RESEARCH ARTICLE OPEN ACCESS

Evolving Swings (topics) from Social Streams using Probability Model

Shweta.G.Hiremath[1], Anand.S.Pashupatimath[2]
[1]PG Student, Department of CSE, SDMCET, Dharwad, Karnataka, India
[2]Assistant.Professor , Department of CSE, SDMCET, Dharwad, Karnataka, India

Abstract:
Evolving swings from social streams is receiving renewed interest and it is motivated by the growth of social media and social streams. Non-conventional based approaches can be appropriate which include text, images, URLs and videos. The focus is on evolving topics by social aspects of the networks and the mentions of user links between users which are generated intentionally or unintentionally through replies, mentions and retweets. A probability model of the mentioning behavior is proposed and the proposed model detects the evolving topic from the anomalies measured. After a several experiments, it shows that mention anomaly based approaches detects the evolving swing as early as text anomaly based approaches.

Index Terms- burst detection, SDNML, social networks, Swing detection

I. INTRODUCTION

Communication between people is increasing through many ways such as phone, internet, newspaper etc. Social networks such as facebook, twitter have become a very vast network where people are discussing about the evolving topics and activities with their friends, families and many other unknown people. Communication over social networks is gaining importance day by day and these networks spread the news in a fast manner and the original talk of the people is spread. The swings discussed can be a hot topic or their personal day today activities. Here, we are interested in the problem of detecting evolving topics from social media and it discovers the market needs or political movements. In the conventional media, social media are unable to capture at the earliest and in the non-conventional media, social media are able to capture at the earliest. A challenging testbed is to detect the evolving swing as early as possible.

In social media, there is existence of mentions. Here, mentions are the links to the other users of same social network; these links can be in the form of reply-to, message-to and retweet-of or explicitly in the text. A single post can contain a number of mentions. Some users include mentions in their posts very rarely, where as some users include regularly. Users like celebrities can be mentioned every minute, for others being mentioned can be rare occasion. Hence, mention is like a language where the number of words is equal to the number of users in social network. We are interested in detecting evolving swings from social streams based on monitoring the mentioning behavior of users. Our assumption is that a evolving topic(swing) is the one which people feel like commenting, discussing or sharing the information to their friends. Conventional based approaches are concerned with the frequencies of words. This approach suffers from ambiguity caused by synonyms or homonyms, and it cannot be applied for the non-textual information

We have considered an example of the evolving topic through the posts on twitter. In the Fig.1 A mentions B and C in his first post, which are both friends of A. A gets a reply from C, and this second post is visible to friends of C and they are not direct friends of A. D one of C's friend retweets the information to his own friends. From the textual information it is not clear about the topic of the conversation because it is shown as link in the text.

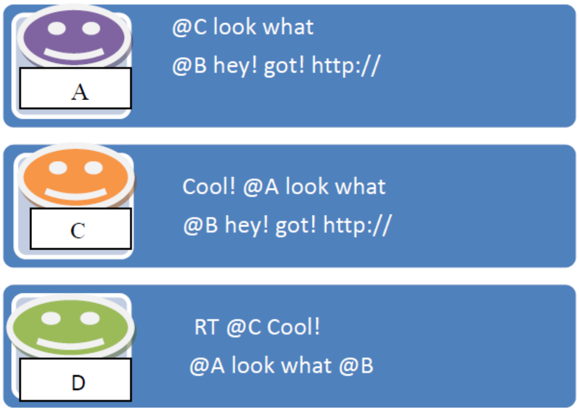


Fig.1. Example of evolving topic in social streams

The normal mentioning behaviour of a user can be captured by our proposed probability model, which consists the number of mentions per post and the frequency of users appeared in that mentions. This model measures the anomaly of future user behavior.

II.LITERATURE REVIEW

Topic Detection and Tracking (TDT) [1] is a DARPA-sponsored initiative to investigate the state of the art in finding and following new events in a stream of broadcast news stories. The TDT Pilot Study ran from September 1996 through October 1997. The primary participants were DARPA, Carnegie Mellon University, Dragon Systems, and the University of Massachusetts at Amherst. This summarizes the findings of the pilot study. The TDT work continues in a new project involving larger training and test corpora, more active participants, and a more broadly defined notion of "topic" than was used in the pilot study.

In 1998, the event based information organization started the topic detection and tracking project (Allan, 2002). The project had a different tasks i) segmentation ii) tracking iii) detection iv) first story detection and v) linking

Tracking: This task detects the stories which discuss previously known target topic..

Detection: Detection of new and the unseen topics. It is called on-line clustering first-story detection. The good of this task is to detect the very first story to discuss the previously unknown event.

A fundamental problem in text data mining is to extract meaningful structure from document streams that arrive continuously over time [2]. E-mail and news articles are two natural examples of such streams, each characterized by topics that appear, grow in intensity for a period of time, and then fade away. The published literature in a particular research field can be seen to exhibit similar phenomena over a much longer time scale. Underlying much of the text mining work in this area is the following intuitive premise --- that the appearance of a topic in a document stream is signaled by a "burst of activity," with certain features rising sharply in frequency as the topic emerges. The goal of this work was to develop a formal approach for modeling such "bursts," in such a way that they can be robustly and efficiently identified,

Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai are exercised in time series with the problem of real-time change-point detection [3]. This technology received vast attentions in the field of data mining as it can be applied to different risk management issues. In this paper by applying SDNML they have proposed a advanced method of real-time change point detection. Here the SDNML is a method used for sequential data compression. It reaches the bottom code length for the sequence and as the time goes the effect of past data is moderately discounted. Hence the data compression can be done robustly to non-stationary data sources. The SDNML is used to study the working of a time series, then each time a change-point score is measured in the form of SDNML code-length.

Model selection by means of the predictive least squares (PLS) [4] principle has been thoroughly studied in the context of regression model selection and autoregressive (AR) model order estimation. J. Rissanen, T. Roos, and P. Myllyma¨ki introduce a new criterion based on sequentially minimized squared deviations, which are smaller than both the usual least squares and the squared prediction errors used in PLS and proved that their criterion has a probabilistic interpretation as a model which is asymptotically optimal within the given class of distributions by reaching the lower bound on the logarithmic prediction errors, given by the so called stochastic complexity, and approximated by BIC.

Syslog monitoring technologies [5] have recently received vast attentions in the areas of network management and network monitoring. They are used to address a wide range of important issues including network failure symptom detection and event correlation discovery. Syslog are intrinsically dynamic in the sense that they form a time series and that their behavior may change over time. This paper proposes a new methodology of dynamic syslog mining in order to detect failure symptoms with higher confidence and to discover sequential alarm patterns among computer devices. The key ideas of dynamic syslog mining are 1) to represent syslog behavior using a mixture of Hidden Markov Models, 2) to adaptively learn the model using an on-line discounting learning algorithm in combination with dynamic selection of the optimal number of mixture components, and 3) to give anomaly scores using universal test statistics with a dynamically optimized threshold.

Text Stream Mining

Petrovic (2010) justified his decision of performing Event detection. Stream models: 1. Stream based machine translation (Levenberg & Osborne, 2009) 2. Approximation kemel matrices of data streams (Shi et al, 2009) topic modeling on streaming document collection (Yao et al, 2009) Gamma (2010) described data stream models to solve text stream event detection problem such as burst detection, change detection, clustering, time series analysis and novelty detection. Rajaram and Ullman(2011) dedicated to the mining of data streams and proposed simple algorithms for filtering counting distinct elements in stream and giving some insights about today’s relevance of data.

SNA

To identify an emerging topic, Agarwal et al (2012) identified a set of keywords which are Temporally correlated and they co-occur in temporally correlated message from the same user.

III.EXISTING SYSTEM

A topic or swing is something people feel like discussing, forwarding or commenting the information further to their friends. Conventional based approaches for topic detection was mainly concerned with frequencies of words (textual). But the frequency based approach suffers from ambiguity caused by synonyms or homonyms. This cannot be applied if the contents of the messages are non-textual information and it requires complicated preprocessing.

IV.PROPOSED SYSTEM

In a commercial era, many areas are dealing with social networks, stream monitoring services and knowledge management. It is important to discover the swings and analyze them in real-time. It is desired in the social stream area to grasp new swings in online. A swing is an event or activity detection. Proposed system overcomes the limitation of existing system. Limitations of an existing system is that the contents of the message can only be a non-textual information, but in our proposed system it works well with both textual and non-textual information such as images, URLs and videos.

The basic idea here is to focus on the social aspects of the posts reflected in the mentioning behaviour of user instead of textual contents. We proposed a probability model which captures the mentions per post and the frequency of mentionee. Proposed system should be able to mine continuously, high volume, social stream documents as they arrive and it will be ready to detect new swings at any time.

This model works by analyzing the content of the message such as text, image or video and calculates the outlier scores which is generated during sharing of posts. Fig.2 shows the overall flow of the proposed method, in the training phase A's posts and B's posts are taken and the mention distribution is known and then the individual score is calculated and are aggregated. The aggregated score is feed into SDNML based change point analysis and then burst detection can be used instead of SDNML.

V.IMPLEMENTATION

Architecture

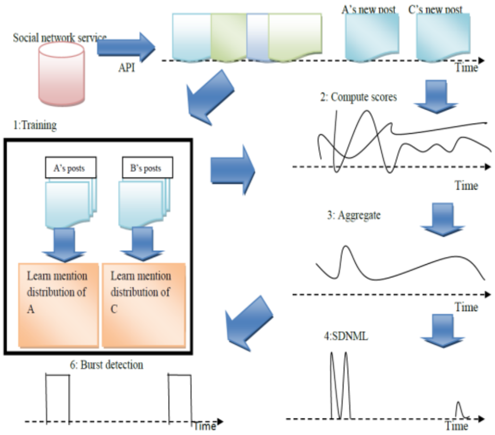


Fig.2. Architecture and overall flow

The proposed work can be described with the system design and system modules.

System design

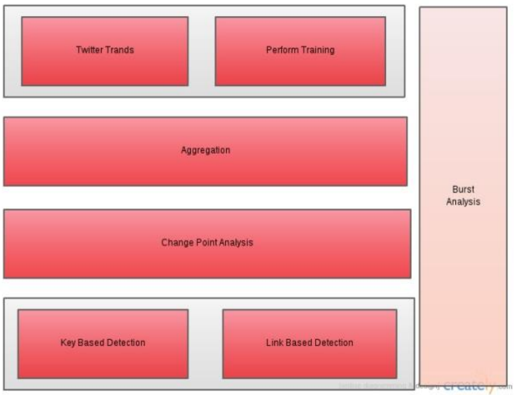


Fig.3. Block diagram

Algorithm

Step 1: The dataset from the social network is obtained using API (Twitter trends are fetched)

Step 2: Sharing pattern of user is analysed.

Step 3: Prediction of sharing of topic is done using change point detection method.

Step 4: Content of the message shared is analysed.

Step 5: Training is performed on fetched data.

Step 6: Aggregation of individual score is obtained.

Step 7: Sharing of topic is predicted using change point and burst analysis.

Step 8: Repeat the steps for link based detection.

Step 9: Resulting bar chart of key based detection and link based detection is displayed.

Fig.3. shows the block diagram of the system and it comprises of fetching of the twitter trends and then performing training. The result of aggregation of individual scores is obtained. The change point analysis through SDNML and burst detection. These steps are repeated for both key based detection and link based detection.

System modules

The model works by analyzing which post is forwarded (which is called retweet in twitter) to many of their friends. Fig.1 shows an example of the evolving topic through twitter. If a particular post is becoming as outlier (anomaly score) then its content is analyzed for text message, images and videos.

The proposed model has following module

1. Training
2. Identify individual Anomaly Score
3. Aggregate
4. Change Point Analysis and DTO
5. Burst Detection

Training phase:

In this training phase, the past behavior is considered and the posts which are shared with their friends are extracted from social network dataset (Twitter dataset) using a social network API for analyzing the retweet or forwarding behavior of user. Here, numbers of user k are mentioned in the post and Ids (names of user mentioned in the post) is taken as set V. The number of users mentioned is limited by geometric distribution.

$$P(k,V|\theta,\{\pi_v\})) = P(k|\theta)\prod_{v'\epsilon V}\pi_{v'} \tag{1}$$

The joint probability distribution is calculated with k and V to predict the probability of each user mentioned in the forwarding list

Suppose there are n posts then predictive distribution is

$$P(k,V|T) = P(k|T)\prod_{v\in V}P(v|T) \tag{2}$$

To measure the general trend, we propose to aggregate the scores obtained for the posts,

$$s'_j = \frac{1}{T}\sum_{t_i\in[T(j-1),t_j} s(x_i) \tag{3}$$

Identify Individual Anomaly score

Here the deviation of user's behaviour from norma mentioning behaviour is computed.

Aggregate

Here, the anomaly scores are combined from different users. It is enumerated for every user based on past behaviour and user's current post.

SDNML- change point detection

This method is used to find change point from sequence of anomaly score for all post and it is done through two layers of process.

First Layer: Collection of aggregated anomaly score and is calculated in specific time period. Using density function outliers is detected.

Second Layer: The outlier which is detected in first layer is used again and the change point is detected.

Consider,

The aggregated anomaly score from time period 1 to j-1. Outlier is detected using the following density function

Finally, by thresholding we need to convert the change point scores into binary alarms.

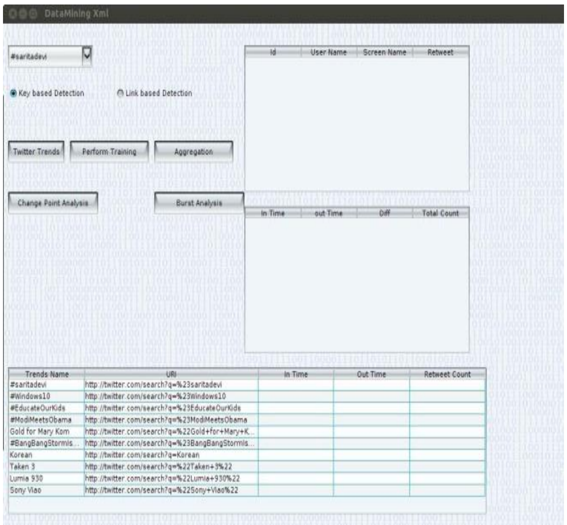
Analyzing the content of post: Once the change point is identified in the aggregated score, using the above two techniques the post can be confirmed as the dynamic post which carries the evolving topic, the content of the post should also be analyzed since the anomaly score is calculated based only on the link that is generated while forwarding or retweeting.

Burst detection:

We have change point detection method and is based on SDNML followed by DTO. The burst detection method is based on an automaton model with two states, burst state and non-burst state. Burst detection method estimates the state transition sequence. An event is defined as point in time when the anomaly score exceeds threshold value.

VI.RESULT ANALYSIS

The following results show the trends name and its ID and shows that link based detection performance are better than key based detection.



Id	User Name	Screen Name	Retweet

In Time	out Time	Diff	Total Count

Trends Name	URI	In Time	Out Time	Retweet Count
#saritadevi	http://twitter.com/search?q=%23saritadevi			
#Windows10	http://twitter.com/search?q=%23Windows10			
#EducateOurKids	http://twitter.com/search?q=%23EducateOurKids			
#ModiMeetsObama	http://twitter.com/search?q=%23ModiMeetsObama			
Gold for Mary Kom	http://twitter.com/search?q=%22Gold+for+Mary+K...			
#BangBangStormis...	http://twitter.com/search?q=%23BangBangStormis...			
Korean	http://twitter.com/search?q=Korean			
Taken 3	http://twitter.com/search?q=%22Taken+3%22			
Lumia 930	http://twitter.com/search?q=%22Lumia+930%22			
Sony Viao	http://twitter.com/search?q=%22Sony+Viao%22			

Fig.4. Fetch the twitter trends

The twitter trends are fetched and are displayed with trends name and its respective URI

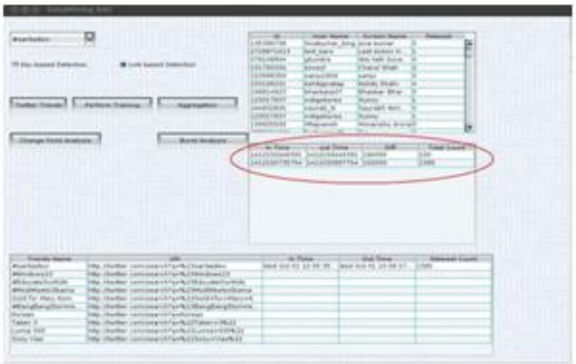


Fig.5. Change point analysis

In the Fig.5 the change point analysis is detected and prior to that the retweet counts of every user is obtained and the score is calculated and the aggregation result with retweet count is obtained. Finally change point detection varies and in link based detection total count varies.

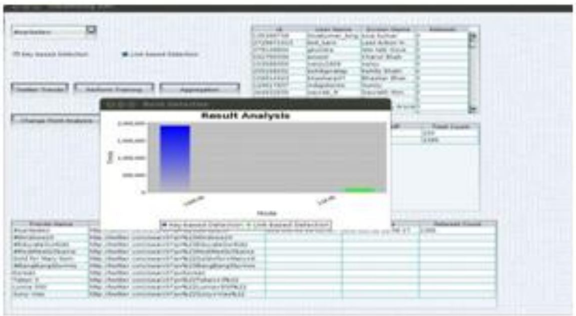


Fig.6. Result analysis

Result analysis between link based detection and key based detection : The blue is shown for key based detection and green is shown for link based detection. By the result analysis shown we proved that link based detection is efficient and give best results compared to key based detection. Link based detection works best with text, images, URLs and videos.

CONCLUSION

In this paper, we have proposed a new approach which evolves the originating topic in a social network stream. We have proposed the probability model which captures the number of mentions per post and also the frequency of mentionee. The individual anomaly score is calculated and then the aggregated score is calculated. We have combined the mention model with SDNML change point detection algorithm [3] and Klienberg's burst detection model [2] to pinpoint the emergence of a topic. The proposed method does not rely on textual contents of social network posts, it can be applied with information other than texts such as video, images, URLs and so on.

In our approach, the twitter trends are fetched through API and the training is performed separately for both key based detection and link based detection. In the link based detection retweet count or forwarding behavior of each user is obtained and the change point analysis is done. The results show that link based detection is efficient and takes less time and it will not rely on text streams. Link based detection performance is better than key based detection. The proposed approach handles the social streams in real time.

ACKNOWLEDGEMENT

I express my deep sense of gratitude to Prof. Anand.S.Pashupatimath, Department staff and college for their valuable guidance, keen interest and encouragement at various stages of my project. Great many thanks to my parents and friends who gave me something to light along my pathway.

REFERENCES

[1] J. Allan et al., "Topic Detection and Tracking Pilot Study: Final Report," Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," Data Mining Knowledge Discovery, vol. 7, no. 4, pp. 373-397, 2003.

[3] Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, "Real-Time Change-Point Detection Using Sequentially Discounting NormalizedmMaximum Likelihood Coding," Proc. 15th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD' 11),2011.

[4] S. Morinaga and K. Yamanishi, "Tracking Dynamics of Topic Trends Using a Finite Mixture Model," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 811-816, 2004.

[5] Q. Mei and C. Zhai, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining, pp. 198-207, 2005.

[6] Krause, J. Leskovec, and C. Guestrin, "Data Association for Topic Intensity Tracking," Proc. 23rd Int'l Conf. Machine Learning (ICML' 06), pp. 497-504, 2006.

[7] D. He and D.S. Parker, "Topic Dynamics: An Alternative Model of Bursts in Streams of Topics," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 443-452, 2010.

[8] H. Small, "Visualizing Science by Citation Mapping," J. Am. Soc. Information Science, vol. 50, no. 9, pp. 799-813, 1999.

[9] D. Aldous, “Exchangeability and Related Topics,” _ Ecole d’ _ Ete´ de Probabilite´s de Saint-Flour XIII—1983, pp. 1-198, Springer, 1985.

[10] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical Dirichlet Processes,” J. Am. Statistical Assoc., vol. 101, no. 476, pp. 1566-1581,2006.

[11] Daniel C. Berrios', Richard M. Keller, “Semantic Analysis of Email Using Domain Ontologies and WordNet”, Source of Acquisition NASA Ames Research Center.

[12] J. Rissanen, T. Roos, and P. Myllyma¨ki, “Model Selection by Sequentially Normalized Least Squares,” J. Multivariate Analysis, vol. 101, no. 4, pp. 839-849, 2010.

[13] C. Giurc_aneanu and S. Razavi, “AR Order Selection in the CaseWhen the Model Parameters Are Estimated by Forgetting Factor Least-Squares Algorithms,” Signal Processing, vol. 90, no. 2, pp. 451-466, 2010.

[14] C. Giurc_aneanu, S. Razavi, and A. Liski, “Variable Selection in Linear Regression: Several Approaches Based on Normalized Maximum Likelihood,” Signal Processing, vol. 91, pp. 1671-1692, 2011.

[15] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi, Member, IEEE, “Discovering Emerging Topics in Social Streams via Link- Anomaly Detection”, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, January 2014.

[16] K. Yamanishi and Y. Maruyama, “Dynamic Syslog Mining for Network Failure Monitoring,” Proc. 11th ACM SIGKDD Int’l Conf. Knowledge Discovery in Data Mining, pp. 499-508, 2005.

[17] J. Rissanen, T. Roos, and P. Myllyma¨ki, “Model Selection by Sequentially Normalized Least Squares,” J. Multivariate Analysis, vol. 101, no. 4, pp. 839-849, 2010.