International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Issues in Mining Techniques in Social Media

Jyoti More
Department of Computer Engineering, Lokmanya Tilak College of Engineering, Navi Mumbai, Maharashtra

ABSTRACT

Social network has acquired substantial attention in the last few decades. Access to social network sites such as Twitter, Facebook, LinkedIn and Google+ through the internet has become more affordable. Social media are the online platforms that provide a way for people to connect with each other and participate actively in the group conversations. For individuals, it is a source of communication that helps sharing contents with friends and like-minded people. For businesses, it provides an access to various issues like what people say about their brand, their product and/or service, to know who are the dominating individuals or possible influencers for their new ideas and then use these findings to make better business strategies. The social media also can be exploited for viral marketing to grow the business spectrum. It is found that there is a growing interest in social networks and people are depending on social networks for information, news and opinion of other users in various fields. The growing trust on social network sites causes people to generate massive data, also called big data. It is typically characterised by three computational issues namely; volume, velocity and variety. These issues in turn make social network data very complex to analyse manually, resulting in the need to use computational means of analysing them. Data mining provides a wide range of techniques for detecting useful knowledge from massive datasets like trends, patterns and rules. In this paper we discuss the different issues with social networks, different approaches, issues, current challenges and trends.

Keywords: Social Networks, Social Network Mining, Social Network Analysis (SNA), Modelling, challenges in social networks

I. INTRODUCTION

Social network analysis (SNA) is the process of finding social structures through the use of networks and graph theory. It can be represented as networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect the nodes, to form the network. Social network analysis has emerged drastically as a significant technique in modern sociology. It has also gained attention in the diverse fields such as in anthropology, biology, communication studies, economics, geography, history, information science, organizational studies, political science, social psychology, development studies, sociolinguistics and computer science, business intelligence, etc. Social media marketing is the use of social media platforms and websites to promote a product or service has also flourished to a large extent. Most of these social media platforms have their own intelligent built-in data analytics tools, which enable companies to track the progress, success, and engagement of advertisement campaigns. Companies access a range of stakeholders through social media marketing including existing and potential customers, employees, journalists, bloggers, and the general public. On a strategic level, social media marketing includes the management of the implementation of a marketing campaign, governance, setting the scope and the establishment of a firm's desired social media usage. To use social media effectively, firms should learn to allow customers and Internet users to post user-generated content and use that data for analysing and improving their business strategies. This kind of strategy can also called mining strategy as it allows to learn the pattern behaviour of the users and take the decisions accordingly.

Some common network analysis applications of social network analysis include data aggregation and mining, network propagation modelling, network modelling and sampling, user attribute and behaviour analysis, community-maintained resource support, location-based interaction analysis, social sharing and filtering, recommender systems development, and link prediction and entity resolution. In the private sector, businesses use social network analysis to support activities such as customer interaction and analysis, information system development analysis, marketing, and business intelligence needs. Some public sector uses include development of leader engagement strategies, analysis of individual and group engagement and media use, and community-based problem solving[1].

II. SOCIAL NETWORKS MODELLING

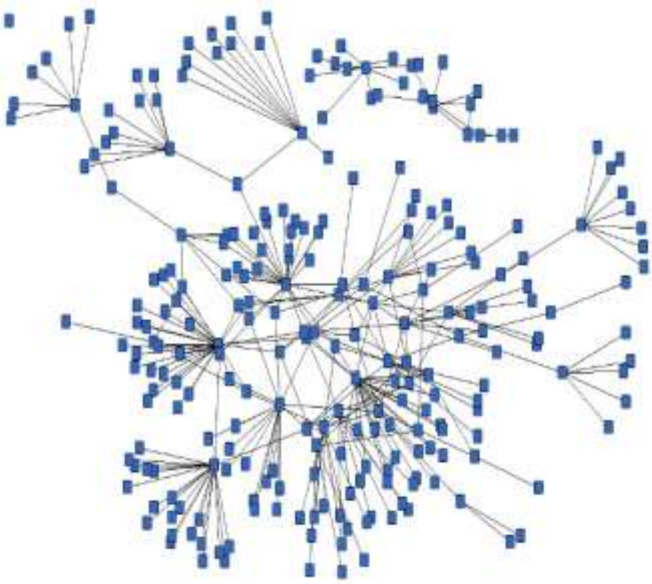

Figure 1. Social Network Visualization

Social media is not about what each one of us does or says, but about what we do and say together, worldwide, to communicate in all directions at any time, by any possible (digital) means. Social Media are the platforms that enable the interactive web by engaging users to participate in, comment on and create content as means of communicating with their social graph, other users and the public. Social media has the following characteristics:

- Encompasses wide variety of content formats including text, video, photographs, audio, PDF and PowerPoint. (Social content is a by-product of creating content with your community.)
- Allows interactions to cross one or more platforms through social sharing, email and feeds.
- Involves different levels of engagement by participants who can create, comment or lurk on social media networks.
- Facilitates enhanced speed and breadth of information dissemination.
- Provides for one-to-one, one-to-many and many-to-many communications.
- Enables communication to take place in real time or asynchronously over time.
- Is device indifferent. It can take place via a computer (including laptops and notebooks), tablets (including iPads, iTouch and others) and mobile phones (particularly smartphones).
- Extends engagement by creating real-time online events, extending online interactions offline, or augmenting live events online

Social network mining is a growing research area which aims at bringing together researchers from different fields such as machine learning, data mining, artificial intelligence, optimization, graph theory, networks, mobile computing and other areas, with the goal of attacking important problems that the birth of social networks has brought into the scientific arena.

SNA provides a rich set of metrics, many of which are used in many social network analysis experiments. They are used to formally define the structural properties of Social Network.

Degree Centrality:

Degree centrality is simply the number of direct relationships that an entity has. An entity with high degree centrality:

- Is generally an active player in the network.
- Is often a connector or hub in the network.
- Is not necessarily the most connected entity in the network (an entity may have a large number of relationships, the majority of which point to low-level entities).
- May be in an advantaged position in the network.
- May have alternative avenues to satisfy organizational needs, and consequently may be less dependent on other individuals.
- Can often be identified as third parties or deal makers.

Betweenness Centrality:

Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network. An entity with a high betweenness centrality generally:
It holds a favoured or powerful position in the network.

- Represents a single point of failure-take the single betweenness spanner out of a network and you sever ties between cliques.
- Has a greater amount of influence over what happens in a network.

Closeness:

Closeness centrality measures how quickly an entity can access more entities in a network. An entity with a high closeness centrality generally:

- Has quick access to other entities in a network.
- Has a short path to other entities.
- Is close to other entities.
- Has high visibility as to what is happening in the network.

Eigenvalue:

Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, Eigenvalue identifies the most central entities in terms of the global or overall makeup of the network. A high Eigenvalue generally:

- Indicates an actor that is more central to the main pattern of distances among all entities.
- Is a reasonable measure of one aspect of centrality in terms of positional advantage.

III. DIFFERENT APPROACHES FOR MINING SOCIAL MEDIA

1. **Graphical Approach:**

Structure mining or structured data mining is the process of finding and extracting useful information from semi-structured data sets. Graph mining, sequential pattern mining and molecule mining are special cases of structured data mining. Jun Zang et. al [2] proposed that the directionality is a significant but inherent property of social ties, though usually ignored in undirected social networks due to its invisibility. Most social ties are natively directed, and the perception of directionality can improve the understanding about the network structures and further benefit other tasks upon social networks.

2. **Community Detection:**

Given a network, it is particularly interesting as well as challenging to detect the inherent and hidden communities. Communities, which have no quantitative definition, are also called clusters. They are usually considered as groups of nodes, in which intra-group connections are much denser than those inter-group ones. There are several algorithms like division algorithms in hierarchy clustering methods, direct partitioning methods, label propogation methods, leadership expansion, clique percolation, matrix blocking, skeleton clusteing etc. All these approaches are well analysed by M. Wang et.al [3].

3. **Recommender System:**

With the advent and popularity of social network, more and more users like to share their experiences, such as ratings, reviews, and blogs. The new factors of social network like interpersonal influence and interest based on circles of friends bring opportunities and challenges for recommender system (RS) to solve the cold start and sparsity problem of datasets. Some of the social factors have been used in RS, but have not been fully considered. Xueming Qian et. Al. [4] their paper, proposed that three social factors, personal interest, interpersonal interest similarity, and interpersonal influence, fuse into a unified personalized recommendation model based on probabilistic matrix factorization.

4. **Opinion Mining:**

It is about understanding the opinions of the general public and consumers toward social events, political movements, company strategies, marketing campaigns, and product preferences. Many new and exciting social, geopolitical, and business-related research questions can be answered by analysing the thousands, even millions, of comments and responses expressed in various blogs (such as the blogosphere), forums (such as Yahoo Forums), social media and social network sites (including YouTube, Facebook, and Flickr), virtual worlds (such as Second Life), and tweets (Twitter). Opinion mining, a sub discipline within data mining and computational linguistics, refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content. Sentiment analysis is often used in opinion mining to identify sentiment, affect, subjectivity, and other emotional states in online text [5].

5. Sentiment Mining:

Long Jin et.al [6] investigated traffic activity from a network perspective. They also concentrated on the characteristics of social behaviours in mobile environments. They reviewed the malicious behaviours of OSN users, and discussed several solutions to detect misbehaving users. Their survey serves the important roles of both providing a systematic exploration of existing research highlights and triggering various potentially significant research**.**

6. Healthcare Mining:

Recent work in machine learning and natural language processing has studied the health content of tweets and demonstrated the potential for extracting useful public health information from their aggregation. Mark Dredze [7] examines the types of health topics discussed on Twitter, and how tweets can both augment existing public health capabilities and enable new ones. The author also discusses key challenges that researchers must address to deliver high-quality tools to the public health community.

IV. ISSUES IN SOCIAL NETWORK MINING

Security Issues in social media can be a great concern to many of the scientists working in this area.

Hongyu Gao et. Al [8] in their paper did the surveys of security issues and available defence mechanisms regarding popular online social networks. It covers a wide variety of attacks and the corresponding defence mechanisms, if available. The authors organize these attacks into four categories - privacy breaches, viral marketing, network structural attacks, and malware attacks - and focus primarily on privacy concerns. They offer an in-depth discussion of each category and analyse the connections among the different security issues involved.

Privacy preservation is another such issue while dealing with social media. Online social networks, such as Facebook, are increasingly utilized by many people. These networks allow users to publish details about themselves and to connect to their friends. Some of the information revealed inside these networks is meant to be private. Yet it is possible to use learning algorithms on released data to predict private information. Raymond Heatherly et. al. [9] in their paper, explore how to launch inference attacks using released social networking data to predict private information. Later they devised three possible sanitization techniques that could be used in various situations. They explored the effectiveness of these techniques and attempted to use methods of collective inference to discover sensitive attributes of the data set.

Geo-social networks (GeoSNs) provide context-aware services that help associate location with users and content. The proliferation of GeoSNs indicates that they're rapidly attracting users. GeoSNs currently offer different types of services, including photo sharing, friend tracking, and "check-ins." However, this ability to reveal users' locations causes new privacy threats, which in turn call for new privacy-protection methods. Carmen Ruiz Vicente et al. [10] study four privacy aspects central to these social networks - location, absence, co-location, and identity privacy - and describe possible means of protecting privacy in these circumstances.

Other privacy issues in social networks include the following[1] Social networking worms, Phishing bait, Trojans, Data leaks, Shortened links, Botnets, Cross-Site Request Forgery (CSRF), Impersonation, etc.

V. CHALLENGES IN MINING SOCIAL MEDIA DATA

Researchers must overcome a number of hurdles in order to find insight in social media data. Social media and Big Data have radically changed how people communicate, interact, work, conduct business, and more. The buzz around these phenomena continues to grow as people and organizations begin to tap into the potential of a socially connected, data-rich world. The excitement is hard to overstate, and there are many examples to illustrate the insights and benefits that can be gained by examining social media data. However, some very real hurdles stand between the unending supply of data and those who would like to mine and use it.Dr. Huan Lui [1] is one of the scholars tackling the challenges inherent in mining social media data. Lui discussed several concrete challenges researchers encounter when they mine social media data for answers to business or other questions:

- **Evaluation dilemma**: Traditional data mining often segregates a portion of the dataset for testing. This provides a means to develop and evaluate models against some kind of ground truth. But with social media data, traditional test data may not be viable, forcing researchers to consider how they will evaluate their claims in the absence of an identified ground truth.

- **Sampling bias**. Because Big Data is so big, APIs and scraping data often return only a small sample of the whole. A relatively small sample size can be biased, threatening the credibility of research results derived from the sample.

- **Noise-removal fallacy**: Posts about what people eat for breakfast are just one source of noise in social media data. But removing the noise can render data from Twitter and other sources useless, according to Lui. The inherently linked nature of social media data further complicates the task and requires researchers to approach noise-removal differently than they would with attribute-value data.

- **Studying distrust in social media**: Trust is a critical human construct that plays a role in many decisions and actions. According to Lui, Distrust may play an important role in consumer consumer decisions. Challenges arise from a lack of computational understanding of distrust with social media data and also as the absence of information during the study.

- **Deception detection**: Information intended to deceive can spread though social media the same as valid information. This raises questions of how to detect different types of deception (e.g., manipulating information, changing context, or outright fabrication) in different social channels and formats (e.g., text, link, audio, photo, video, multimedia).

VI. CONCLUSION

Social media mining uses a range of basic concepts from computer science, data mining, machine learning and statistics. Social media miners develop algorithms suitable for investigating massive files of social media data. Social media mining is based on theories and methodologies from social network analysis, network science, sociology, ethnography, optimization and mathematics. Mining is all about study of different patterns. These patterns and trends are of interest to companies, governments and not-for-profit organizations, as these organizations can use these patterns and trends to design their strategies or introduce new programs (or, for companies, new products, processes and services). Hence we need to thoroughly understand the significance and spectrum of social media. This paper attempts to provide an overview of all the basics of social media analytics, possible approaches, scope of mining, issues and different challenges.

VII. REFERENCES

[1]. Reza Zafarani, Mohammad Ali Abbasi, Huan Liu, "Social Media Mining", Cambridge University Press, 2014

[2]. Jun Zhang, Chaokun Wang, Jianmin Wang, Jeffrey Xu Yu, Jun Chen, Changping Wang, "Inferring Directions of Undirected Social Ties", IEEE Transactions on Knowledge and Data Engineering , Volume: 28, Issue: 12, Dec. 1 2016

[3]. M Wang, C Wang, JX Yu, J Zhang, "Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework", Proceedings of the VLDB Endowment 8 (10), 2015

[4]. Xueming Qian, He Feng, Guoshuai Zhao, Tao Mei, "Personalized Recommendation Combining User Interest and Social Circle" IEEE Transactions on Knowledge and Data Engineering, Volume: 26, Issue: 7, July 2014

[5]. Hsinchun Chen David Zimbra, "AI and Opinion Mining", IEEE Intelligent Systems, Volume: 25, Issue: 3, May-June 2010.

[6]. Long Jin, Yang Chen, Tianyi Wang Pan, Hui Athanasios, V. Vasilakos, "Understanding user behavior in online social networks: a survey", IEEE Communications Magazine, Volume: 51, Issue: 9, September 2013.

[7]. Mark Dredze, "How Social Media Will Change Public Health", IEEE Intelligent Systems, Volume: 27, Issue: 4, July-Aug. 2012.

[8]. Hongyu Gao, Jun Hu, Tuo Huang, Jingnan Wang, Yan Chen, "Security Issues in Online Social Networks" IEEE Internet Computing, Volume: 15, Issue: 4, July-Aug 2011

[9]. Raymond Heatherly, Murat Kantarcioglu, Bhavani Thuraisingham, "Preventing Private Information Inference Attacks on Social

Networks", IEEE transactions on Knowledge and Data Engineering, Volume: 25, Issue: 8, Aug. 2013.

[10]. Carmen Ruiz Vicente, Dario Freni, Claudio Bettini, Christian S. Jensen, "Location-Related Privacy in Geo-Social Networks", IEEE Internet Computing, Volume: 15, Issue: 3, May-June 2011