A Literature Review on Social Networks Data Analysis

Abstract
For one thing, with the rapid development and innovation of Internet, we are surrounded by large index data. Every day a large number of data from the electronic commerce, science, social networks and other massive amounts of data are stored in the Internet and various kinds of data storage devices. For another, the development of mobile Internet promoted the development of the social network, we can use Twitter, Facebook, Weibo and other social networking platform to publish ourselves views, opinions and mood, through social networks make our contacts more richly. The richness of social network provides unprecedented opportunities for data analytics. In this paper, we review different information analysis techniques; starting with the analysis of different hashtags, network-topology of social networks, event spread over the network, identification of influence, and finally analysis of sentiment.

Keywords: Data analysis, big data, social network, Twitter,

Introduction

The growing phenomena of social networks, such as: Facebook, Twitter, Weibo, and Instagram, with each one has its own characteristics and its usages, are constantly affecting society. Facebook, for example, is considered as a social network where everyone in the network has a reciprocated relationship with another one in the same network. The relationship in this case is undirected. Conversely, in Twitter everyone in the network does not necessarily have a reciprocated relationship with others. In this case, the relationship is either directed or undirected. Future research and development work will be addressed.

In this paper, we focus on social networks for data analysis. For example, Twitter is an online news and social networking service where users post and interact with messages, known as "Tweets." These messages were originally restricted to 140 characters, but now, the limit was doubled to 280 characters for all languages except Japanese, Korean and Chinese [1]. And Facebook is an online social media and social networking service, it may be accessed by a large range of devices with Internet connectivity, such as desktop, laptop and tablet computers, and smartphones [2]. Compare with Twitter to Facebook, Twitter is accessible for unregistered users to read and monitor most tweets, unlike Facebook where users can control the privacy of their profiles. Twitter is also a large social networking microblogging site. The massive information provided by Twitter such as tweet messages, user profile information, and the number of followers/ followings in the network play a significant role in data analysis, which in return make most studies investigate and examine various analysis techniques to grasp the recent used technologies.

The rest of the paper proceeds as follows: Firstly, we discuss various methods used to

retrieve data, social networks user rankings, and the network topology. Secondly, we discuss some techniques used in information diffusion such as the hashtag life cycle, the network topology, and the forwarding amount. Thirdly, we discuss how other studies gauge the user influence on social networks. Fourthly, we review two approaches for sentiment analysis namely "Natural Language Processing" and "Machine Learning".

Methods

To track and monitor different datasets, most studies [3], [4] began with collecting the desired datasets from social networks, and applied filtering techniques to remove redundant data or spam messages. Then parsed the data into a structured form. Finally analyzed the data. Below we review several types of analyses that most researchers have used.

Datasets

Analyzing structured data have been widely used. In such case, the traditional Relational Database Management System (RDBMS) can deal with the data. With the increasing amounts of unstructured data on various sources (e.g. Web, Social network, and Blog data) that are considered as Big Data, a single computer processor cannot process such huge amount of data. Hence, the RDBMS cannot deal with the unstructured data; a nontraditional database is needed to process the data, which is called NoSQL database. There are different types of social networks data such as user profile data and messages. The former is considered static, while the latter is dynamic.

Data Retrieval

Before retrieving the data, some questions should be addressed: What are the characteristics of the data? Is the data static, such as the profile user information "name, user Id, and bio"; or dynamic such as user's messages, and user's network? Why is the data important? How is the data will be used? And how big the data is? It is important to note that it is easier to track a certain keyword attached to a hashtag rather than a keyword not attached to it. API of social network is a widely used application to retrieve, read and write its data. Other studies, as in [5], have used GNU/GPL application like YourTwapperKeeper tool, which is a web-based application that stores social network data in MySQL tables. However, YourTwapperKeeper in storing and handling large size of data exhibits some limitations in using, as MySQL and spreadsheets databases can only store a limited size of data. Using a hybrid big data technology might address such limitations as we suggested above.

Ranking and Classifying Users (e.g. Twitter)

There are different types of user's networks; a network of users within a specific event (hashtag), a network of users in a specific user's account, and a network of users within a group in the network, that is, Twitter Lists. Lists are used to group sets of users into topical or other categories to better organize and filter incoming tweets [6]. To rank Twitter users, it is important to study the characteristics of Twitter by studying the

network-topology (number of followers/ followed) for each user in the dataset. Many techniques have been employed in ranking analysis.

Homophily

Homophily is defined as the tendency that contacts among similar users occur at a higher rate than among dissimilar users [3], that is, similar users tend to follow each other. It requires studying the static characteristics of social networks data, such as the profile name and the geographic feature of each user in social networks.

Reciprocity

Social networks have made most studies analyze reciprocity. Reciprocity is the property of following a user and being followed back (mutual relationship). For instance, celebrities tend to follow each other, so are politicians, bloggers, and ordinary users. From [3], [6] we can conclude that homophily and reciprocity have the same logical behavior. In [3], the reciprocal relationship is measured by analyzing the number of followers, PageRank, and forwarding amount. Additional methodology is investigated in [6], where the users follower-graph is studied to infer users reciprocities.

Information Diffusion

Since there are different kinds of information spread over social networks, there is no agreement on what kind of information is more widely spread than others. There is also no agreement on how messages are spread over social networks. In this area many studies have attempted to address those questions by studying the First-network topology, and by measuring the forwarding amount [5], [7], [8].

Event Life Cycle

To analyze the life cycle of an event, it is important to choose the measurements of the life cycle such as measuring the number of messages over a period of time, and the number of users in the network. In [5], the life cycle of five different hashtags were demonstrated and analyzed by tracking the most uprising political events. Regarding the difficulty of tracking a specific event for a long period of time, [9] followed an effective technique by tracking a specific hashtag on different times and employed a comparison between them to examine the fluctuation of the event life cycle as they investigated three metrics to track each hashtag.

Network-Topology Analysis

Networks consist of levels of a hierarchal fashion, that is a first-network topology, a second-network topology of the first-network topology, and so on. Most studies have focused on the first-network topology for analyzing information diffusion over social networks. In [3], [5], for instance, studied the first-network topology to examine information spread. In [10], a hybrid methodology had been investigated to analyze the message content, besides analyzing the network-topology, by employing a linear-regression model to predict the speed of message propagation for each crawled hashtag.

Influence on Social Networks (e.g. Twitter)

Social influence occurs when an individual's thoughts or actions are affected by other people [8]. Examining the influential users is related by the message propagation by answering on the following questions; who are the originators of the tweets, how many audiences they have, and what is the retweet rate of the original tweet. Most studies agreed on analyzing the network-topology and the retweet rate to identify the influential users. Additional methodology had been used to examine the influence by studying the retweet mechanism through the "Centrality measures" technique [11].

[11] used the "Degree Centrality" by counting the number of links attached to the node (user) in case of directed graph. Also employed the "Eigenvector Centrality" by answering the question of "how many users retweeted this node?" As [6], [12] agreed on identifying the influential users by ranking the users using the number of followers, the PageRank, and the retweet rate. Additional method had been employed by [9], which is studying the reply influence metric and identifying the number of replies to the original tweet. In addition to analyze the network-topology, the authors in [13] investigated another methodology by analyzing the number of tweets, the date of joining, and the previous history of the influential users.

Sentiment Analysis

Sentiment analysis is the measure of people's opinions on the level of agreement on a specific topic, a product, or a service, or even elections. Two approaches had been employed to study the sentiment analysis: natural language processing, and machine learning algorithms.

To assess the customers' opinions in the past some paper-based surveys had been used, but it is difficult to monitor and collect all customers' opinions. With the increasing phenomena of social network it has become easier and more accessible to crawl all customers' feedbacks and analyze their sentiments as positive or negative.

Natural Language Processing Approach

According to [14], natural language processing (NLP) is the interaction between computers and human (natural) languages. To evaluate sentiment of users online, particularly on social networks, effective sentiment annotation should be used. Most studies use the three common sentiment labels: positive, neutral, and negative. In [15], new feature had been used to effectively annotate sentiments of users; "Mixed Sentiment label", it exists in messages that have two different meanings. For example "I love iPhone, but I hate iPad". "iPhone" entity is annotated with positive sentiment label, and "iPad" entity is annotated with negative sentiment label, that means the message has a mixed sentiment

Machine Learning Approach

According to [16], machine learning (ML) is a scientific discipline that explores the construction and the study of algorithms that can learn from data.[17]-[20] used the

machine learning approach in analyzing the sentiment of social networks users. A hybrid method had been used by [18] since an advanced classifier was employed for sentiment analysis "The Latent Dirichlet Allocation Model", in which a topic has probabilities of generating various words.

Conclusion

The sheer amount and the different types of data on social networks and the public nature of messages have allowed exploiting social network information in data analysis. Firstly, by measuring the life cycle of a specific topic by measuring the number of messages over a period of time. Secondly by measuring the sentiment of users towards a specific topic through NLP and ML algorithms. Our aim is to enhance the analysis of social network data for specific events to measure the effect and the behavior of users towards different events categories. A successive work will focus on studying the data and its attributes, and investigating modeling techniques to identify the frequency distribution for each event.

References

[1] Twitter. From https://en.wikipedia.org/wiki/Twitter

[2] Facebook. From https://en.wikipedia.org/wiki/Facebook

[3] Bastos, M. T., Travitzki, R., & Puschmann, C. (2012). What sticks with whom? Twitter follower- followee networks and news classification. *Proceedings* of *6th International AAAI Conference on Weblogs and Social Media—Workshop on the Potential of Social Media Tools and Data for Journalists in the News Media Industry.*

[4] Hajibagheri, A., & Sukthankar, G. (2014). Political polarization over global warming: Analyzing twitter data on climate change. *Academy of Science and Engineering (ASE), USA.*

[5] Bastos, M. T., Travitzki, R., & Raimundo, R. (2012). Tweeting political dissent: Retweets as pamphlets in #FreeIran, #FreeVenzuela, #Jan25, #SpanishRevolution and #OccupyWallSt. University of Oxford.

[6] Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. *Proceedings of the 20th International Conference on World Wide Web.* ACM New York, NY, USA.

[7] Bongwon, S., Lichan, H., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting Retweet in Twitter network. *Proceedings of the 2010 IEEE Second International Conference on Social Computing* (pp. 177-184).

[8] Ye, S., & Wu, F. (2013). Measuring message propagation and social influence on Twitter.com. *International Journal of Communication Networks and Distributed System, 11(1),* 59-76.

[9] Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology, 16(2)*, 91-108.

[10] Tsur, O., & Rappoport, A. (2012). What's in a Hashtag? Content based prediction of spread of ideas in microblogging communities. *Proceedings of the Fifth ACM international Conference on Web Search and Data Mining* (pp. 643–652). ACM New York, NY, USA.

[11] Kumar, S., Morstatter, F., & Liu, H. (2014). *Twitter Data Analytics*. Springer, New York.

[12] Kwak, H., Lee, C., & Park, H. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web.* Raleigh, North Carolina, USA.

[13] Romero, D. M., Medeer, B., & Kleiberg, J. (2011). Differences in the mechanics of information diffusion topics: Idioms, political Hashtags, and complex contagion on twitter. *Proceedings of the 20th International Conference on World Wide Web* (pp. 695-704).

[14] Natural language processing. From https://en.wikipedia.org/wiki/Natural_language_processing

[15] Saif, H., Fernandaz, M., & Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the STS-Gold. *Proceedings of 1st International Workshop Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI* (ESSEM 2013). Turin, Italy.

[16] Machine learning. From https://en.wikipedia.org/wiki/Machine_learning

[17] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical Report, Stanford Digital Library Technologies Project.

[18] Jahanbakhsh, K., & Moon, Y. (2014). The predictive power of social media: On the predictability of U.S presidential elections using Twitter. *arXiv preprint arXiv: 1407.0622.*

[19] Johnson, C., Shukla, P., & Shukla, S. (2012). On classifying the political sentiment of tweets. *Cs.utexas.edu.*

[20] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. *The Semantic Web* (pp. 508– 524). ISWC.