International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

Identification of Current Events and Control Spamming from Social Networking Sites- A Review

Roshani M. Shete
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Wardha, India

Prof. Sudhir W. Mohod
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Wardha, India

***Abstract* - Nowadays all kinds of real world events generate reach and timely information can generates from online social media. However large amount of user generated data is available which requires various ways for filtering to get related events and topics. Therefore, detecting trending topics is perfect to summarize information getting from social media. But unfortunately social services are opportunity for spamming which greatly affect on value of real time search so control spamming is necessary. There are different several ways, techniques, methods which affects on result quality. It is observed that how preprocessing of data, considered event nature, sampling procedure of data and volume of activity by considering time affect result quality of detected topic. It also depends on which methods being used. So by considering all above, in proposed work different data mining concepts like natural language process, machine learning, support vector data regression and topic K-means will be used.**

Keywords - Event detection, control spamming, social media, text mining, information filtering

I. INTRODUCTION

Popularity of social media is increasing day by day. The pervasiveness is expanding of online media. The importance of information is increasing dramatically, as communications and interactions using chats, texts always reflect on dynamics and real time events. Public base on online media gets more active in producing stuff on real world events. Because social networking sites content become exact and accurate sensor of real time events.

More than 150 million people are using social networking sites to be connected to their co-worker, friends and family members. There are so many subjects, events discussed by people on it.

Few events get more attention whether some get less. So, there is need to invent a system that tells us which topic is hot on the social media and why.

Social media data mining is nothing but process which understands and predicts meaning of online user generated words. It is study of dynamics of real world. As the Arab riot [2] the disasters of nature [3], opinion tasks of political process [4] are example of events which are reported by using content of social media. To describing applications like urban monitoring [5], computational journalism [6], [7] implicate by the capacity to monitor such type of phenomena. This public generated data streams can help detecting and preventing the misuses and illegal activity on online media [8], [9].

So for producing this kind of dimensions in real time there is need of collection of user generated data on social networking sites. There are two approaches in the proposed work, identifying current and control spamming.

Identifying the topic which is currently public is discussing by posting, commenting, and texting on social media. This kind of tasks [10] tackled before but there are many difficulties aroused like spamming, noise of content, time resolution, burstiness of events and fragmentation etc.

The Blogosphere has unfortunately been infected by several varieties of spam like content. Spam means unwanted behavior of spammer i.e. irrelevant comment or post, posting useless crap, unwanted data and ads, repetition of comment. It affect on search machine, page ranking. So for removing this type of behavior there is need of controlling spamming.

The methods which will be used in proposed work make sure to avoid all above problems. So, natural language processing will be used for preprocessing, then machine learning for predicting and suggesting term which is previously appeared, also techniques like support vector data regression and topic K-means will be used for feature extraction, topic detection, topic ranking, frequent terms, time relations etc. Data will be generated from news portals and facebook, twitter etc. social networking sites.

II. RELATED WORK

Topic Detection and Tracking extract event from public generated data on social sources and identify the trend in term of time [10]. In this the public generated data means posts uploaded by them.

In clustering of all the data streams on social media there are two type of methods one cluster by document or cluster by feature. The first is document pivot and the other is feature pivot method. Both approaches are presented by authors differently in previous work. So these two used methods and their work explained below.

Feature pivot method means the term or keyword will be considered while clustering but the drawback of that is it capture misleading term. For example if we want to search “definition of class”, it will shows extra result like

subclass, superclass etc. So the accuracy of result is very less, also redundancy and ambiguity gets formed in this.

M. Cataldi, C. Schifanella and L. Di Caro [11] proposed two measures, term frequency to calculate nutrition for each word and a page rank measure. After that Bursty keywords are obtained using nutrition trend. Then by using graph based approach for bursty keywords generates the topic boundary. Sayyadi, Maykov and Hurst [12] used graph approach in which clustering of keywords is done by matching pairs. They used community detection algorithm in which made a graph whose nodes are clustered. Also the topic extraction is carried out by identifying document with similar term. Lehmann, Kleinberg and Backstorm [13] have used the the graph for short phrases. Phrases are connected by edges.

One of the method modeled called Latent Dirichlet Allocation (LDA) [14], the idea of knowing the most breaking news by calculating the bursty terms in document [15]. This avoids the other topics by capturing the high peak [16]. So first find bursty term then cluster them for event detection. In some graph based approach, the first step is to tag the terms, then group it and then find the interest in social media [17]. Chang and Lim [18] proposed the Discrete Fourier Transform in which the term separated on the basis of periodicity and power. Then, depending on term according to time the cluster get formed.

Document pivot method has problem of cluster fragmentation. At the time of context streaming it depends on the arbitrary thresholds while including new document.

Phuvipadawat and Murata have presented method of detecting breaking news by clustering the document on the bases of similar metric [19]. Petrovic, Lavrenko and Osborne discussed First Story Detection. Aim is to find first document the public is discussing from huge corpus. The new topic is detected by calculating low similarity in clusters. Also the concept of Hashing is proposed. In that the nearest neighbors retrieved from document. But it form difficulty when nearest neighbors are not same [20].

Using hashtags the post, text or you can say tweets retrieved are processed and weighted with tf-idf (term frequency-inverse term frequency) sort countries name, stars. Tweets are then combined into incoming similar content and clusters which are already exist. The same approach is explained in literature [21], [22]. Becker, Gravano and Naaman [22] invented the classification method. Tweets are classified according to the features like social, temporal and topical features. From that the real world events are identified. In this, the real time event and non-event messages are analyzed. The drawback of that is test samples and manual annotation of training.

The other dimension than text for this purpose is temporal relations between tweets. Samet, Shankarnarayanan and Teitler [23] described extraction from noise, concept to deal with several qualities. In this the centroid of cluster is calculated by tf-idf and the average post time. They discussed and studied about tweets, retweets, follower, and friend also all tweeter services. The problem of this method is noise sensitivity and fragmentation. Gabriel Pui Cheong Fung and Hongjun Lu [24] proposed the new system which finds the bursty event from bursty documents. It determined the bursty features occur in time window. The tasks have done in this are finding positive features, enlarge it, find negative features and classify the text.

III. PROPOSED WORK

In proposed work, address the task of detecting topics in (near) real time from social media streams. To keep our approach general, consider the stream is made of (short) pieces of text generated by social media users (posts, messages, or tweets in the case of social media). The flow of proposed work will be data processing, topic allocation, topic detection, frequent pattern, topic trend. In proposed work, the keyword based mining will be used for identifying current event and rule based mining will be used in control spamming. There will be a threshold value. Then after probability calculation, if probability one of repeated word is near to threshold then it is nothing but the current event. Also the words which are not important in sentences will be removed.

In proposed work, the concept of Machine Learning and Natural Language Processing will be used. Machine Learning is the system which learns from past experiences and improves the performances of intelligent programs. So, in proposed system the frequent appearing words will be copied in the database and next time that words will be used. The Natural Language Processing is devoted to make computer "understand" statement written in human language. In Natural Language Processing there are different methods like Lexical analysis which divide text into paragraph, sentence and words. Then, Syntactic analysis in which analysis of word in a sentence to know grammatical structure and semantic analysis in which it derives an absolute meaning of context.

First from given data cluster will be formed, and then natural language processing will be done for classification. The flow of work is explained below.

A. Data Processing

In data processing filtering of all data will be done. In that the punctuation, symbols, deletion of email ids etc. which is not important in current event detection will be removed. After removing unwanted material next process will be carried out.

B. Control Spamming

First by using lexical analyzer, the words will be clustered then that words will be compared with the collection of words in database if they match with it then that words will be blocked. Spammer clustering will be done by using support vector data regression.

C. Topic Allocation

Topic allocation means allocating data in the form of field like we do in our PC; we allocate movies as per the category e.g. Hollywood movies, Bollywood movies, Animated movies etc. So like this there is need of allocating topics category wise. This work will be done in proposed work.

D. Topic Detection

Topic detection will be done after topic allocation for that topic K-means will be used. Topic K-means will use for feature extraction.

E. Frequent Pattern

In this, the words or phrase which is appearing constantly or you can say the word which is having more frequency will be detected. And then the task of ranking will be done.

F. Topic Trend

This is nothing but our main objective in which the detection of current event will be done. It means after all processing the trending topic will be detected.

IV. CONCLUSION

Identifying the current event is complex procedure to deal with the dimensions that extract the story on social networking sites. The main aspects which will be considered in this work are text generated by user, time relations and event nature. The identification of current events from user generated textual content and controlling spamming by using different data mining techniques is the main objective of proposed work. Using these techniques, the aim is to overcome previous drawbacks in proposed work.

REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker and Ioannis Kompatsiaris, "Sensing Trending Topics in Twitter", IEEE Transactions on Multimedia, Vol.15, No.6, October 2013.

[2] A. Panisson, "Visualization of Egyptian Revolution on Twitter", Feb.2011.

[3] T. Sakali, M. Okazaki and Y. Mastuo, "Earthquake shakes twitter users: real time event detection by social sensors", Int. Conf. 2010.

[4] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini and F. Menczer, "Predicting the political alignment of twitter users", in Proc. SocialCom: 3rd IEEE Int. Conf. Oct. 2011.

[5] S. Cohen, J. T. Hamilton and F. Turner, "Computational Journalism", commun. ACM, Vol. 54, pp. 66-71, Oct. 2011.

[6] New York Times Cascade Project-nytlabs.com/projects/cascade.html, 2011.

[7] D. Querecia, J. Ellis, L. Capra and J. Crowcroft, "Tracking "Gross community happiness" from tweets", in Proc. ACM Conf. 2012.

[8] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in Proc. ICSWM: 5th Int. AAAI Conf. Weblogs and Social Media, 2011.

[9] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: When bots socialize for fame and money," in Proc. ACSAC: 27th Annual Computer Security Applications Conf., New York, NY, USA, 2011, pp. 93–102, ACM.

[10] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are strange when you're a stranger: Impact and influence of bots on social networks," in Proc. ICWSM: 6th AAAI Int. Conf. Weblogs and Social Media. AAAI, 2012, pp. 10–17.

[11] M. Cataldi, L. Di Caro and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in Proc. MDMKDD: 10th Int. Workshop Multimedia Data Mining, New York, NY, USA, 2010, pp. 4:1–4:10, ACM.

[12] Sayyadi, M. Hurst and A.Maykov, "Event detection and tracking in social streams," in ICWSM, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds. Palo Alto, CA, USA: AAAI Press, 2009.

[13] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proc. KDD: 15th ACM Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, 2009, pp. 497–506.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003

[15] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and persistence: Modeling the shape of microblog conversations," in Proc. CSCW: ACM Conf. Computer Supported Cooperative Work, New York, NY, USA, 2011, pp. 355–358

[16] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in Proc. WSDM: 4th ACM Int. Conf. Web Search and Data Mining, New York, NY, USA, 2011, pp. 177–186.

[17] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali, "A graph-based clustering scheme for identifying related tags in folksonomies," in Proc. DaWaK: 12th Int. Conf. Data Warehousing and Knowledge Discovery. Berlin, Germany: Springer-Verlag, 2010, pp. 65–76.

[18] Q. He, K. Chang, and E.-P. Lim, "Analyzing feature trajectories for event detection," in Proc. SIGIR: 30th Annual Int. ACMConf. Research and Development in Information Retrieval, NewYork, NY, USA, 2007, pp. 207–214.

[19] S. Phuvipadawat and T.Murata, "Breaking news detection and tracking in Twitter," in Proc.Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM Int. Conf., 2010, vol. 3, pp. 120–123.

[20] S. Petrović, M. Osborne and V. Lavrenko, "Streaming first story detection with application to Twitter," in Proc. HLT: Annual Conf. North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 181–189.

[21] B. O'Connor, M. Krieger, and D. Ahn, "TweetMotif: Exploratory search and topic summarization for Twitter," in ICWSM, W. W. Cohen, S. Gosling, W. W. Cohen, and S. Gosling, Eds. Palo Alto, CA, USA: AAAI Press, 2010.

[22] H. Becker, M. Naaman and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," in Proc. ICWSM: 5th Int. AAAI Conf. Weblogs and Social Media, 2011.

[23] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman and J. Sperling, "Twitter stand: News in tweets," in Proc. GIS: 17th ACM Int. Conf. Advances in Geographic Information Systems, New York, NY, USA, 2009, pp. 42–51.

[24] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in Proc. VLDB: 31st Int. Conf. Very Large Data Bases, 2005, pp. 181–192, VLDB Endowment.

[25] M. McCord and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers", CSE Dept Lehigh University 19 Memorial Drive West Bethlehem, PA 18015, USA.

[26] Fabr´ıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virg´ılio Almeida, "Detecting Spammers on Twitter", Computer Science Department, Universidade Federal de Minas Gerais Belo Horizonte, Brazil.

[27] RC Chakraborty, "Natural Language Processing", AI course lecture 41, notes, June 1, 2010.

[28] Pedro Domingos, "A Few Useful Things to Know about Machine Learning", Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, U.S.A.