A Survey on Uncovering Trending Stories from Twitter by Extracting Ground Truth from Datasets

Shinde Vishal Dilip
Student
Department of Computer Engineering
DGOI COE, Bhigwan, Pune
vishalshinde.it@gmail.com

Dhaigude Tanaji A.
Assistant Professor
Department of Computer Engineering
DGOI COE Bhigwan, Pune
tanajidhaigude@gmail.com

***Abstract*—**Today's online social networking services generates series of conversation that shows the all kinds of real-world events, however the large amount of data are available on social network. This data can be filtered for finding trending topics using standard natural language processing techniques. An Uncovering trending stories is therefore a building block is to extract and summarizes the information raised from social networking services, this building block is very useful to find trending stories and its initiator .There are verity of methods that improves quality of result.

This paper explores about different Topic detection method for uncovering trending topics from twitter datasets, such as Document-Pivot methods, Feature-Pivot methods, Frequent Pattern Mining, Soft Frequent Pattern Mining and BNgram.

Keywords*-Trending Topic, API-Aplicaion Programming Interface*

*****

I. INTRODUCTION

The online social media has seen rapidly expanded in recent years .As social networking services spread rapidly in more geographically area of world. The huge amount of facts discussion, user interaction and communication happen on social networking services and this reflects real world events and trending topics; user is more active for producing content about real-world events, so social media become accurate area where we can drill down for real-time stories

The social networking service has becomes mainstream, as number of user increasing rapidly and there effective conversation. Twitter is openly access for posting information on breaking news and ongoing events, near about 500 million users and more than 400 million short messages known as tweets [13], These tweets contains conversation, thoughts, impacts of current affairs.

Uncovering trending stories from dataset that contains data stream of user tweets with time stamps is helpful for breaking stories broadcaster, social issues analyzer. [13]And it is very useful for government in tracking issues in certain events for security issues when uncovered stories comes with initiator user

In this work, we review different methods for extracting trending topics from Twitter datasets and need of work.

The reminder of this paper is as follows. Next, we provide background on the use of Twitter. Then we introduce types of trends, datasets related work and detailed view of methods for uncovering trending stories.

II. BACKGROUND

In this section we provide background of twitter related to this work, describing the way they interact with each other. Then details of trends in twitter which we are using input to methods for uncovering ground truth from trending stories

- Twitter:

  Twitter is popular social networking service where large numbers of users contribute tweets on daily. The success of twitter is due to two reasons first, shortness of tweets, which cannot exceeds 140 characters that create and share minimum period of time .and second is spreading those messages to a large number of user in very little time. The twitter has established syntax for interaction with one another, which syntax adopted by developers .Most major Twitter clients have implemented this as well. The standard in the interaction syntax include [13]:

  - *User mentions*: when a user mentions another user in their tweet, an at-sign is placed before the corresponding username.
  - *Replies*: when a user wants to direct to another user, or reply to an earlier tweet, they place the *@username* mention at the beginning the tweet.
  - *Retweet*: a retweet is a re-share of a tweet posted by another user.retweets,the new tweet copies the original one in it, then the retweet attaches a *RT* and the @username of the. user who posted the original tweet at the beginning of the retweet.

- *Hashtags*: it is same as tagging facilities on other social networking service, hashtags included in a tweet to mention other user.

- Trending Stories:

  One of the main feature on the homepage of twitter shows a list of top terms so-called trending topics at all times. These terms reflect that are being discussed most. Twitter focuses on topics that are being discussed much more than usual. Trending topics have attracted big interest not for only user mainly for other information consumers such as journalists, real-time application and social media researcher [13].However, no further evidence is know about the algorithm that extracts trending topics[13].

III. TYPES OF TRENDING TOICS

Next, we treat trending topics in following categories [13]:

- News: On many occasions news break on Twitter before any news agency. We define that a trending topic can be categorized as news when that gives present information.
- Ongoing events: The trending topic is in ongoing event when information is posted by community of users tweeting about an ongoing event.
- Memes: Also trending topic is in memes which is posted by either individual or community with viral ideas. It can be from a funny message that attract user to repost.
- Commemoratives: Last type of trending topic which produced by individual for congratulating celebrity their birthday or anniversary r any memorable day such as Independence Day, Republic Day.

IV. DATASETS

Twitter employs search API for the latest tweets containing the topic as a query term and some other information text, timestamp, user and language for each tweet. [13]

V. RELEATED WORK

This section describes about various technique in uncovering trending topics. In Sensing Trending Topic in twitter [1], Three Twitter datasets are used to extract trending topic detection and it is extracted by BNgram Method. In Emerging Topic detection on Twitter based on social terms evaluation [2], recognize the primary role of twitter and they propose a novel topic detection technique that permits to retrieve in real-time the most emergent topic expressed by the communities of users, They define a directed graph of active authors based on their authority by relying on the well-known page algorithm [2]. In another work TwitterMonitor: Trend Detection over the Twitter Stream [3], they represent TwitterMonitor, a system that provides meaningful analytics that synthesize an accurate description of each topic using Twitter API, Another work using Twitter API is TwitterStand: News in Tweets[4] to build a news processing system. In Detecting and tracking political Abuse in Social media [5], describe a machine learning framework that combines topological, content-based and crowd sourced features using Twitter API.

In Predicting political preference of Twitter users [6], they can predict from their interaction with political parties by building prediction model based on a verity of contextual and behavioral feature training the models by restoring to a distance supervision approach. In another work beyond trending topics: Real-world event identification on twitter [7], explores approach the stream of twitter message to distinguish between message about real world events and non-event messages sing cluster-level event features based on Temporal, social, topical, twitter-centric. In next approach, Taking Topic Detection from Evaluation to Practice [8], avoids generating garbage clusters, they had revert to different approach. In mining Newsworthy Topics from Social media [9], demonstrate by analyzing tweets corresponding to events drawn from the word of politics and sport using BNgram method. Also In breaking news detection and Tracking in twitter [10], propose a method to collect, group, rank and track breaking news in twitter, each group is ranked based on popularity and reliability factors.

In recent work, Real-Time Classification of Twitter Trends [13] uses Twitter API for first obtaining top ten trending topic and second obtaining trending topic with text, timestamp, user, and language for each of the underlying tweets.

VI. METHODS FOR UNCOVERING TRENDING STORIES

A. *Latent Dirichlet Allocation(LDA)*

Topic extraction in textual corpora can be addressed through probabilistic topic models [1]. In this method, every document is considered as bag of words or terms. The topic distribution per document and the term distribution per topic are instead hidden and have to be estimated through Bayesian inference algorithm [1].LDA [14] is best known and mostly used topic model so we are using it as baseline to compare our method against.

B. *Document-Pivote Topic Detection(Doc-p)*

The second method we reviewed, the class of his method based on the work by Petorovic [11], which uses LSH for immediately retrieve the nearest neighbor of document. The principle behind this method is that same as in near-duplicate detection in the similarity-based aggregation.

It works as , First it perform online clustering of post by computing cosine similarity of the *tf-idf* [15] if similarity to the best matching post is above some threshold $\Theta_{tf\text{-}idf}$, Assign term to the same cluster as its exact match, if not then create new cluster of new term. In next step filter out cluster with

term count smaller than Θ , and for each cluster, compute a score as ,

$$score_c = \sum_{i=1}^{|Docs_c|} \sum_{j=1}^{|words_i|} \exp(-p(w_{ij}))$$

where w_{ij} is the j^{th} term appearing in the i^{th} document the cluster. In last step clusters are sorted according to their score and the cluster are returns.

The advantage of using LSH is that it can immediately provide the nearest neighbors with respect to cosine similarity in large collection of document [1].

C. *Graph-Based Feature-Pivot Topic Detection(GrFe-p)*

This method uses clustering on features with the Structural Clustering Algorithm for Networks (SCAN) [16].The SCAN graph-based clustering algorithm steps as follows. First is Selection in that top *k* terms are selected using the ratio of likehoods and a node for each of them is created in the graph G. in next step linking the nodes of G are connected using a term linking strategy first similarity measure for pair of terms is selected, in clustering the SCAN algorithm is applied to the graph for detecting each communities.

D. *Frequent Pattern Mining(FPM)*

This method mainly consist two methods first Frequent Pattern Detection (FPD) and second one is Frequent Pattern Ranking (FPR), In first step FPD requires three rounds of Map-Reduce processing, keyword list, Parallel construction of an FP-tree data structure and frequent pattern mining .In next processing step of FPR, once a set of frequent pattern has been extracted from the dataset, they are ranked and the top N result are returned as candidate topics [1].

E. *Soft Frequent Pattern Mmning(SFPM)*

In FPM approach provide the solution to problem of feature pivot method that account only pairwise cooccurrence between terms.FPM examines cooccurrence between any numbers of terms. This method combines features of these two methods, this method is a soft version of Frequent Pattern Mining.

This approach work by maintaining set of terms and add new terms greedy manner, it maintain vector between terms and candidate term for matching cooccurrence between them ,another vector is binary indicator that represent whether the term occurs in document or not. we achieve the "soft" matching between terms that is considered for expansion and set of terms. we can use greedy approach that expands the set of terms with the best matching term[1].

F. *BNgram*

The both method FPM and SFPM uses simultaneous cooccurrence between more than two terms, It is also possible to get similar result by using n-gram instead of unigram[1].Using n-grams makes particular sense for Twitter, since large number of status updates are just copies or retweet of previous message, so important n-gram tend to be frequent[1].This method uses a new feature selection method, that focuses on changing frequency of terms over time as a useful to detect trending topic. The main goal of this method is to find trending topics in post by comparing the term frequencies from the current slots with preceding time slots. Incoming term from the post arranges in n-grams based on their *df-idf* and storing in clusters based on their similarity and ranking them to find emerging topic.

VII. CONCLUSION

The objective of our work is to provide a need of uncovering trending stories and study different methods for uncovering tending stories from twitter. Various methods are introduced in that work which is emerged in recent years. This analysis shows that different technologies used in all the paper with taking different way for detecting trending topic for various purpose. Although applying these methods along with preprocessing to uncovering trending stories from Twitter by extracting ground truth.

REFERNCES

[1] LucaMaria Aiello,Sensing Trending Topics in Twitter, IEEE Transaction on Multimedia, Vol. 15 No.6, Oct. 2013 pp. 1268-1282.

[2] 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.

[3] . 271–350. M. Mathioudakis and N. Koudas, "Twitter monitor: Trend detection over the Twitter stream," in *Proc. SIGMOD: Int. Conf. Management of Data*, New York, NY, USA, 2010, pp. 1155–1158, ACM.

[4] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, andJ. 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.

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

[6] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *Proc.SocialCom: 3rd IEEE Int. Conf. Social Computing*, Boston, MA, USA,Oct. 2011.

[7] 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.

[8] James Allan,Stephen Harding and Devid Fisher, Taking Topic Detection From Evaluation to practce,

[9] Carlos Martin, David Corney, Ayse Goker and Andrew MacFarlane, Miniing Newsworthy Topic from Social Media.

[10] JjS. 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. 120–114.

[11] 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.

[12] Charu C. Aggarwal and Karthik Subbain, Event Detection in Social Stream

[13] Arkatiz Zubiaga, Damiano Spaina, and Victor Fresno, Real-Time Classification of Twitter , trends in American Socity for Infrmation Science and Technology,

[14] 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.

[15] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*.New York, NY, USA: McGraw-Hill, 1986.

[16] X.Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: A structural clustering algorithmfor networks," in *Proc. KDD: 13th ACM Int. Conf. Knowledge Discovery and Data Mining*, New York, NY, USA, 2007, pp. 824–833