Search to Unfold Topics in the Stream of Social Networks

P Geethanjally[1] | Keerthana Devaraj[2] | M Pavithra[3] | Shilpa Xavier[4] |
D Tamilselvi[5] | R Dinesh Kumar[6]

[1,2,3,4] UG Scholars, ,[5,6]Professor,Department of Computer Science and Engineering, KTVR Knowledge Park for Engineering and Technology, Coimbatore, India, me.dineshkumar@gmail.com[6])

Abstract— *Search to find the topic at peak discussion has now become a new trend among common people using social stream. The new topic is chosen based on the common agreement that a number of people in the same social stream come in as conclusion. Topics are either in the form of text, images, URLs and videos. Extraction of the URL is regardless of the post's format. The posts are discovered depending on the number of mentions it has obtained. In cases of similar post an aggregate is computed of the number of mentions received by them. These detected topics are not just circulated among friends or friends of friends but shared among all the users in the same social stream. Link between the users in the same social network is made through messaging, replying, and retweet or explicitly through text. The topics are unfold based on what people discuss on and forward to friends in their circle. Recognizing of new (emerging) topics is through conventional based approach.*

Keywords— Social Stream, Social Networking, Search Engines

1. INTRODUCTION

Social Networks is an environment to build relationship among people who share their views, events, resources or real status. Social networking service includes circumstances and direction. The content which is exchanged over the network is considered to be the resource. Information exchanged may also be relation with money or goods or services in the real life.

We are concerned about finding out the emerging topics in social streams as soon as possible and bring them to light. This is done based on the mentioning behavior of users where some users have frequent mentioning and others do not. Link between the users of the same social network is made through messaging, replying, and retweet or explicitly through text. In this paper we unfold the topics that people discuss on and forward to friends in their circle. Recognizing of new(emerging) topics is through conventional based approach.

In this paper, we propose a model (probability model), in which we consider total number of mentions for a particular post and the frequent appearance of the user for the same. An average count is evaluated using similar posts of hundreds of users. The post having the highest average count is extracted. Extraction of such new topic using the aggregate is done using change-point detection technique.

In the Fig 1, the link from Albert has been shared to his direct friend Jobs and that passes on to Bitto who is direct friend of Jobs and finally to Sam where Albert and Sam are not direct friends. Likewise, the links pass on to the entire network.

Fig 1 Link over the network

2.PROPOSED METHOD

2.1 Existing model

In the Fig 2, there exist three different friends circle say A, B, and C in which the facts and figures get circulated among them. Thus each of the circles is unaware of the emerging subject. Moreover, the non-textual information cannot be processed in this model.

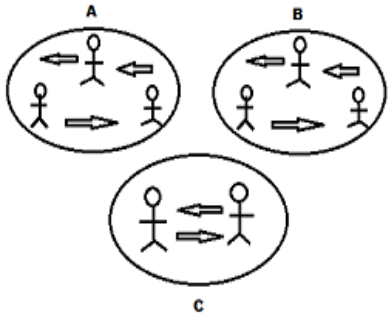

Fig 2 Existing Model

2.2 Proposed model

The Fig 3 shows the same friends circle of the existing model but the resource is being circulated among different circle who are not the direct friends of each other. Here the subject breaks the circle and it gets notified to all the users in the entire social stream. Apart from text other formats like images, videos, audios and animated sequences are extracted in the form of URL. The conventional-term-frequency is more appropriate.

This method is implemented using Kleinberg's burst detection method rather than using SDNML Change-point analysis technique. Here Kleinberg's burst detection model is using a two state version since there is no hierarchical structure.

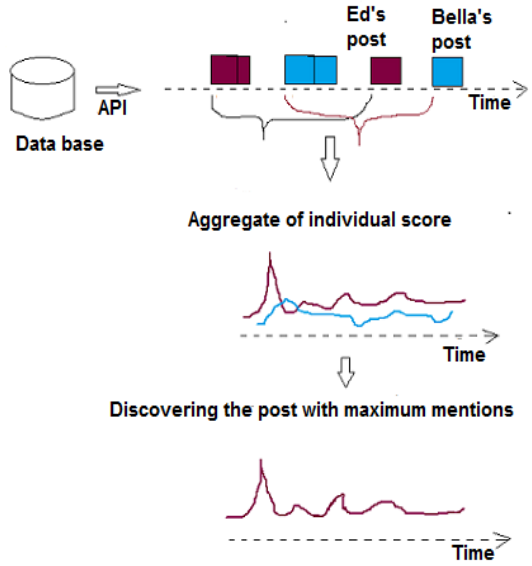

Fig 4 Overall flow of proposed method

2.2 Probability Model

Probability model here is used to identify the usual mentioning behavior of a user. The categorization of the post in social stream is through the number of mentions k (Post has) and a set v that holds the names of the mentioned. In particular, this probability model is concerned on setting a limit to a particular user (i.e. how many times a user can be a part in the post).

2.3 Computing the link-anomaly score

This technique is to describe how to calculate the difference in user's behavior from the usual behavior. This is done in order to find the score for the new post x= (t, u, k, V) where user u, time t, k mentions and users V (i.e. by user u to user V in time t having k mentions) and the probability in calculated with set T (t) u. The set is a collection of the posts made by the user u say in a time period [t-T, t].

2.4 Combining Anomaly Scores from Different Users

Anomaly detection is otherwise known as outlier detection.In this section we consider the aggregate of the scores that is computed from different users. The score for each user is found depending on the user u current post and the past behaviour T (t) u of the user. To identify the general behaviour of the user the average scores are obtained for posts(x1.....x) of the user.

2.5 Burst Detection Method

The time interval between the mentions may differ from place to place. This method isto consolidate the time difference in order to maintain the correctdetectionon the scores of the mentions for the particular post.

CHANGE-POINTDETECTION ALGORITHM

Change point detection algorithm alerts with an alarm as an evidence when it attains a threshold. This algorithm can be done both online or offline. When the scenario is online the change point is detected immediately once a change has occurred, and when the scenario is offline the goal is to compute the change-point in time sequence.

From the calculated aggregate the change point is discovered. Extension of Change Finder is Change point detection. Here the structure of time series is monitored. Detection of change point is done through two layers of scoring processes. Layer one is to find out the outliers whereas the second layer is to find out the change points. As a criteria for scoring autoregressive model is used. For certain applications it is necessary to calculate location or time of the change-point at some level of confidence.

This algorithm uses the fundamental idea of Bayesian statistics to calculate predictive distribution.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

here H stands for hypothesis , E is for evidence, P(H|E) refers to the probability of hypothesis for a given evidence and is called as the Posterior probability, P(H) is the prior probability, P(E|H) is called as the likelihood, P(E) is called the normalizing constant. This detection algorithm helps in improving the accuracy of the predictions. It identifies the probability of the change point in the recently occurring mention.

KLEINBERG'S BURST DETECTION ALGORTHIM

To discover that any two or more strangers are connected with each other through some chain network on basis of common information. A two state version on Kleinberg's burst detection model is used because a hierarchical structure is not expected. The algorithm can occur in two states, the first is the burst state and the second is non-burst state.

Considering two strangers where source person is to be Adam and the target person is Dave. The information forwarded is known only through first-name basis. This process can be proceeded using local knowledge of links

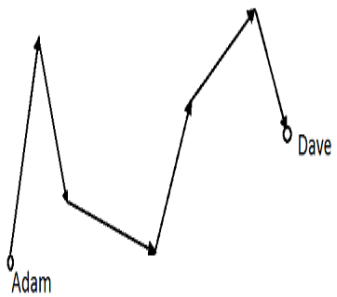

Fig 5 Link between two strangers

3. CONCLUSION

In this paper, we have considered an approach to unfold and pop out the emerging topics that are at the importance of discussion. The mentioning behaviour of the users is taken into account on basis of the number of mentions and the appearance of the mentioned for a particular post. Change-point detection algorithm is used in

order to identify the emerging topic. The proposed methods rely on the URL of the post regardless of the format.

4. FUTURE WORK

As an extension to this paper, future work will include certain methodologies, Firstly to remove noisy data both in mentions and the posts. Secondly, to provide a privatization to the users mentioning the post by wrapping their name.

REFERENCES

[1] Toshimitsu TakahashI, Ryota Tomioka and Kenji Yamanishi,"Discovering Emerging Topics in Social Streams via Link-Anomaly Detection", 2014.
[2] A.Krause, J.Leskovec and C.Guestrin,"Data Association for Topic Intensity Tracking", 2006.
[3] C.Giurcaneanu and S.Razavi,"AR Order Selection in the Case When the Model Parameters are Estimated by Forgetting Factor Least-Squares Algorithms", 2010.
[4] D.He and D.S Parker,"An Alternative Model of 'Bursts' in Streams of Topics",2010.
[5] Q.Mei and C.Zhai, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining",2005.
[6] K.Yamanashi and Y.Maruyama,"Dynamic Syslog Mining for Network Failure Monitoring", 2005
[7] S.Morinaga and K. Yamanishi, "Tracking Dynamics of TopicTrends Using a Finite Mixture Model",2004.
[8]Y.Teh, M.Jordan, M.Beal and D.Blei, "Hierarchical DirichletProcess",2006.
[9] S.Sujitha and S.selvi, "A Model of Textual Emotion Mining From Text Document", 2014.
[10] Neil Shah, Alex Beute, Brian Gallagher and Christos Faloutsos, "Spotting Suspicious Link Behaviour with fBox: An Adversarial Perspective", 2014.
[11] S.Saranya, R.Rajeshkumar and S.Shanthi, "A Survey on Anomaly Detection for Discovering Emerging Topics", 2014.
[12] Chandan MG and Chandra Naik, "Survey on Link Anomaly Detection for Textual Stream in Online Social Network", 2014.
[13] Bimal Viswanath and M.Ahmad Bashir, "Towards Detecting Anomalous User Behaviour in Online Social Networks", 2013.
[14] Robert Dale, Sabine Geldof and Jean Philippe Prost, "Using Natural Language Generation in Automatic Route Description", 2012.
[15] A.A.Sattikar and R.V.Kulkarni, "Natural Language Processing for Content Analysis in Social Networking", 2012.
[16] So Hirai and Kenji Yamanishi, "Normalized Maximum Likelihood Coding for Exponential Family with its Applications to Optimal Clustering", 2012.
[17] J.Rissanen,T.Roos,and P.Myllymaki, "Model Selection by Sequentially Normalized Least Squares",2009
[18] Lin Li, Anna Scaglione, Ananthram Swami and Qing Zhao, "Phase Transition in Opinion Diffusion in Social Networks", 2012.
[19] Gabriel Weimann, "Terrorists Using Online Social Networking", 2012.
[20] Chong Wang and John Paisley, "Online Variational Inference for the Hierarchial Dirichlet Process", 2010.
[21] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", 2004.
[22] Russell Swan and James Allan, "Automatic Generartion of Overview Timelines", 2000.
[11] A.Trouve and Y.Yu, "Unsupervised clustering trees by nonlinear principal component analysis," Pattern Recognit. Image Anal., vol. 2, pp. 108-112, 2001.
[12] D. T. Larose, Data Mining Methods and Models. Hoboken, NJ, USA: Wiley, 2006.
[13] D. Arthur, B. Manthey, and H. Roglin, "Smoothed analysis of the k-means method," J.ACM, vol. 58, no. 5,pp. 19:1-19:31, Oct. 2011[Online]. Available: http://doi.acm.org/10.1145/2027217.
[14] C. Sorg et al., "Increased intrinsic brain activity in the striatum reflects symptom dimensions in schizophrenia," Schizophr Bill., vol.39, no. 2, pp. 413-421, NOV.2007.
[15] M. Halkidi, Y.Batistakis, and M. Vazirgiannis, "On clustering validation techniques," J. Intell. Inf. Syst., vol. 17, no. 2-3, pp. 107-145, 2001.
[16] B. Scholkopf, A. J. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput., vol. 10, no.5, pp. 1299-1319, 1998.