Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.697 – 703

RESEARCH ARTICLE

Exploring Hottest Trends at Twitter

Saikiran Chepuri[1], Vaibhav Wadhai[2], Suhaib M Ansari[3]

[1]M.Tech & VIT University, India

[2]M.Tech & VIT University, India

[3]M.Tech & VIT University, India

[1] saikiranchepuri35@gmail.com; [2] vibs.wadhai@gmail.com; [3] es_em31@yahoo.co.in

Abstract*: *Twitter is one of the rising social media platforms which enable users to post and send the short-messages considered as tweets. Detecting hottest trends at twitter via re-tweeting disseminates the information throughout the world. Traditional term-frequency approach considering with fewer regions makes the information to pass through some extent only. The focus of this paper is about extending region by taking into account of the re-tweet count on proposed Linkbased method assuming the increase in the geo-graphical region leads to gain more importance of trends. In this paper, a proposal on Increased Region Space leads in detection of exploring more trends leads to a huge change compared to previous Keybased method. Trends are changing over the time we fetch the twitter trends and then aggregating on individual candidate dataset for the detection on the time intervals between each post at twitter; which shows that the hottest trending topics only based on the reply/mention relationships in social-network posts. The paper determines the technique of gathering several real data sets from Twitter; the experiments show that the proposed Linkbased approach can detect new trending topics at least as earliest as Textbased approach.

Keywords: social network, twitter trends, re-tweet count, region expansion, reply mentions

I. INTRODUCTION

With the rapid advancement of Internet technology the communication over online social network community permit clients to spread data in quicker and faster way. Social Network Analysis (SNA) is a discipline of examining to provide on a set of tools and theoretical approaches for holistic discovery of the communication and relations of social systems. Deeper understanding of SNA leads to detect the communities, friendship relations and the mutual relationships between the users connected at one network say, social platform [4].

Twitter is the current micro blogging and person to person communication meting out to present social platform those open doors for looking at regular language behavior (i.e., the NLP) and machine learning [2]. Since the data exchanged over the organizations of social networks is messages as well as URLs, pictures, and features, they are tested on analysis for the study of data mining. In particular, we are keen interested in the problem of detecting (hottest) trending topics from various places of world via social streams.

Twitter displays pretty opportunities for NLP applications and machine learning. One such part of Twitter that gives opportunities-is the drifting points - words and expressions, highlighted on the fundamental page of Twitter, that are at present prevalent in clients' tweets. Trending topics are recognized as for as long an

hours, day and week [1]. The development of Twitter exhibits a few difficulties for uses of NLP and machine adapting. Communication over social networks like Face book, Twitter, YouTube, Flicker, and Oracle gets more and more consideration consistently. Each and every day, the attention of social network data is turning into a more specific subject in the field of data mining [4]. The analysis of Twitter's network is that in order to discover and identify trending topics is that one must understand and have to estimate its huge quantity at online.

II. MOTIVATION

The main aim of this paper is to spread the information as far as possible. An assumption made to the traditional method is to limit a boundary value (which to be considered as less region initially) then on proposed Linkbased method we try to increase the scalability of the system as it demonstrates to more region when compared to earlier one, the Keybased approach. As the following figure shows the re-tweeting enables the fast spreading of information, this is possible through the user reply/mention of messages in the form of comments, re-tweets, likes , adding favorites, tagging the users with hottest trending topic etc.

Fig 1: Dissemination of information through re-tweets

Eg: If A updates his status as "Fabulous India at #WC2015 ", his friend B comments the A post as "great by #Virat at #WC2015 " so the same post to be followed to B friend's who is not in touch with A. However the some people be using #WC2015 and can re-tweet to previous post and it continues to grow at network which finally becomes the most trending topic at twitter.

III. LITERATURE SURVEY

Although there is a significant work pertaining to natural language processing (NLP) techniques. While the twitter have been used to examine the dynamics of social networks, mainly on the temporal behavior of social networks at some stage like in disasters, such as earthquakes and hurricanes etc...

In this paper, an analysis of prior work done on detection of streaming trend topics is done through the tf-idf term weighting, and term frequency analysis to get the trending topics are identified as spikes in the form of both unigrams and bigrams as discussed in[1]. The discussion is gone on implementing NLP through a step refinement of processing stop-lists and stemming algorithm to generate tweets.

Another work related to rumor spreading at online social network which was taken the data set from Sina micro-blogging [2]. This study shows how the topic is spreader out due to increase in the size of community.

Another study shows on how important people responsible for the growth of community by spreading the information through re-tweeting i.e., through the reply/mention of user tweets in the networks, this work done by Kretschmer as reviewed in [3].

A work related to ontology representation for expanding the network by applying clustering algorithm regarding the trends and clusters of network which is discussed in [4].

IV. PROPOSED WORK

In this paper, the focus is on expanding the network so that each and every corner of the global area can identify the trend which detects the hottest topic has been reached to some extent. For this work, 4 modules are considered along with the existing and proposed method. It has of: Social network dataset preparation, Training on candidate datasets, Aggregating on individual topic, and Breakpoint analysis. A total count has been well thought-out to measure the number of tweets have been growing on dynamically. For Keybased method, a threshold value to be included so that it specifies to limited tweets in limited region only. In this we go through the Linkbased method which is measured through the more reply mentions of the user by re-tweeting as by forwarding the discussion to other users at network.

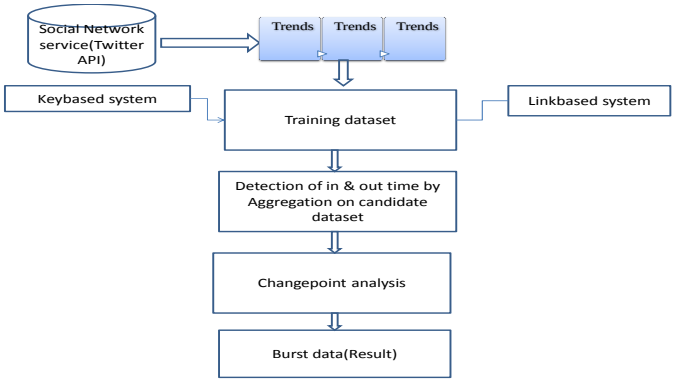


Fig 2. Proposed architecture

V. ALGORITHM

Proposed method on analysis of re-tweet count
Steps involved:
Input : Social Network dataset
Output: Variation between keybased and linkbased method in terms of total count
Steps:
1. Social Network dataset preparation
2. Training on twitter trends
3. Aggregate on user dataset for time interval detection
4. Change point Analysis:
Total count: from
   A. Keybased method
   B. Linkbased method in retweets/reply.
5. Result estimation through Bar chart.

A. Social Network Dataset preparation

This module prepares up extracting the twitter trends containing #hash taglines which are created from 4-5 days from Internet connecting to twitter API through OAuth protocol mechanism. It is not possible to extract the tweets older than 6 days but in few cases retrieving tweets is considered for its higher participation at twitter.We characterize a post in a social network stream by the number of mentions k it contains, and the set V of names (IDs) of the mentionees (users who are mentioned in the post). There are two types of infinity we have to take into account here. The first is about the number k of users mentioned in a post. Although, in practice a user cannot mention hundreds of other users in a post, we would like to avoid putting an artificial limit on the number of users mentioned in a post. Instead, we will assume a geometric distribution and integrate out the parameter to avoid even an implicit limitation through the parameter. The second type is extending the behaviour of post thus by increasing scalability of the system.

B. Training and Aggregation on Candidate dataset

In this subsection, we eliminate stop-words, punctuations and website address (usually starts with “http(s)”) and numbers, symbols except “#” and letting the trends be identified with the start of a symbol of “#”. So, only user id, user name and screen name (their tagged line at twitter e.g.: @SaikiranChepuri).Now, it shows how many numbers of persons are tweeting on a particular topic, and thus we come to know how much the conversation between/among the persons is being going in particular interval of time which can be done with the help of next section of performing aggregation on selected candidate dataset.

On aggregating for selected candidate dataset we detect the time interval for each post of candidate on particular twitter trend. A re-tweet count is being considered for the Linkbased method which is nothing but defines to “spreading of information”. Eg: If someone posted/commenting on a particular #hash

tagline, his friends is also able to re-tweet that and further can forward to the network. So, a re-tweet count is created dynamically for Linkbased approach.

C. Breakpoint analysis and Burst data

In this section, it shows a breakpoint for two methods i.e., Keybased and Linkbased. For Keybased an artificial limit has been restricted to have certain count value is considered while for Linkbased method, we extend the scalability of the network as we are considering a re-tweet count which the total counts of initial Keybased approach will be a high value.

The result between the two methods is demonstrated in the form of Bar chart with the view of comparison to the total count in the two methods. Time and count considered as an X-Y dimension of bar chart gets to know how the topic has been spreading even at remote areas of world by expanding the geometrical region/space.

EVALUATION AND EXPERIMENTAL RESULTS

Following figure shows the trending topics at twitter, from which the #HappyBirthdaySachin in India is getting more spikes which conducted the experiment on April 24th 2015

Fig 3. Fetched trends from Twitter at Internet

We used OAuth protocol which used to get twitter trends and designed a GUI with the help of Java swings the following panel showing the twitter trends along with corresponding URI, In & Out time and Re-tweet count columns.

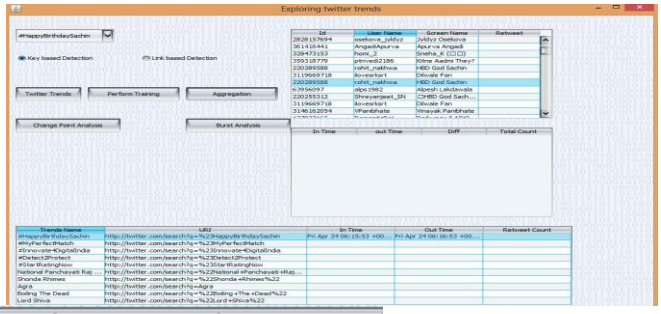
Trends Name	URI	In Time	Out Time	Retweet Count
#HappyBirthdaySachin	http://twitter.com/search?q=%23HappyBirthdaySachin			
#MyPerfectMatch	http://twitter.com/search?q=%23MyPerfectMatch			
#StartRatingNow	http://twitter.com/search?q=%23StartRatingNow			
National Panchayati Raj ...	http://twitter.com/search?q=%22National+Panchayati+Raj...			
#GreysAnatomy	http://twitter.com/search?q=%23GreysAnatomy			
#BombayVelvetMusic	http://twitter.com/search?q=%23BombayVelvetMusic			
Boiling The Dead	http://twitter.com/search?q=%22Boiling+The+Dead%22			
Subramanian Swamy	http://twitter.com/search?q=%22Subramanian+Swamy%22			
Happy Wedding Anniver...	http://twitter.com/search?q=%22Happy+Wedding+Anniver...			
Agra	http://twitter.com/search?q=Agra			

Here am giving a detailed step parallel process of trending topic along with real-time trends at internet and manual made GUI. On performing training at candidate datasets, one candidate say @rohit_nakhwa has updated his status with trending topic as shown in below fig.

Here the screen name representing the group at twitter.

Id	User Name	Screen Name	Retweet
2828157694	osekova_jyldyz	Jyldyz Osekova	
361416441	AngadiApurva	Apurva Angadi	
328473153	homi_2	Sneha_K (□□)	
359318779	ptrivedi2186	Kitne Aadmi They?	
220289588	rohit_nakhwa	HBD God Sachin	
3119669718	ilovesrksrt	Dilwale Fan	
220289588	rohit_nakhwa	HBD God Sachin	
63956097	alps1982	Alpesh Lakdawala	
220255312	Shreyanjeet_SN	□HBD God Sach...	
3119669718	ilovesrksrt	Dilwale Fan	
3146162054	VPanibhate	Vinayak Panibhate	

On performing keybased method (below at radio button selection) by Aggregating the @rohit_nakhwa user it detects the time interval and total count of 100 which is shown in fig 1 and fig 2



In Time	out Time	Diff	Total Count
1398300353527	1398300413527	60000	100

While repeating the same procedure for linkbased method(proposed one) we can observe the Retweet count for our user as shown in below fig.

Linkbased

Id	User Name	Screen Name	Retweet
2834905028	iEkansh29	Ekansh Kapoor	6
2374946683	bhuvansingh14	bhuvan singh	1
1212105139	iamrvsc	(^o^)	4
1687924130	rotahuabacha	Dr. Zidhu	0
124473367	bhanu_tarak9999	HappyBirthdayS...	0
148655912	TIKKI16	TIKKI	0
1356050886	saliseemala	Charan_Ntr	5
44574946	sourabhmalhotra	TheGoodCounsel	24
220255312	Shreyanjeet_SN	HBD God Sach...	4
285549010	sandyrockyd	Santhosh R	26
301974922	mesurendernegi	Surender Negi	0

The below figs shows the proof of demonstration which is collected from connecting to Internet.

After performing Aggregation process which shows the change in total count within the interval of time period as shown in below fig.
Difference in total count:

In Time	out Time	Diff	Total Count
1398300353527	1398300413527	60000	100
1398301018273	1398301101273	83000	7856

A bar chart progressed about the variation of Time to Total count of two methods. The first one specifying about the less number of count value spending lot of time while the second one (green colored) demonstrates the more number in count value within less time period.

VI. RESULT ANALYSIS

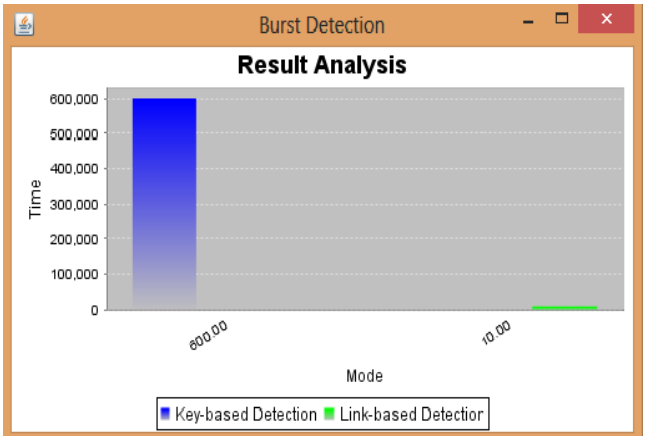


Fig 4: A sample bar chart demonstrating two methods

Since, all this is made under the assumption of disseminating the information at each corner of the world through re-tweet count. However, if we go through the Outlier detection methodology at Data mining area will get accurate results.

At glance of internet if we increase the space region across from remote area to metros of city(say from Australia region to US/Green Land region) we let to know how much the topic is been getting importance around the world.

VII. CONCLUSION AND FUTUREWORK

In this paper, discussions made on the importance of twitter trends which are trending over the time. A brief introduction on social network analysis is made initially at the Literature Survey which demonstrated on how communities are formed and about the user relationships between nodes at social network graph. Then the topic went on Keybased and Linkbased approaches performed on individual candidate datasets which results in a variation on total count of two methods. Specifically the re-tweet resulted to gain more count which specifying that increasing the geometrical region can explore the importance of trend. In the future, the work can be done by classifying the network region to identify how the topic is praised/ rumor on a particular trend.

ACKNOWLEDGEMENT

We are very much thankful to our guide Prof.Prabhavathy.P who is currently working as Assistant Professor at VIT University, SITE department for timely and exclusive access to gain information on various Data mining and Data processing issues.

REFERENCES

[1] James Benhardus, "Streaming Trend Detection in Twitter," UCCS REU for Artificial Intelligence, Natural language processing and Information Retrieval,2010.

[2] LIU Ling-han, QI Jia-yin, YU Chao-yang , " How Does Praise and Rumor Spread Online? Analyzing User Relationships on Social Networks: A Case Study of Haidilao," International Conference on Management Science & Engineering,vol. 13,2013.

[3] Zudha Aulia Rachman, Warih Maharani, Adiwijaya ," The Analysis and Implementation of Degree Centrality in Weighted Graph in Social Network Analysis" International Conference of Information and Communication Technology, Vol. 13,,2013.

[4] Fakhri Hasanzadeh, Mehrdad Jalali, Majid Vafaei Jahan " Detecting Communities in Social Networks by Techniques of Clustering and Analysis of Communications" Institute of Electrical and Electronics Engineers , Vol. 14, 2014.

[5] Shokoufeh salem minab, Mehrdad jalali, "Online Analyzing of Texts in Social Network of Twitter" International Congress on Technology, Communication and Knowledge and IEEE, Vol.14,ppno. 26-27,2014 .

[6] Rocío Abascal-Mena, Rose Lema, Florence Sèdes, "From Tweet to Graph: Social Network Analysis for Semantic Information Extraction" Institute of Electrical and Electronics Engineers, Vol.14,2014.

[7] Xufeng Shang, Yubo Yuan, "Social Network Analysis in Multiple Social Networks Data for Criminal Group Discovery," International Conference on Cyber-Enabled Distributed Computing and Knowledge Discove,2012.

