“Data Mining in Social Media Data: A Review”

Kalokhe Anil Sopan, Pawar Mahesh Dattatray

1 Asst. Professor, Department of Computer Application, Vidya Pratishthan’s, Arts, Science and Commerce College, Baramati, MH (India).
2 Research Student, Swami Ramanand Teerth Marathwada University, Nanded.
Email: anil.kalokhe@gmail.com, mdpawar77@gmail.com

Abstract: Data Mining is a powerful tool which helps the extraction of projecting information from large data sets which helps to incorporate to focus on the most important information in data storeroom. Data mining techniques provides practitioners and researchers the tools which are needed to analyze large, complex and frequently changing social media data. It uses various statistical, graphical methods and machine learning methods to separate the knowledge in to a form which is useful for many real life applications. Social network analysis has become a very popular field of a modern research because it is highly useful for many applications. In this paper we have reviewed various data mining techniques used for social media analysis.

Keywords: Data mining, social media, Social Networking sites, Social Network Analysis

Introduction:
Now a days, People uses Social media platform to discuss their ideas, issues and opinions. Social Media refers to the interactions among peoples in which they create share, exchange and comments contents among themselves in virtual communities and networks. Today Social media is an important part of our life from shopping to electronics mails, education and business rules. Social media plays an important role for transforming people’s life style. Social media includes social networking sites like Facebook, Twitter, Whatsapp and blogs where peoples can communicate with each other [3].With the huge use of social media via the internet, an unprecedented amount of data is available [5]. These data can be analyzed using different data mining tools and techniques. Social Network Analysis (SNA) is defined as the study of social networks in order to understand social network structure and behavior [1].

1. Social Networks:
A social network is a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors[6]. A Social network is a heterogeneous and multi relational dataset represented by a graph. Vertexes represent the objects (entities), edges represent the links (relationship or interaction) and both objects and links may have attributes. Social networks are usually very large for representing many real world applications [1]. Network construction from general, real world data presents various unexpected challenges owing to the data domains themselves (e.g., information extraction and preprocessing) and to the data structures used to knowledge representation and storage [2]. Social media platforms allow users to have conversations, share information and create web content.
A Social network can be generally understood to be a kind of computer application which facilitates the creation or definition of social relations among the people based on general interest, activities, professional interests, associative relations, family and so on[1]. Finding and representing a social network from a data source can be a difficult problem. This challenge is due to many factors namely the ambiguity of human language, incompatible representation of information, multiple aliases for the same user and the ambiguity of relationships between individuals.

Social media refers to a variety of information services used to collaboratively by many people placed into the subcategories shown in following table [5].

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogs</td>
<td>Blogger, LiveJournal, Wordpress</td>
</tr>
<tr>
<td>Microblogs</td>
<td>Twitter, GoogleBuzz</td>
</tr>
<tr>
<td>Opinion Mining</td>
<td>Epinions,Yelp</td>
</tr>
<tr>
<td>Photo and Video Sharing</td>
<td>Flickr, YouTube</td>
</tr>
<tr>
<td>Social bookmarking</td>
<td>Delicious, StumbleUpon</td>
</tr>
<tr>
<td>Social Networking sites</td>
<td>Facebook, LinkedIn, MySpace</td>
</tr>
<tr>
<td>Social news</td>
<td>Digg, Slashdot</td>
</tr>
<tr>
<td>Wikis</td>
<td>Scholarpedia, Wikihow, Wikipedia, Event maps</td>
</tr>
</tbody>
</table>

The rise and popularity of social media is Surprising and shocking. For Example, Consider the popular social networking site facebook. During the last 14 years of operation Facebook reached over 2000 million active users [7]. Figure 1.2 illustrates the exponential growth of Facebook during its last thirteen years [17]. The network statistics provides the ranking information about social media websites. Facebook is ranked 1st in the world for internet sites based on the amount of daily internet to the site [7].
2. Pros and Cons of Social Networking:

2.1 Pros of Social Networking [9]
- Ability to connect to other people all over the world.
- Easy and instant communication.
- Real-time news and information discovery.
- Great opportunities for business owners.
- General fun and enjoyment.

2.2 Cons of Social Networking
3. Need of Data mining:

We need data mining to obtain fruitful knowledge from our noisy data and it is achieved through a step by step process. Data mining helps to remove inconsistent data, extract patterns relevant to the analysis task, transform data into forms appropriate for KDD (Knowledge Discovery in Databases) and finally present the minded knowledge to the user through visualization. It helps to sort out the raw data, analyze it and process it into valid, relevant and accurate data effectively; data mining techniques have to be employed. Data mining is the process of analyzing large amounts of data in an effort to find correlations, patterns, and insights [13]

Following diagram shows different types of data mining processes.

The main idea of data mining comes under following two categories [1].

a. Predictive data mining: It creates the model of the system from the given data.

b. Descriptive data mining: It generates significant data sets from the existing data.

4. Data Mining Techniques:

The aim of these above ideas is achieved by the following data mining techniques [1].

- **Characterization**: It is used to generalize, summarize and possibly different data characteristics.

- **Classification**: Data Classification is the process in which the given data is classified into different classes.

- **Regression**: This process is similar to classification. The major difference is that the object to be predicted is continuous rather than discrete.
• **Association**: It discovers the association between various data bases and the association between the attributes of single database.

• **Clustering**: It involves grouping of data into several classes such that it describes the data. It breaks the large data set into smaller groups to make the designing and implementation process to be simple.

• **Change Detection**: This method identifies the significant changes in the data from the previously measured values.

• **Deviation Detection**: It focuses on the deviations between the actual values of the objects and its expected values. This method finds out the deviation according to the time as well the deviation among different subsets of data.

• **Link Analysis**: It traces the connections between the objects to develop models based on the patterns in the relationships by applying graph theory techniques.

• **Sequential Pattern Mining**: This method involves the discovery of the frequently occurring patterns in the data.

Following diagram shows some of the data mining techniques which are applied on social media for mining data.

![Data Mining Techniques Model](image)

**Figure 4.1 Data Mining Techniques Model**

**5. Social Network Analysis and Data Mining:**

Data mining tools can answer industry questions that traditionally were to time consuming to resolve. Data mining of social networks can be done using the graph mining methods such as prediction, classification/topologies, efficiency, pattern detection, measurement and metrics, modeling, evolution and structure, data processing and communities [2]. Here to extract the information represented in graphs we need to define the metrics that describe the global structure of graphs, find the community structure of the network and define metrics that describe the patterns of local interaction in the graph, develop efficient algorithms for mining data on networks and understand the model of generation of graphs [1].

Social network and its analysis is an important part and it is widely spread among many young researches. Social network research emerged from psychology, sociology, statistics and graph theory. Based on graph theoretical concepts a social networks interprets the social
relationships as points and their relationships as the lines connecting them [2]. The various types of social network analysis are

5.1 Socio-Centric(Whole) network analysis

The focus of a whole network analysis is on measuring the structural patterns of those interactions and how those patterns explain outcomes, like the concentration of power or other resources, within the group. The underlying assumption is that members of a group interact more than a randomly selected group of similar size. Socio centric network analysts are interested in identifying structural patterns in cases that can be generalized in a whole network study. The actors of the network are usually known or easily determined. This is because a socio centric network study usually focuses on “closed” networks implying that the boundaries of a whole network are a priori defined [16].

Following are the key elements of socio centric network analysis.

- Emerged in sociology.
- Involves quantification of interaction among a socially well defined group of people.
- Focus on identifying global structural patterns.
- Personal network analysis
- Involves concentrations on interactions between an individual and all other Person related to ego.

5.2 Knowledge(Egocentric/Personal) based network analysis

In the social network parlance, the person we are interested in is referred to as the “ego” and the people referred to by the “ego” as his affiliate, advisor, friend, or relative, are known as “alters”[16].

Following are the key elements of egocentric based network analysis.

- Emerged in Computer Science
- Make generalizations of features found in personal networks.
- Knowledge discovery based on entities associated with actors in the social network.

Conclusion:

This study surveyed on social media, social analysis, data mining, data collection, and social data analysis methods. In this Valuable information is hidden in vast amounts of social media data, presenting ample opportunities for collaboration between computer scientists, social scientists to use data mining technologies. This paper also focuses on how active users on social media are increased day to day. Data mining provides proficient way to execute and make use of database. In this paper we have briefly reviewed the various data mining techniques which are used for social network analysis. Social media data are vast, noisy, distributed, unstructured, and dynamic, which poses novel challenges for data mining. In this paper, we offer a brief introduction to mining social media.
References:

[1]: Saima Jan, Rahila Ruby, Peer Taha Najeeb, Mudasir Ahmed Muttoo, Social network analysis and data mining, Volume.6, Issue 6, June 2017
[3]: Shabnoor Siddiqui, Tajinder Singh, Social Media its Impact with Positive and Negative Aspects Volume 5– Issue 2, 71 - 75, 2016, ISSN:- 2319–8656
[5]: Geoffrey Barbier, Huan Liu, Data Mining in Social Media
[7]: https://www.google.co.in/search?q=active+user+on+facebook+images&tbm=isch&sa=X&ved=0ahUKEwicwMfao5PZAhWKLo8KHch-ANAJQ9QEJkDA#imgrc=qMTH3Mxj85lR4M:
[10]: Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, Frederic Stahl, A Survey of Data Mining Techniques for Social Network Analysis
[11]: France Cheong, Christopher Cheong, Social Media Data Mining: A Social Network Analysis of Tweets during the 2010-2011
[14]: https://link.springer.com/article/10.1007/s00146-014-0549-4
[15]: Annie Syrien, M. Hanumanthappa , A Study on Social Network Analysis through Data Mining Techniques -A detailed Survey
[16]: Kenneth K S Chung, Liaquat Hossain, Joseph Davis, Exploring Socio centric and Egocentric Approaches for Social Network Analysis
[17]: https://www.google.co.in/search?q=active+user+on+twitter+from+foundation+in+graph&oq=active+user+on+twitter+from+foundation+in+graph&gs_l=unsupported-result-0-21134837618:21134837618-0-13893661593-0-13893661593-0-0-0-502-1-162-3-0-0-0-0a8 websitelinksrecognizable-0-1191-a:wy6QnZQyL-0