

SAA Hridoy et al [5] demonstrated the methodology to determine the opinion or sentiment of a product at different locations across male or female users on Twitter. iPhone 6 was chosen as their research product reason being availability of sufficient data. The data was collected through Twitter public API which allowed developers to extract data programmatically. The collected data was filtered due to the casual nature of tweets. NamSor tool was used classify the gender of each tweet. The work was carried out by various stages such as data extraction, data preprocessing, and implementation phases. The data extracted through twitter public API, "twitteroauth" version was used, the twitteroauth is implemented in the PHP so it gave direct access through

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

local host and web servers. Based on the set of parameters the query is structured, the relevant data is produced as an output to the browser which is directly stored under MYSQL database. The research work was focused on region USA and major cities of USA. The comparisons of graphically presented data were done with respect to the real world scenarios to find the perfection of methodology. The drawback is the demonstrated work was reasonably good except for good data availability and challenges of data preprocessing. Better NLP tools can be introduced to reduce the percentage of unfit data.

Bogdan Batrinca et al [6] 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.

Wei-Hao Chang et al [7] demonstrates data collection and analysis from social network which is termed as Profile Analyzer System (PAS), in the Profile Analyzer system, they analyses individual insights such as personal behavior, habits, personality so on. In their research paper they talk about the following:

1. Data collection from websites
2. Kind of data source and data type is accepted in their research
3. Data dimension description

Data Collection is carried through several Data crawler tools such as WebHarvest (Java) and Crawler4j (Java). WebHarvest (Java) which focuses on html/xml websites. Crawler4j (Java) tool is executed with apache ant which is helpful in multi-thread tasks. It also has an important feature which can automatically detect character issues.

Data type and Data Sources: In PAS system the raw data is segregated as primary, secondary and ternary information. The first set of information is about Personal information such as birth date, general location or habits etc. The second set of data is about like, weibo posts, personal tages, re-twitter counts. Finally the information related friends, which is achieved through XFN (Xhtml Friends Network).

Dimension Efficiency: There are various dimensions considered for the data influences, few are:

- Social Network Effect
- Education/Job/Living Background Effect
- Brand Effect
- Buying Behavior/Consumption Concept Effect
- Religion Belief Effect and so on

The process of Profiling Analysis took place in different stages such as Keywords and characteristics mapping, weight calculations, sketching portraits and portrait visualization. In keywords and characteristic Mapping kNN (K nearest neighbor algorithm) is used to filtrate the collected datasets. After which matching process is executed through a mapping library, if the matching does not correspond to the characters then fuzzy mapping is used for matching process.

Direct and indirect relation weighs calculation acquitted through Perception Learning Algorithm (PLA), finally profiling scratch is expressed through Random Forest Classification algorithms and the analyzed data is represented through Visualizing techniques.

Julian Buhler et al [8] in his research paper Big Data, Big Opportunities provides explorative case study and proposed the hypothesis such as H1, H2, H3 and H4. Whereas H1: are users of entertainment-oriented social media services; H2: Asian users of entertainment-oriented social media services; H3: Availability in a native language moderates the effect of users striving for business-oriented social media; H4: Age moderates the effect of striving for social media presence for all users. Finally it concludes Business- Oriented social media services are less present in Asia. In contrast, business-oriented service providers would utilize premium functions and account features

Summary of Literature Review

References	Year	Model	Algorithm and tools used	Domain
Haishuai Wang et al [4]	2015	Social Analysis	Peak identification algorithm, VSM, String matching, Levenshtein distance and SimHash	Product
SAA Hridoy et al [5]	2015	Twitter	Namsor, POS tag, SNLP	Phone
Bogdan Batrinca et al [6]	2014	Social Media Analytics	-	-
Wei-Hao Chang et al [7]	2014	PAS	kNN algorithm, Perception Learning algorithm, Random Forest Classification, WebHarvest, Crawler4j	Product
Julian Buhler et al [8]	2015	Hypothetical	Case study	Product

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

CONCLUSION

This study surveyed on social media, social analysis, data mining, big data, data collection, data extraction and preprocessing methods. In this review the works carried out for SocialAnalysis, Social media, Profile Analyzer System are discussed, Algorithms such as Peak identification algorithm, VSM, String matching, Levenshtein distance, SimHash, Namsor, POS tag, SNLP, kNN algorithm, Perception Learning algorithm, Random Forest Classification, WebHarvest, Crawler4j were identified as important algorithms to be used for the Social Analysis.

REFERENCES

[1] Donghee Sinn and Sue Yeon Syn, "Personal documentation on a social network site: Facebook, a collection of moments from your life?," Archival Science, vol. 14, no. 2, pp. 95-124, 2014.
[2] Irena Pletikosa Cvijikj, Erica Dubach Spiegler, and Florian Michahelles, "Evaluation framework for social media brand presence," Social Network Analysis and Mining, vol. 3, no. 4, pp. 1325-1349, 2013.
[3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, no. 3, p. 37, 1996.
[4] Haishuai Wang, Peng Zhang, Ling Chen, and Chengqi Zhang, "SocialAnalysis: A Real-Time Query and Mining System from Social Media Data Streams," in Australasian Database Conference, 2015, pp. 318-322.
[5] Syed Akib Anwar Hridoy, M Tahmid Ekram, Mohammad Samiul Islam, Faysal Ahmed, and Rashedur M Rahman, "Localized twitter opinion mining using sentiment analysis," Decision Analytics, vol. 2, no. 1, p. 1, 2015.
[6] 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.
[7] Wei-Hao Chang, Bing Li, and Xin Fang, "Data Collection and Analysis from Social Network--Profile Analyzer System (PAS)," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2014, pp. 765-772.
[8] Julian Buhler, Aaron W Baur, Markus Bick, and Jimin Shi, "Big Data, Big Opportunities: Revenue Sources of Social Media Services besides Advertising," in Conference on e-Business, e-Services and e-Society, 2015, pp. 183-199.
[9] Sophia Karagiorgou, Dieter Pfoser, and Dimitrios Skoutas, "Geosemantic network-of-interest construction using social media data," in International Conference on Geographic Information Science, 2014, pp. 109-125.
[10] Giacomo Inches and Fabio Crestani, "An introduction to the novel challenges in information retrieval for social media," in Bridging between Information Retrieval and Databases: Springer, 2014, pp. 1-30.