UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

Survey on Analytical Techniques for Social Networking Service

Hyeuk Kim[1]

Assistant Professor, Division of Global Management Engineering, Hoseo University, Asan, Korea[1]

Abstract: We are living in big data era. Massive data are generated every day as the internet develops. The development of internet yields new service which is called social networking service. New analytic methods are also developed to analyse data from the social networking service. We review text mining, sentiment analysis, social network analysis, and text clustering which are widely used for the analysis of social networking service. We also point out the characteristics of the above techniques.

Keywords: Social networking service, Text mining, Sentiment analysis, Social Network analysis, Text clustering.

I. INTRODUCTION

One of the most interesting issues is Big Data in the twenty-first century. Huge data are generated every day in the digital world. Gartner defines that big data is high-volume, high-velocity, and high-variety information that can enhance the procedure and create value in 2012 [1]. 3V [2] is a word that describes the characteristics of big data. The first V means the volume. It is the physical size of data. The size of analysed data increases from a terabyte to several petabytes. The second V means the variety of data. Big data is categorized into three types: structured data, semi-structured data, and unstructured data [3]. Structured data is usually fixed in the form like a relational database management system and spreadsheets. Semi-structured data is not maintained in the fixed fields, but they conclude meta data or schema such as XML and HTML. Unstructured data is generated and available nowadays. They are not saved in the fixed fields and can be found on the internet as the forms of documents, images, videos, and speeches. The portion of unstructured data is growing very quickly. The third V is velocity. It is about generating and handling data. Data have to be handled and analysed fast since they are generated much and quickly. Twitter generates over 7 terabytes for users' conversations and Facebook handles over 10 terabytes by users every day. Therefore, near real-time or real-time data processing is essential nowadays. It has passed several years since big data becomes popular and another V(value) is appeared to describe the characteristic of big data. In the past, people focus on the technical aspects of big data and the interest on big data is moved since many applications are developed. Social networking service is the main cause of the development of big data and many techniques are developed and used for SNS data. In the research, we investigate and summarize major skills for analysing social networking service data. The paper is constituted of four sections. This section is about an introductory part and belongs to the first section. The second section introduces the environment for big data analysis. The analytical techniques for social networking service are described in the third section. It concludes in the last section.

II. BIG DATA ENVIRONMENT

The system environment is very important to analyse big data since it require huge resources. The environment can be divided into five specific parts such as distributed computing, cloud computing, Hadoop [4], NoSQL, and R. One of the most serious problems is huge volume of data in big data. Computer capacity must be increased to handle and analyse data as the size of data increases. Cluster-based distributed computing is an easy solution for handling huge size of data. The distributed computing is the computer system that connects multiple servers through network. Researcher can increase the capacity by adding a node whenever he needs extra capacity for data. The approach has high availability since other node can be used instead of broken node without stopping operation. Recovery speed is also fast when nodes are used in a cluster-based distributed computing. A repairman fixes not the whole system but the defected node. A distributed computing skill is one of the most important essentials to support applications for big data. It is the technique that the combined computers act like one computer. The second part which constitutes big data environment is cloud computing. Cloud means internet and cloud computing is the computing technique which is based on internet. Cloud computing is the environment that user can use any application without installing in his own computer and share data through network equipment simultaneously. For example, contents like documents, pictures, and videos are saved outside a computer in cloud computing. The advantages of saving contents outside are as follows. Users can access and modify data on any time and any place. Service is still available or damaged partly even though there is a natural disaster. The key skill in cloud computing is to use multiple servers to keep and handle massive data. It is the

DOI10.17148/IARJSET.2017.4820

distributed computing which is mentioned above. Therefore, both skills are complementary. Hadoop, which is an open-source platform, has been developed for handling massive data. The name was decided by the developer Doug Cutting and from the toy elephant's name. Hadoop consists of HDFS(Hadoop Distributed File System) and MapReduce [5]. Data is saved by the form of HDFS and handled through MapReduce. HDFS is composed of Job Tracker and Task Tracker and performs to read and work data from the saved files. Job Tracker is usually located on the server like a name node and Task Tracker is located on the same server with data node in HDFS. The inputted data are separated out several pieces and processed in parallel. The action of MapReduce consists of two parts such as map and reduce. All input and output data in map and reduce procedures are processed by the pairs of key and value. NoSQL is an abbreviation of 'Not Only SQL' and means non-relational database unlike the traditional relational database. NoSQL makes user to handle structured and unstructured data easily and effectively. The popular application of NoSQL is MongoDB, HBASE, and so on. R is an open-source application which supports statistical and graphical analysis for data [6]. R becomes popular since it is general public license and many packages for extra analysis are developed and available free. R is a kind of programming language and handles the tasks from the basic statistical analysis to the recent machine learning techniques like Deep Learning [7]. Another advantage of R is that it can be connected to other programming language such as Java, C, and Python. Huge IT companies such as Google, Facebook, and Amazon use R for massive data analysis since it provides the distributed processing in Hadoop environment. The logos for Hadoop and R are descried in Figure 1 and 2.

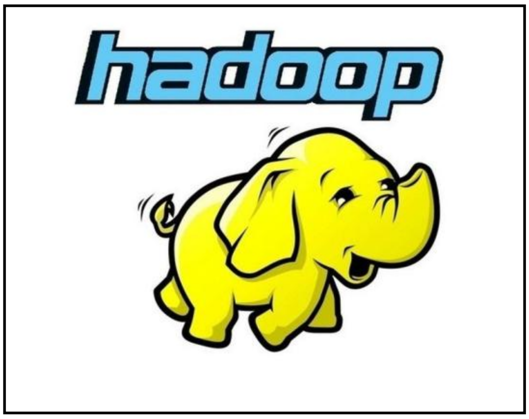

Fig. 1 Logo of Hadoop

Fig. 2 Logo of R

III. ANALYTICAL TECHNIQUES FOR SNS DATA

Text mining is a mining technique which extracts pattern or relationship from unstructured text data and discovers meaningful information and value [8]. The technique is based on natural language processing skill. Basically, data from social media are analysed by text mining because they belong to unstructured data. Text mining is the process to collect and analyse massive unstructured data form SNS, websites in real time or near real time, and finally figure out the sentiment or the intent of users. The meaning of text mining is a little different from the meaning of data mining. Data mining is the technique to extract information from the structured data in a relational database, whereas text mining is the technique to extract information from the unstructured data. A graphical method in text mining is a word cloud. It is the text mining technique which displays the keywords which are appeared frequently in text. The word is written larger as its frequency is higher. Therefore, a reader can find that which words are used frequently in the document.

DOI10.17148/IARJSET.2017.4820

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

Twitter text mining is very popular methods in text mining since the documents in Twitter has the limitation in the size and many people in the word use it. It handles the sentences from Twitter. Some terms used in Twitter are summarized in Table 1 [9].

TABLE I TERMINOLOGIES IN TWEETER

Term	Description
Tweet	It is the document written in Twitter. It has to be written within 140 characters.
Following	It is about adding friends. A user can add other user as a friend.
Follower	It is the opposite notion of following. It describes other user who adds a user as his friend.
Retweet(RT)	It is the document from other's tweet. Its function is to help a user's followers share a specific tweet written by other user.
Timeline	Tweets and Followings' tweets are displayed in the space called Timeline. They are in chronological order.

Twitter also supports the application programming interface for an analyst to use Twitter's information outside. Many companies use Twitter's data for their service and develop a client application. Therefore, a researcher does not need to get a permission to use others information from Twitter when he collects information through Twitter API. A researcher, who wants to collect data through Twitter API, has to register his account on the homepage of Twitter first and pass an open authorization.

Sentiment analysis is the method to classify writers' opinions into one of two sides [10]. They agree or disagree for the specific subject. It is also called opinion mining. Sentiment analysis is to guess a writer's emotion from his document and it focuses on not the subject that he mentions but his emotion for a specific topic. The analysis is usually applied to evaluations about movies, books, goods. The range of applications is getting broader to the various areas as social network services become popular. One of the examples by using sentiment analysis is as follows. A pharmaceutical company advertises an antidepressant treatment and decides to reflect the result of sentiment analysis. The moods of the target people are different based on days. Their moods are best on Friday in one week. Their feelings are worse as the day goes on. On Monday, they are in the most negative moods. Therefore, it is effective for a pharmaceutical company to advertise its product on Saturday or Monday. The basic concept in sentiment analysis is a sentiment score. It is defined as follows. The sentiment score is the remained number of positive words after deleting the number of negative words. A sentiment is in positive opinion if a sentiment score is greater than zero, a sentiment is in negative opinion if a sentiment score is less than zero, and a sentiment is in neutral status if a sentiment score is around zero. An opinion lexicon is needed to decide whether a certain word in the document means positive concept or not. The disadvantage of an opinion lexicon is that it cannot handle polysemy such as irony and sarcasm even though it is convenient for a researcher to use.

Social network analysis is the skill to analyse the social relationships among individuals or groups structurally and figure out intrinsic relationship [11]. Network is composed of multiple lines which connect distinct points. A point is called a node or a vertex and a line is called a link or an edge in graph theory. The individual such as a person, an organization, and an object is also called an actor. Each actor is denoted by a point and the connection between two actors are linked by a line. There are two types of network which are directed network and undirected network.

Social network is the network that the actors are connected socially. Therefore, SNA(social network analysis) is to analyse the social relationship among the social individuals structurally and figure out the embedded relationship. It is applied to not only social science but also various areas such as economics, biology, medical science, and engineering.

We introduce the basic concepts of graph theory. A graph is the set consists of a vertex and an edge which connects individuals. There are several types of graphs. The directed graph has a direction. An example of the directed graph is the relationship in Twitter which is mentioned above. The opposite type of the directed graph is the undirected graph. The complete graph is defined that the number of edges is same with the number of all possible links from vertices. When the undirected graph has n vertices, the maximum of edges is n(n-1)/2. The maximum of edges for the directed graph is 2 times of the case of the undirected graph, n(n-1). The weighted graph is to assign weights to the edges which are connected between vertices. The weight means the degree of the difference between edges.

Text clustering is to apply cluster analysis to text document. It calculates the similarity between texts and combines the texts with the similar similarity into the same cluster [12]. It is also called document clustering. For example, some people usually talk about picture and camera and other people usually talk about car in Twitter. We can divide whole people into two groups based on their hobbies through text clustering. The individual can be a document, a paragraph, a sentence, and a term in text clustering. The applications of text clustering are near duplicates detection, search engine optimization, recommendation system, and document summarization. The brief explanations of applications are

DOI10.17148/IARJSET.2017.4820

described below. Near duplicates is the documents which are not exactly same but almost same. Text clustering researcher says that 20~30% documents in blogs belong to near duplicates. The search engine optimization is the task that a website is displayed in the front page when it is searched in search engine. The recommender system is to recommend the interesting materials such as movie, music, book, news, and image to users. Users who choose contents based on the recommender system do not need to consider massive materials for selection and producers offer personalized contents and increase the profit with a new system. The document summarization is to summarize long documents. It is useful on an electronic library, knowledge management system, and the tasks related to books. When we compress or summarize the text, the notion of TF(term frequency) is widely used. TF means the number of appearances in the document. The word is much related to a certain document if it appears frequently in the corresponding document. TF is expressed by term-document matrix of which row is for term and column is for document. The advantage using TF is that it is very simple and able to quantify information. The disadvantage using TF is that it is hard to distinguish two documents with the same sets of TF even though they have different meanings. Standardization and normalization for TF are applied to reinforce TF method.

IV. CONCLUSION

Huge data have been generated in the past, but they were erased without analysis. Nowadays people start to analyze and use data as the techniques for hardware and software develop. The amount of data generated one year is over 3 zeta bytes. Social network service has been started from 1994. They have been developed fast since then, and they influence our life much now. Many new techniques are developed because data form from the social networking service are different from the ordinary data. The common point of them is that their algorithms are from the various fields such as mathematics, statistics, and computer science. Another common point is that they can be developed more since their history is short. Text mining, sentiment analysis, social network analysis, and text clustering are introduced as the techniques to analyze the social networking service in the paper. The definition and some important concepts are described. We are able to apply one of the techniques or multiple techniques which are introduced in the paper whenever we encounter data for social networking service and create value which is essential to Big data era.

REFERENCES

[1] Gartner, “What is big data,” Gartner IT Glossary, 2017.
[2] D. Laney, “3D data management: Controlling data volume, variety and velocity,” Meta Group, 2001.
[3] W. Khattak and P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, 1st ed., Prentice Hall, 2016.
[4] “Hadoop Releases,” Apache Software Foundation, 2014.
[5] Google, “MapReduce,” Google Research Publication, 2016.
[6] K. Hornik, “R FAQ,” Comprehensive R Archive Network, 2015.
[7] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” IEEE Transactions on Pattern Analysis and Machine Learning, 35(8), pp. 1798-1828, 2013.
[8] S. M. Weiss and N. Indurkhya, Fundamentals of Predictive Text Mining, 1st ed., Springer, 2015.
[9] I. Lamont, Twitter in 30 minutes: How to connect with interesting people, write great tweets, and find informat, 1st ed., i30 Media Corporation, 2013.
[10] P. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” Proceedings of the Association for Computational Linguistics, pp. 417-424, 2002.
[11] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, 1st ed., Cambridge University Press, 1994.
[12] C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, 1st ed., MIT Press, 1999.

DOI10.17148/IARJSET.2017.4820