OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

TOPIC DETECTION FROM TWITTER

Ms. Jasmin Jamadar
Department of Computer Engineering, Parvatibai Genba Moze College of Engineering, Wagholi Pune, India
jasminjamadar@gmail.com

Abstract: **From last few decades there is wide spread usage of social network platforms such as twitter or other micro blogging systems which contains huge amount of timely generated data. Tweeter is fastest means of information sharing where user shares event/news which take place in front of them. Thus Tweeter act as news portal where news reaches to the people within fraction of seconds. Extracting valuable information in timely manner is important because this wealthy information is useful for companies, government agencies and health organizations. Topic detection is the new research area in data mining and knowledge discovery where extracting useful and valuable information from timely generated online streams is the new challenge. In this article we survey the different algorithms used for trending topic and event detection using social media data and propose new system for news event detection from social media data.**

Keywords – **Twitter; Topic Detection; Term Aging; Term Co-occurrences; PageRank; Burstiness**

I INTRODUCTION

Topic detection and topic modeling is the new research area in data mining and information retrieval which has attracted several communities in last few decades. Social media is mostly used for online conversation. Some conversations are user specific and large part of conversations are responses which triggered by events e.g. political events (presidential polls), disasters (earth quack, terrorist attack), protest and marches. Twitter, Facebook, Tumblr are the popular microblogging system where people share their experience, exchange opinions in terms of short messages. Among these Twitter is the powerful means of fastest information service in the world. Billions of messages appear daily reacting to the real world event and cultural event. It is popularly known as best news paper as it informs the web community about emerging topics and events. Extracting valuable information in timely manner has gained the popularity because this wealthy information is useful for companies, government agencies and health organizations. Extracting valuable information from huge amount of data is challenging due its informal structure and diversity.

Numerous research papers have concentrated on presenting methods and systems for extracting emerging topics from social media but the proposed approach is not able to distinguish between a news event and irrelevant hot topics. Many existing system designed to detect the topics which is discussed among social media users. The limitation of existing systems is that they cannot differentiate the news events (e.g. terrorist attack) from entity events (e.g. celebrity death). In this sense, we propose an algorithm in which we distinguish news events through the entity and event detection algorithms.

II LITURATURE REVIEW

Mario Cataldi [1] studied the role and topology of Twitter. Authors analyzed the user's interest by extracting content of his generated tweets then using social relationship, directed graph of online community having active users is designed. Authors' monitored streams generated by the entire network and studied the behavior of each term using aging model. In this model authors compute the burstiness of each term using life status (i.e. Burstiness value) of each term authors ranked terms. Then using keyword graph based approach minimal set of terms is retrieved which connects the extracted terms. Considering the time frame of active user tweets authors analyzed the content of tweets, this information is used to represent the topic. The proposed approach is not able to distinguish between a news event and irrelevant hot topics.

Luca Maria Aiello [2] presented a comparative study of six different topic detection methods on three Tweeter data collections authors observed the factors like pre-processing methods of data, structure of data, outliers in the data and user activity over the time, which greatly affect quality of the final result. Authors tested six methods using three different classes namely, feature pivot methods, probabilistic model (Latent Dirichlet Allocation), Classical Topic detection and Tracking (document- Pivot approach), and proposed the new algorithm which combine the n-gram with df-idf topic ranking algorithm which give best result among the other state-of-art techniques. As an extension to this work authors suggested that this technique would be used to detect the most interesting topic within events and thus able to notify only more related topic/events which occurs.

Marta Arias [3] proposed a method that includes pre-processing of Tweeter data, building sentiment indicator and sentiment index from Tweeter data. Authors built and utilized the machine learning models i.e. linear models, neural network, several experiments has been conducted involving models with sentiment index and each of these model is trained with and without Tweeter data and performance comparison of each model is made using these experimental results authors developed the decision tree which represents the information which they termed as summary tree. Using summary tree they proved that forecasting prediction with Tweeter data is accurate as compared to forecasting prediction without Tweeter data. In the proposed work authors found that nonlinear models i.e. SVM and neural network performs better as predictor, but linear regression model unable to utilize any sentiment indices.

Hassan Sayyadi [4] suggested KeyGraph method for topic detection and tracking. KeyGraph is directed graph that represents document collection as a keyword concurrences graph. Authors then applied community detection algorithm to form the cluster of keywords into community. Each community forms grouping of keywords that represents the topic. Authors compared the accuracy and execution time of KeyGraph with other topic modelling technique like LDA-GS on spinn3r dataset.

Nargis Parvin [5] described novel method and its implementation for the detection of trending topics. Authors designed novel and context sensitive algorithm ’TrendMiner’ for detecting trending topics in microblog post streams. The proposed work has several shortcomings; first, the current approach does not consider the order of the words in trending topics. Second, when the cluster size increases, the precision of the TrendMiner method decreases.

Xiangmin Zhou [6] proposed a novel approach to detect composite social events over streams, which utilizes the information of social data over multiple dimensions. Authors implemented a graphical model called location-time constrained topic (LTT) which extracts the content, time, and location of social messages. A social message is represented as a probability distribution over a set of topics by using location-time constrained topic, and to measure the similarity between two messages the distance between their distributions is calculated. Finally by using efficient similarity joins over social media streams events are then identified. Variable dimensional extendible hash over social streams is utilized to accelerate the similarity join.

Erich Schubert [7] proposed a measure by drawing upon experience from outlier detection that can be used to detect emerging topics early, long before they become hot topic, Secondly, by using hash tables in a heavy-hitters type algorithm for establishing a noise baseline. Finally by applying clustering algorithm the detected co-topics are aggregated into larger topics, as often as a single event will cause multiple word combinations to trend at the same time.

Feng Chen [8] has suggested an approach based on Non-Parametric Heterogeneous Graph Scan (NPHGS), which detect the event by considering the entire heterogeneous network: author using sensor networking concept in which each node is sensor senses its environment and reports an empirical p- value. Then, they efficiently maximize a nonparametric scan statistic over connected sub graphs to identify the most anomalous network clusters. At this point author represented the event by each cluster which is represented with information such as type of event, geographical spot, time, and participants. The limit of NPHGS is it can’t provide rich domain knowledge naturally.

Tim Althoff [9] studied three major online and social media streams, Twitter, Google, and Wikipedia, covering thousands of trending topics during an observation period of an entire year Author presented a novel technique for forecasting the life cycle of trending topics in the very moment they emerge. In this, nearest neighbor forecasting technique is used based on author’s assumption that similar behavior topics are semantically similar topics. The proposed model is unable to explicitly detect and exploit seasonality as well as incorporate global changes in viewing statistics.

Junjie Yao [10] presented a novel approach called a unified user-temporal mixture model for detecting stable and temporal topics simultaneously from social media data. Proposed method distinguishes temporal topics from stable topics. Authors improved the performance of their model by designing a regularization framework that utilizes prior spatial information in a social network, as well as a burst-weighted smoothing scheme that exploits temporal prior information in the time dimension.

Ceren Budak [11] proposed two novel structural trend definitions called coordinated and uncoordinated trends that use friendship information to identify topics that are discussed among clustered and distributed users respectively. Author also introduced a novel information diffusion model called Independent Trend Formation Model (ITFM) that distinguishes viral diffusion of information. Author also exploits a sampling technique for structural trend detection that provides computational gain as well as a solution within an acceptable error bound. Unfortunately, the proposed techniques cannot be used as a substitute for traditional trend detection but rather as a compliment.

III PROBLEM STATEMENT

Social media platform works as news channel which informs its user about event that happened. Twitter is the popular micro blogging media through which user shares information. Many research works are done to extract useful and valuable information from huge and unstructured data. Topic detection is the new research area where topics which are discussed among the social community users are detected. As the social data i.e. tweets contains both entity (person) and event specific information, many existing system detects the most emergent topics but they cannot differentiate topic such as person related and event related topics. In this sense we propose a method that detects the most emergent topics and also differentiate news events from irrelevant topics. Our aim is to form the clusters of news event which considers only the news event(e.g earth quake, terrorist attack, cyclone etc).

IV PRAPOSED FRAMEWORK

By extending the work of Mario Cataldi [1] we initially analyze user interests by extracting and formalizing the content of her generated tweets. We then model the social community as a directed graph of the active authors based on their social relationships, and calculate their authority by relying on the well-known PageRank algorithm. In the proposed framework we calculate the burstiness value of each term expressed by user in the tweets and also compute reputation of each author. We select the set of most emerging keywords by dynamically ranking the terms depending on their life status (defined through a burstiness value). We represent each related to each emergent term by constructing and analyzing a keyword graph which links the extracted emerging terms with all their co-occurrence keywords. At this point, by applying the traversal method on keyword graph we create the set of emerging topics. Finally using clustering algorithm the news events are grouped together to form news event cluster. Various steps of proposed work are shown in fig.1 and in the following section report these steps in detail

A. Pre-processing

- Stop words removal: In the proposed framework we apply preprocessing techniques like stop word removal. We remove unwanted words or stop words from data.
- Stemming: Stemming usually refers to a heuristic process that chops off the ends of words, including the removal of derivational affixes. We also apply stemming operations on every word.

B. Term aging model

The basic step for information retrieval is the extraction of relevant keywords from the tweets. Considering time interval It, we associate each tweet twj with tweet vector Vtj. Each component of the vector Vtj represents a weighted term extracted from the related tweet twj .

- Weight- tf-idf: This measure is used to calculate the importance of term to the document in this case it is tweet. We define the tf-idf for terms in the tweet as

  $w_{x,j} = tf\text{-}idf = tf_{x,j} \times idf_{x,j}$ Eq. 1

  Where, wx, j of the x-th vocabulary term in j-th tweet. tfx,j is term frequency value of the x-th vocabulary term in j-th tweet and idfx,j is inverse document frequency x-th vocabulary term in j-th tweet. The tweet vector Vt can obtained using tf-idf value for the jth tweet as

  $V_{tj} = \{w_{1,j}, w_{2,j}, \ldots\ldots. w_{v,j}\}$ Eq. 2
- Nutrition Value: We calculate the nutrition value of term in the tweets generated by different users as follows.

  $nutr_k^t = \sum_{tw_j \in TW} w_{k,j} * rep(user(w_j))$ Eq.3

  where wk,j represents the weight of the term k in the tweet vector Vtj, the function user(wj) returns the author u of the tweet twj, and rep(u) returns the reputation value associated to u.

This nutrition formula evaluates the usage of term by considering its frequency in the tweets that mention it.

While emerging topic detection it is important to consider the source of messages which are the set of users. To extract the contents from the tweets is necessary to figure out the level of importance of source i.e. user. So we compute the reputation of user by analyzing community graph.

- Reputation Calculation: We calculate the reputation for every user using number of followers and there weighted words. The formula for computation of reputation of user is as follows,

  $rep(u_i) = d \times \sum_{u_j \in follower(u_i)} rep(u_j) / following(u_j) + (1 - d)$ Eq.4

  Where d∊(0, 1) is a dumping factor, follower(ui) is a function that returns the set of users following uj, and following(uj) returns the set of users that uj follows.

After calculating nutrition value we map this value to the value of burstiness.

- Term Burstiness Values: The burstiness value of a term indicates its actual contribution (i.e., how much it is emergent) in the corpus of tweets. Our idea is that the temporal information associated to the tweets can be used as distinguish function in that sense.

  $burst_k^t = \sum_{x=t-s}^{t-1} (((nutr_k^t) - (nutr_k^x)).1/\log(t-x+1))$ Eq.5

This formula provides a way to find the usage of a given keyword k with respect to its previous usages in a limited number of time intervals.

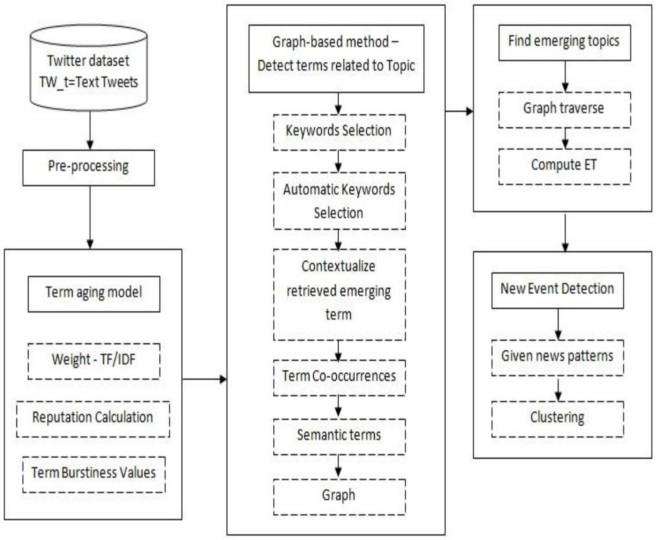


Figure 1 Proposed Framework

C. Detect terms related to Topic

After evaluating burstiness value of the terms we select a limited set of the emerging term. There are two methods for keyword selection: user driven and fully automatic technique.

- User driven keyword selection: The first approach for the selection of emerging terms relies on a user-specified threshold parameter. Critical drop value used to find the term which most emergent and is given as

  dropt = δ · ∑k∈Kt (bursttk)/|Kt| Eq.6

  where δ ≥1.
- Automatic keyword selection: In order to cope with this, we provide a completely automatic model that works as follows.
  1. The system first ranks the keywords in descending order of burstiness value previously calculated in.
  2. It computes the maximum drop between consecutive entries and identifies the corresponding drop point.
  3. It computes the average drop (between consecutive entities) for all those keywords that are ranked before the identified maximum drop point.
  4. The first drop which is higher than the computed average drop is called the critical drop.

  At the final step, the keywords ranked before the critical drop are emerging keywords and set of emerging keywords is EKt at time interval It.
- Contextualize retrieved emerging term: Considering the given corpus of tweet, tweets are extracted within the time interval, in this step we study the semantic relationships that exist among the keywords in it, in order to retrieve the topics related to each emerging term.
- Term Co-occurrences: We get various terms and calculate the co-occurrences for each word. Using similarity function we calculate the semantic relationship between the two keywords.
- Graph: We create the graph by using User, word, TF co-occurrences. We construct a keyword-based topic graph in the form of a directed, node-labeled, edge-weighted graph, TGt(Kt, E, ρ) Kt be a set of vertices, where each vertex kϵ Kt represents the keyword extracted during the time interval It. E is the edge that represents the relationship that exist between the two keywords. ρ represents relative weight of one keyword with respect to other keyword. Following fig2 represents the graph which shows the relationship that exists among various terms. The strong relationship between the terms is represented by dark arrows.

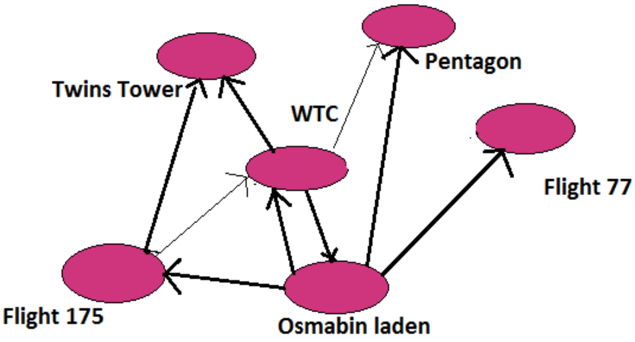


Figure 2 A topic graph representing emerging topic

D. Find Emerging Topics

We traverse the graph using depth first search traversal and find the set of emergent topics. Also get the news patterns and find the news events from twitter.

Whatever emergent topics are detected we order them by ranking. We calculate ranking value for topic by using formula,

rankET=∑(burstkt)/|Kzt| Eq.7

E. Clustering

At this point, we apply the clustering algorithm to the set of emergent topics in order to differentiate the news event from irrelevant topics and create clusters of news events. In the proposed framework we use density based clustering algorithm. As there are many clustering algorithms, like connectivity based clustering, centroid based clustering, Distribution based clustering. For our work we choose DBSCAN, density based clustering algorithm because DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. In our works after detecting emergent topic we can start forming cluster. Again DBSCAN is significantly more effective in discovering clusters of arbitrary shape.

V MATHEMATICAL MODEL

A. Set Theory

1. Let S be a system that describes Topic detection system that detect emerging topics from online social media.

S= {................}

2. Identify input as I

S= {I......

I = {Q | where Q is input dataset

3. Identify output as O

S= {I, O .....

O={ Ri | Ri is output emerging topic that represents event in which users are interested

4. Identify the processes as P.

S= {I, O, P..........

P= {PD, TAM, GG, ETD, NEC}

* PD is pre-processing of data

* TAM is term aging modeling.

* GG is graph generation

*ETD is emerging topic detection

*NEC is news event clustering

5. PD = {Dataset , PDataset}

Dataset is data of tweeter given as input

PDataset is output of pre-processing phase

6. TAM = { PDataset, Formlae ,WT}

PDataset is input which is output of PD is given to TAM

Formulae are set of equations used to calculated weights of term and reputation of user

WT is output of TAM

7. GG = {PDataset, WT, G}

PDataset and WT are given as input to generating graph

G is the graph which is output of graph generation

8. ETD = {G, ET}

G is graph given as input to emerging topic detection.

ET is the emerging topics which is the output of emerging topic detection

9. NEC = {ET, NE}

ET is the emerging topics given as input to news event clustering

NE is the news events which the output.

8. Identify failure cases as F.

S = {I, O, P, F...

Failure occurs when O≠Ri

9. Identify Success cases as s

S= {I, O, P, F, s...

Success is defined as O= Ri

10. Identify initial condition IC

S= {I, O, P, F, s, IC}

There is no initial condition IC= {null}

VI EXPERIMENTAL EVALUATION

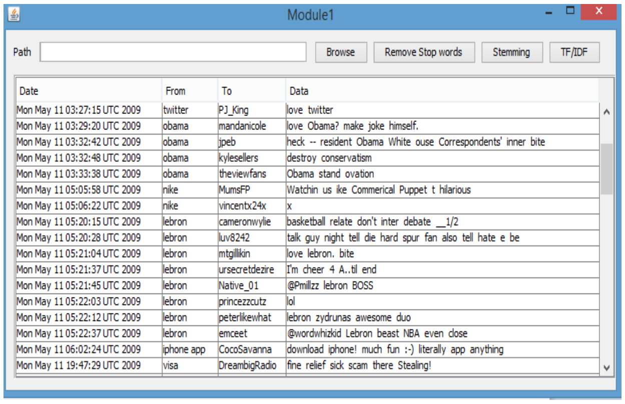
A. EXPERIMENTAL SETUP

We evaluate the proposed method by analyzing real case studies. In particular, we conduct several experiments by monitoring the twitter community during the period included from 11th to 24th May 2009.

Considering the topic detection method introduced in this paper, we analyze different experiments: we firstly analyze a real case study by evaluating the emerging topics retrieved by the system within the considered time interval. Then, we vary the number of considered time slots in order to study how this parameter affects the quality of the retrieved topics. Secondly, considering the users' interests, we ask different users to provide a real feedback about the retrieved personalized topics for evaluating the precision of the presented methods. Finally, we study the difference between the supervised and unsupervised selections methods evaluating their impact on the resulting emerging topics.

B. EXPERIMENTAL RESULTS

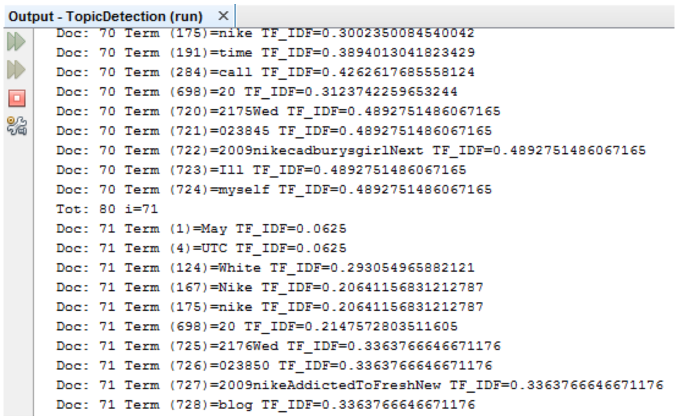
Following figure 3 shows the result obtained after preprocessing of tweeter dataset. It includes the removal of stop words and stemming. For stemming, WordNet dictionary is used.



Date	From	To	Data
Mon May 11 03:27:15 UTC 2009	twitter	PJ_King	love twitter
Mon May 11 03:29:20 UTC 2009	obama	mandanicole	love Obama? make joke himself.
Mon May 11 03:32:42 UTC 2009	obama	jpeb	heck -- resident Obama White ouse Correspondents' inner bite
Mon May 11 03:32:48 UTC 2009	obama	kylesellers	destroy conservatism
Mon May 11 03:33:38 UTC 2009	obama	theviewfans	Obama stand ovation
Mon May 11 05:05:58 UTC 2009	nike	MumsFP	Watchin us ike Commerical Puppet t hilarious
Mon May 11 05:06:22 UTC 2009	nike	vincentx24x	x
Mon May 11 05:20:15 UTC 2009	lebron	cameronwylie	basketball relate don't inter debate __1/2
Mon May 11 05:20:28 UTC 2009	lebron	luv8242	talk guy night tell die hard spur fan also tell hate e be
Mon May 11 05:21:04 UTC 2009	lebron	mtgillikin	love lebron. bite
Mon May 11 05:21:37 UTC 2009	lebron	ursecretdezire	I'm cheer 4 A..til end
Mon May 11 05:21:45 UTC 2009	lebron	Native_01	@Pmillzz lebron BOSS
Mon May 11 05:22:03 UTC 2009	lebron	princezzcutz	lol
Mon May 11 05:22:12 UTC 2009	lebron	peterlikewhat	lebron zydrunas awesome duo
Mon May 11 05:22:37 UTC 2009	lebron	emceet	@wordwhizkid Lebron beast NBA even close
Mon May 11 06:02:24 UTC 2009	iphone app	CocoSavanna	download iphone! much fun :-) literally app anything
Mon May 11 19:47:29 UTC 2009	visa	DreambigRadio	fine relief sick scam there Stealing!

Figure 3 Output after preprocessing

The following figure 4 shows the output of term weighting using tf_idf algorithm. Document number represents the tweet number to which term belongs also position of term in the tweet is shown along with the tf-idf value.



```
Doc: 70 Term (175)=nike TF_IDF=0.3002350084540042
Doc: 70 Term (191)=time TF_IDF=0.3894013041823429
Doc: 70 Term (284)=call TF_IDF=0.4262617685558124
Doc: 70 Term (698)=20 TF_IDF=0.3123742259653244
Doc: 70 Term (720)=2175Wed TF_IDF=0.4892751486067165
Doc: 70 Term (721)=023845 TF_IDF=0.4892751486067165
Doc: 70 Term (722)=2009nikecadburysgirlNext TF_IDF=0.4892751486067165
Doc: 70 Term (723)=Ill TF_IDF=0.4892751486067165
Doc: 70 Term (724)=myself TF_IDF=0.4892751486067165
Tot: 80 i=71
Doc: 71 Term (1)=May TF_IDF=0.0625
Doc: 71 Term (4)=UTC TF_IDF=0.0625
Doc: 71 Term (124)=White TF_IDF=0.293054965882121
Doc: 71 Term (167)=Nike TF_IDF=0.20641156831212787
Doc: 71 Term (175)=nike TF_IDF=0.20641156831212787
Doc: 71 Term (698)=20 TF_IDF=0.2147572803511605
Doc: 71 Term (725)=2176Wed TF_IDF=0.3363766644671176
Doc: 71 Term (726)=023850 TF_IDF=0.3363766644671176
Doc: 71 Term (727)=2009nikeAddictedToFreshNew TF_IDF=0.3363766644671176
Doc: 71 Term (728)=blog TF_IDF=0.3363766644671176
```

Figure 4 Output after weighting the terms

VII CONCLUSION

The related work presented in the paper showed that Twitter is the ideal scenario for the study of real-time information spreading phenomena. All the proposed methods are not able to distinguish between a news event and irrelevant hot topics (e.g., discussions on facts about celebrities). Considering this we work on this direction. In this paper we presented approach to detect in real-time emerging topics on Twitter. We formalize the keyword life cycle using burstiness value intended to mine terms that frequently occur in the specified time interval and they are relatively rare in the past. We also study the social relationships in the community network in order to evaluate the importance of each analyzed contents. At this stage, we formalize a keyword-based topic graph which connects the emerging terms with their co-occurrent ones, allowing the detection of emerging topics under user-specified time constraints. Finally by using clustering algorithm we differentiate the news events form irrelevant topics.

REFERENCES

[1] Mario Cataldi, Luigi Di Caro and Claudio Schifanella, “Personalized Emerging Topic Detection Based on a Term Aging Model”, ACM Transactions on Intelligent Systems and Technology, Vol. 5, No. 1, Article 7, December 2013.

[2] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, “Sensing Trending Topics in Twitter”, IEEE Transactions on multimedia, Vol. 15, NO. 6, October 2013.

[3] Marta Arias, Argimiro Arratia, and Ramon Xuriguera, “Forecasting with Twitter Data”, ACM Trans. Intell. Syst.Technol. 5, 1, Article 8, December 2013.

[4] Hassan Sayyadi and Louiqa Raschid, “A Graph Analytical Approach for Topic Detection”, ACM Trans. Internet Technol. 13, 2, Article 4 December 2013.

[5] Pervin, N., Fang, F., Datta, A., Dutta, K., and Vandermeer, “Fast, scalable, and context-sensitive detection of trending topics in microblog post streams”, ACM Trans. Manage. Inf. Syst. 3, 4, Article 19, January 2013.

[6] Xiangmin Zhou, Lei Chen, “Event detection over twitter social media streams”, © Springer-Verlag Berlin Heidelberg 2013.

[7] Erich Schubert, Michael Weiler, Hans-Peter Kriegel,” SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds”, ACM Transactions on Intelligent Systems, August 2014.

[8] Feng Chen, Daniel B. Neill,” Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs”, ACM, August 2014.

[9] Tim Althoff Damian Borth Jörn Hees Andreas Dengel, “Analysis and Forecasting of Trending Topics in Online Media Streams”, ACM MM’13, October 21–25, 2013.

[10] Junjie Yao, "A unified model for stable and temporal topic detection from social media data", ICDE, 2013, 2013 29th IEEE International Conference on Data Engineering (ICDE 2013), 2013 29th IEEE International Conference on Data Engineering (ICDE 2013) 2013.

[11] Ceren Budak, “ Structural Trend Analysis for Online Social Networks”, 2011 VLDB Endowment.