Detection of Likelihood Topics Trends on Internet

Preeti Sao
Shri Shankaracharya Group of Institutions
Dept. of Computer Science and Engineering
Bhilai, Chhattisgarh, India
preetisao7587@gmail.com

Prof. Sampada Vishwas Massey
Shri Shankaracharya Group of Institutions
Dept. of Computer Science and Engineering
Bhilai, Chhattisgarh, India
sampada.satav@gmail.com

Abstract

Discovery of likelihood topics becomes more interested in the rapid growth of social networking sites. The information exchanged in the social networking site's post involve not only the text contents, but also images, URL's and video hence conventional-term-frequency-based approaches may not be appropriate in this context. Emergence of topics is mainly focused by the social aspects of these networks. In this paper we propose a novel method where the user tweets are analyzed by the probabilistic function. We propose joint probabilistic based probability calculation of user tweets, re-tweets and mentions. The proposed algorithm outperforms the traditional key based method in terms of start and end time calculation of tweets with particular hashtag.

Keywords— *Topic detection, Social Networks, Anomaly detection, Burst Detection.*

I. INTRODUCTION

Nowadays, communication over social networking sites such as Facebook, Twitter, etc. has been increasing its value. The development of social networking sites increases the messages or data information exchanged between the users are not only the text contents, but also images, videos and URLs. Particularly, we are involved in the discovery of likelihood areas from social media streams which are posted by thousands of users. It can be used to capture the unedited voice of normal or ordinary people via social media, which helps to create automated "breaking news" , or underground political movements or discover hidden market needs. Hence compared to conventional media, the social networking sites are able to capture the unedited voice of ordinary people as early as possible. Therefore, the challenge is to discover the emergence of topics earliest possible at a moderate number of false positives.

The major difference that makes social networking sites more social and popular is the existence of mentions. Here, we mean by existence of mentions is to link with the other users of the same social networking sites in the form of reply-to, message-to, retweet-of or explicitly in the form of text. Single post may contain a number of mentions. There are some user may include mentions in their posts rarely; there are other users too which may mention their friends all the time. Some users such as celebrities may receive mentions every minute; and for other being mentioned might be a rare occasion. Therefore, in social media, mention is like a language with the number of words equal to the number of users.

The huge popularity of social networking sites such as blogs, comments and posts represent significant opportunities. A vast volume of information is generated everyday by bloggers and other contents produced over worldwide, therefore providing an important real time view, opinion, feelings, comments, activities, intention and trends of individuals as well as group across the globe. These data may enable early for the detection of likelihood topics, issues and trends for the considerable value of interest. However, it is very difficult for the user to find the signature of likelihood topics and trend which are buried in the massive and largely irrelevant data. Therefore, to discover useful single topic rapidly out of millions of online data generated daily by the social networking sites is extremely difficult.

We are interested to discover likelihood topics from social networking streams which is based on monitoring the mention behavior of different users. Here is our basic assumption is that a new likelihood topic is something in which people like commenting, discussing, or forwarding the useful information further to their different friends. Therefore, the conventional approaches for the topic detection which have mainly been concerned with the frequency of textual words. Frequency of textual words based approach could suffer from the ambiguity caused by the homonyms or

synonyms. The target language may also require complicated preprocessing. Moreover, when the content of messages is non-textual information then it is not applied. On the other hand, the unique word formed by the mentions requires little preprocessing to obtain the data or information is separated from the contents and is available in spite of the nature of the contents.

A Social networking site has some challenges like discovery of topics, theme pattern from text, bursts, change points and outlier detection. To overcome these disputes, there are various methods and different models have been proposed. The challenging task is to find the anomaly.

Anomaly detection deal with finding the patterns present in a data-sets whose behavior is unexpected which means not normal. These types of unexpected behaviors of the patterns in a data-set are also called as anomalies or outliers. Each and every anomaly cannot be always detected or categorized as an attack but it can be categorized as a surprising behavior which is previously not detected or known. That surprising behavior may be or may not be that much harmful.

The anomaly detection gives very significant as well as critical information in a variety of applications, for example identity thefts or the Credit card thefts. When data has to be analyzed with respect to find the relationship or to calculate or predict the known or unknown data mining techniques are used. And therefore, this includes the clustering of data-sets, classification and also machine based learning techniques. In order to attain higher level of accuracy, the Hybrid approaches are also being created on detection of anomalies. This approach help us to try to combine the existing data mining algorithms which helps to generate better results. Therefore detecting the abnormal or the unexpected behavior or anomalies will produce to study and this will categorize it as into a new type of attacks or any particular type of interruption.

II. LITERATURE SURVEY

Xiaomeng Wan [1], the proposed link-based anomaly detection method helps to consider deviations from individual patterns by taking into consideration the behavioral pattern of the cluster to which the individual belongs. Cluster can be formed based on a specific attribute or by a standard clustering procedure depending on the dataset shows the experiment that this method performs well on the data available by the network traffic and the data available by the email communication.

Richard Colbaugh [2],propose paper consider the problem of monitoring the social networking sites to spot emerging memes – different phrases which act as "tracers" for distinct cultural units - as a means of speedy detecting new topics and trends. Author recently developed a method for the prediction by which memes will propagate widely and which will not, and therefore it enables the discovery of significant topics. In these authors demonstrate the efficiency of this approach through case studies involving memes associated with an emerging cyber threat and political memes.

Jiangfeng Chen [3], focuses on the problem of predicting emerging hot topics. To discover the hot topics, previous prediction models usually focus on building the content profile, the social networking sites post in the form of contents, images or videos may neglect. By introducing a combined model using the connection information and content, author define the concept of topic hotness which helps to introduce the algorithm that calculating the hotness with content based hotness and connection based hotness to evaluate hot topics, and finally predict those evaluate hot topics by the hotness evolution model.

Toshimitsu Takahashi [4], user discovers the emerging topics from the social networking sites such as Twitter, Facebook, etc. To exchange the information in the social networking sites post includes not only the text, but also URLs, images and video hence conventional-term i.e., frequency-based approaches may not be appropriate in this context. There are hundreds of users respond in social networking sites post is used to detect the emergence of new topics. The proposed model helps to capture a number of mentions present in per post and the frequency of different user's occur in the mention.

Stephen Bonner [5], focuses on a new method for storing datasets and also querying medical RDF datasets using Hadoop Map / Reduce model. This approach inherently exploits the parallelism found within RDF datasets and queries, which helps to allow to scale up with both datasets and system size. The previous solutions provides the framework which uses highly optimized (SPARQL) joining strategies, the intelligent data caching as well as the use of a super-query to enable the completion of eight different SPARQL lookups, comprising over eighty different joins, in only two MapReduce iterations. Results are providing by

comparing both the Jena and also a previous Hadoop implementation which demonstrating the superior efficiency of the new methodology. The result of new method is shown to be five times faster than Jena and as fast as twice as the previous approach.

III. METHOD USED

In this section, we will discuss about the method used for analyzing the anomaly of twitter users. The algorithm is used here is Join Probabilistic Model. This model is used for training of twitter dataset.

It is sued to capture the mentions, re-tweets probability of users. It is based on probabilistic theory. For each tweets the probability are calculated and based on that the users are identified with same tweets.

The equation of Join probabilistic model is presented below.

$$P(k, V \mid \theta, \{\pi_v\}) = P(k \mid \theta) \prod_{v' \in V} \pi_{v'}$$

IV. METHODOLOGY ADOPTED

`The proposed work flow of anomaly detection framework is presented in fig. fig.1. There are various modules. The modules are described in this section.

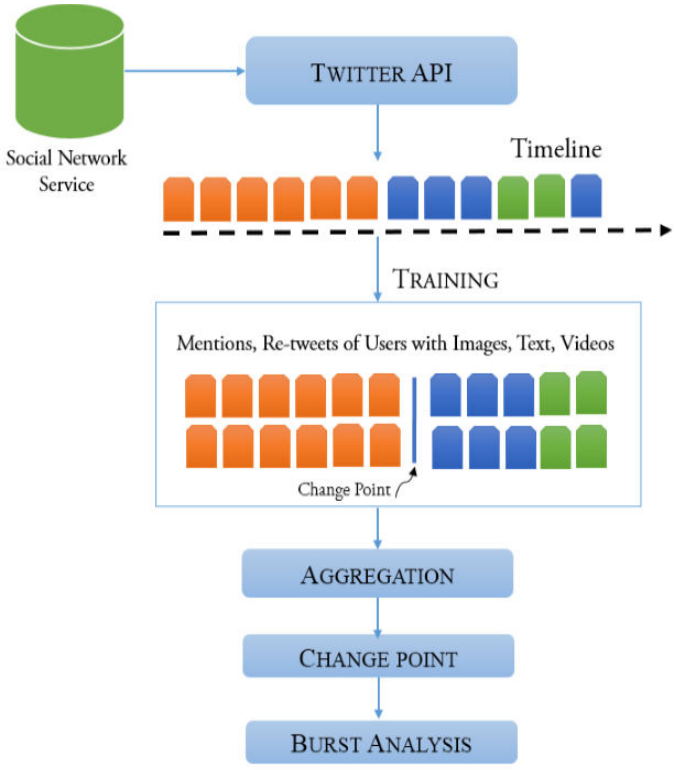

Fig. 1. Workflow of Proposed Framework

A. Social Network Service

It is a service provided by many social network sites. The services like posting, commenting, and chatting are the services offered by many Social networking sites.

B. Twitter API

The twitter API is used for downloading all the tweets from the twitter account. The following information are downloaded.

a. Screen ID
b. Screen Name
c. Tweets
d. Photo
e. Mentions

C. Timeline

Timeline is the collection of the tweets where all the tweets are present. It may contain text, images and videos with particular hashtag.

D. Training

The training part is for finding the users tweets probability. The join probabilistic model is used.
It considers Images, Text, Video files for training of tweets.

E. Aggregation

Aggreagtion module is used for calculation of average time delay for each tweets. The time delay is calculated in this section and outputted to change point module.

F. Change Point

It is used to calculate the start and end time of the tweets. It calculates the time of start of that tweet and calculates the end of the tweets.

G. Burst Analysis

This modules comapres the time taken by key based and link based anomaly detection algorithm. The key based doesnot consider the re-tweets and link based considers all the aspect of tweets such as re-tweets and mentions.

V. RESULT AND DISCUSSION

The experiments are performed in the Eclipse IDE using Java language. The #IndVBan hashtag is considered for analysis. Fig. 2 shows the comparison of key based and link based anomaly algorithms.

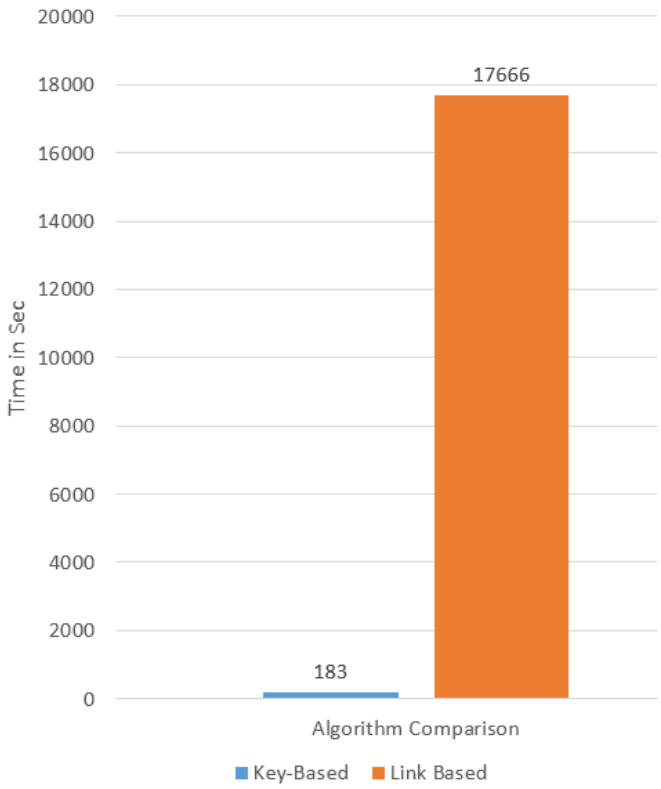

Fig. 2. Comparison of Algorithms in capturing tweets time

In paper [2] proposed by Richard, address the problem of sporting and monitoring memes. The result is presented in fig. 3. For τ = 24 hrs. The amount of accuracy achieved is 92.8 %.

In paper [3] proposed by Jianfeng, address the problem of anomaly in documents. The model proposed by Jianfeng achieved upto accuracy of 91 %.

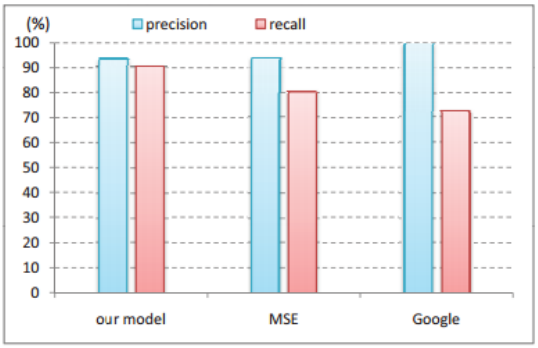

Fig. 3. Shows the performance of [3]

In the proposed approach the accuracy is achieved from the traditional key based approach is more than 95.2% for 24 hr. The ground truth for tweets is 37,000 tweets, out classifier measures it as 35,233 tweets.

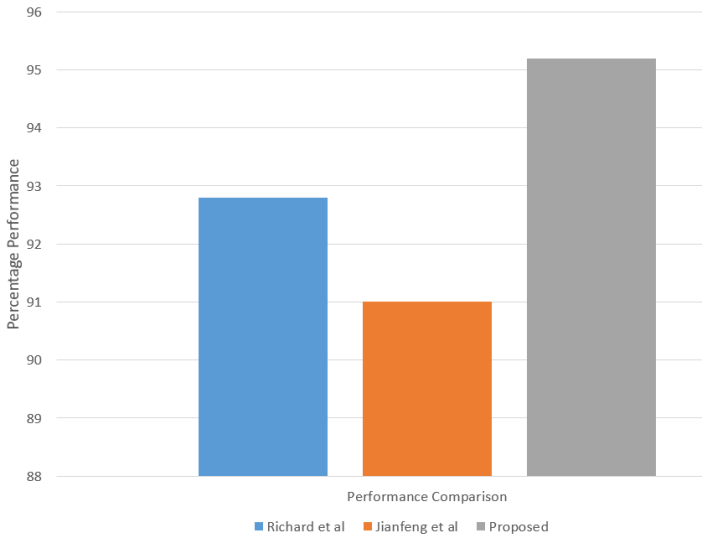

Fig. 4. Shows the performance of various author and proposed one

VI. CONCLUSION

In this paper we compare two main methods i.e. key based and link based. The key based algorithm cannot identify the tweets of the users who have re-tweets or mention some other person into it. The link based algorithm which identifies based on re-tweets and mention in the tweets outperforms the key based algorithm in term of finding start and end time.

REFERENCES

[1] Xiaomeng Wan, Evangelos Milios, Nauzer Kalyaniwalla and Jeannette Janssen, "Link-based Anomaly Detection in Communication Networks", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[2] Richard Colbaugh, Kristin Glass, "Detecting Emerging Topics and Trends Via Predictive Analysis of 'Meme' Dynamics", 2011 IEEE.

[3] Jiangfeng Chen, Jianjun Yu, Yi Shen, "Towards Topic Trend Prediction on a Topic Evolution Model

with Social Connection", 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[4] Toshimitsu Takahashi, Ryota Tomioka, Kenji Yamanishi, "Discovering Emerging Topics in Social Streams via Link Anomaly Detection", 2012 IEEE Transactions on Knowledge and Data Engineering.

[5] Stephen Bonner, Andrew Stephen McGough, Ibad Kureshi , John Brennan , Georgios Theodoropoulos, "Data Quality Assessment and Anomaly Detection Via Map/Reduce and Linked Data: A Case Study in the Medical Domain", 2015 IEEE International Conference on Big Data (Big Data).

[6] C. Budak, D. Agrawal, and A. E. Abbadi, "Structural Trend Analysis for Online Social Networks," Proceedings of the VLDB Endowment, Volume 4, Issue 10, 2011.

[7] Y.-N. Tu and J.-L. Seng, "Indices of Novelty for Emerging Topic Detection," Information Processing and Management, 2011.

[8] D. He and D. S. Parker, "Topic Dynamics: an Alternative Model of Bursts in Streams of Topics," In the Proceedings of KDD'10, 2010.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," The Journal of Machine Learning Research, 2003.

[10] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, Volume 30, Issues 17, 1998, 107117.