

DISCOVER EMERGING TOPICS IN SOCIAL NETWORKING

By

A.HarithaDeepthi, V.Aravindkarthick, P.S.Clinton

harithadeepthibtech@gmail.com, aravindkarthick7@gmail.com, clintonshaju7@gmail.com

ABSTRACT

Web Mining is the retrieval of useful information from web. The information exchanged in social network posts includes not only text but also images, URLs, and videos. This system focuses on mentions of users – links between users that are generated through replies, mentions, and retweets. The probability model is proposed to detect the emergence of a new topic from the anomalies measured through the model. The technique demonstrated on trained dataset using web application. Aggregating anomaly scores from hundreds of users, shows that detecting emerging topics only based on the reply/mention relationships in social network posts. The experiments show that the proposed mention-anomaly-based approaches can detect new topics at least as early as text-anomaly-based approaches.

1. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Detection of emerging topics is now receiving renewed interest motivated by the rapid growth of social networks. The information exchanged via social network posts include not only text but also images, URLs, and videos. This system focuses on emergence of topics signaled by social aspects of networks by considering mentions of users and links between users that are generated dynamically to find the topics

2. DETECTION OF EMERGING TOPICS

Communication over social networks, such as Facebook and Twitter, is gaining its importance in our daily life. Since the information exchanged over social networks are not only texts but also URLs, images, and videos, they are challenging test beds for the study of data mining. In particular, we are interested in the problem of detecting emerging topics from social streams, which can be used to create automated “breaking news”, or discover hidden market needs or underground political movements. Compared to conventional media, social media are able to capture the earliest, unedited voice of ordinary people. Therefore, the challenge is to detect the emergence of a topic as early as possible at a moderate number of false positives. Another difference that makes social media social is the existence of mentions. Mentions links to other users of the same social network in the form of message-to, reply-to, retweet-of, or explicitly in the text. One post may contains a number of mentions. Some users may include mentions in their posts rarely other users may be mentioning their friends all the time. Some users (like celebrities) may receive

mentions every minute, for others, being mentioned might be a rare occasion. In this sense, mention is like a language with the number of words equal to the number of users in a social network. The main aim of this project is to detect emerging topics as early as the keyword based methods and evaluate whether the proposed approach can detect the emergence of the topics recognized and collected by people.

3. TEXT CLASSIFICATION ALGORITHMS

The problem of classification has been widely studied in the database, data mining, and information retrieval communities. The problem of classification is defined as follows. We have a set of training records as follow $D = \{X_1, \dots, X_N\}$, such that each record is labeled with a class value drawn from a set of k different discrete values indexed by $\{1 \dots k\}$. The training data is used in order to construct a classification model, which relates the features in the underlying record to one of the class labels. For a given test instance for which the class is unknown, the training model is used to predict a class label for this instance. In the hard the classification problem, a particular label is explicitly assigned to the instance, whereas in the soft version of the classification problem, a probability value is assigned to the test instance. Other variations of the classification problem allow ranking of different class choices for a test instance, or allow the assignment of multiple labels to a test instance.

4. RELATED WORK

The aim of this project is to detect the emerging topics receiving renewed interest motivated by the rapid growth of social network. Conventional term frequently based approaches may not be appropriated in the context, because the information exchanged in social network posts includes not only text but also images, URL's and videos etc...focused on emergence of topic signaled by social aspects networks. Specifically, focused on mention of users links between users that are generated dynamically through replies, mentions and retweets. It propose a probability model of the mentioning behavior of a social network user, and propose to detect the emergence of a new topic from the anomalies measured through the model.

The overview of the project is to aggregating anomaly scores from hundreds of users, It show that it can detect emerging topics only based on the reply mention relationship in social-network posts. In this project it uses a technique called Text Classification Algorithm in several real data sets. This project shows that the proposal mention anomaly based approaches can detect new topics at least as early as text anomaly based approaches and in some case much earlier when the topic is poorly identified by the textual contents in posts.

5. PROBABILITY MODEL

Probability model used to capture the normal mentioning behavior of a user and describes how to train the model. Also characterized a post in a social network stream by the number of mentions k it contains, and the set V of names (IDs) of the mentioned (users who are mentioned in the post). There are two types of infinity taken into account here. The first is the number k of users mentioned in a post. Although, in practice a user cannot mention hundreds of other users in a post, this system would like to avoid putting an artificial limit on the number of users mentioned in a post. Instead, assume a geometric distribution and integrate out the parameter to avoid even an implicit

A Computing the link-anomaly score

This subsection describes how to compute the deviation of a user's behavior from the normal mentioning behavior modeled in order to compute the anomaly score of a new post $x = (t, u, k, V)$ by user u at time t containing k mentions to users V , and compute the probability with the training set $T(t, u)$, which is the collection of posts by user u in the time period $[t-T, t]$ (we use $T = 30$ days in this paper). Accordingly the link-anomaly score is defined $ass(x) = -\log(P(k|T_u^{(t)}) \prod_{v \in V} P(v|T_u^{(t)}))$
 $s(x) = -\log P(k|T_u^{(t)}) - \sum_{v \in V} \log P(v|T_u^{(t)}) - 1$

The two terms in the above equation can be computed via the predictive distribution of the number of mentions, and the predictive distribution of the mention.

B Combining Anomaly Scores from Different Users

This section describes how to combine the anomaly scores from different users. The anomaly score is computed for each user depending on the current post of user u and his/her past behaviour $T(t, u)$. In order to measure the general trend of user behaviour, aggregated anomaly scores obtained for posts $x_1, \dots, x_n$ using a discretization of window size $\tau > 0$ as follows:

$$S'_j = \frac{1}{\tau} \sum_{t_i \in [T(j-1), T(j)]} S(x_i) - 2$$

Where $x_i = (t_i, u_i, k_i, V_i)$ is the post at time t_i by user u_i including k_i mentions to users V_i .

C Burst detection method

In addition to the change-point detection based on SDNML followed by DTO was described, also test the combination of our method with Kleinberg's burst detection method. More specifically, implemented a two-state version of Kleinberg's burst detection model. The reason to choose the two-state version was because in this experiment expect no hierarchical structure. The burst detection method is based on a probabilistic automaton model with two states, burst state and nonburst state. Some events (e.g., arrival of posts) are assumed to happen according to a time varying Poisson processes whose rate parameter depends on the current state.

D Scalability of the proposed algorithm

The proposed link-anomaly based change-point detection is highly scalable. Every step described in the previous subsections requires only linear time against the length of the analyzed time period. Computation of the predictive distribution for the

number of mentions can be computed in linear time against the number of mentions. Computation of the predictive distribution for the mention probability in and can be efficiently performed using a hash table. Aggregation of the anomaly scores from different users takes linear time against the number of users, which could be a computational bottle neck but can be easily parallelized. SDNML-based change-point detection effects as change of day light or moving shadows.

VI DATA FLOW DIAGRAM

LEVEL 0

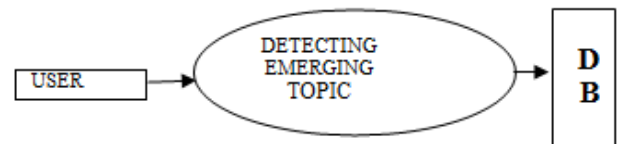


Fig:1 User defining the topics

The above figure explains the user detecting the topics from the database.

LEVEL 1

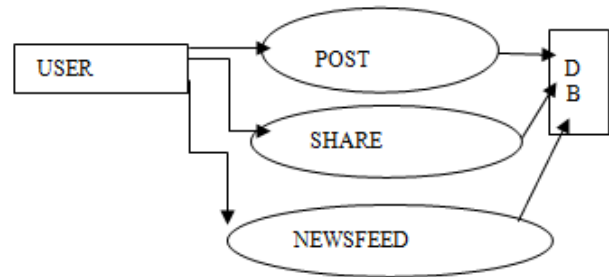


Fig2: User sharing and posting

The figure explains that the user logged in can post a newsfeed and can share the newsfeed the user has posted, he also can share the newsfeed that other users have posted in the web application. The posts and the newsfeed that has been posted by the users in the web application are stored in the database.

6. SYSTEM ARCHITECTURE DESIGN

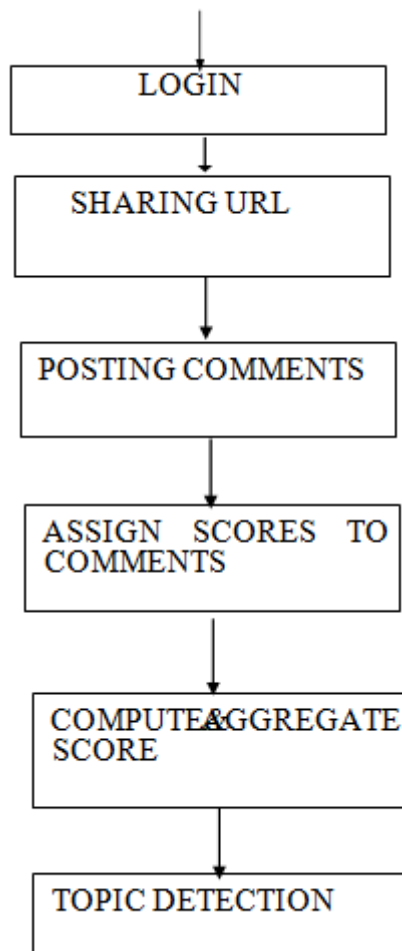
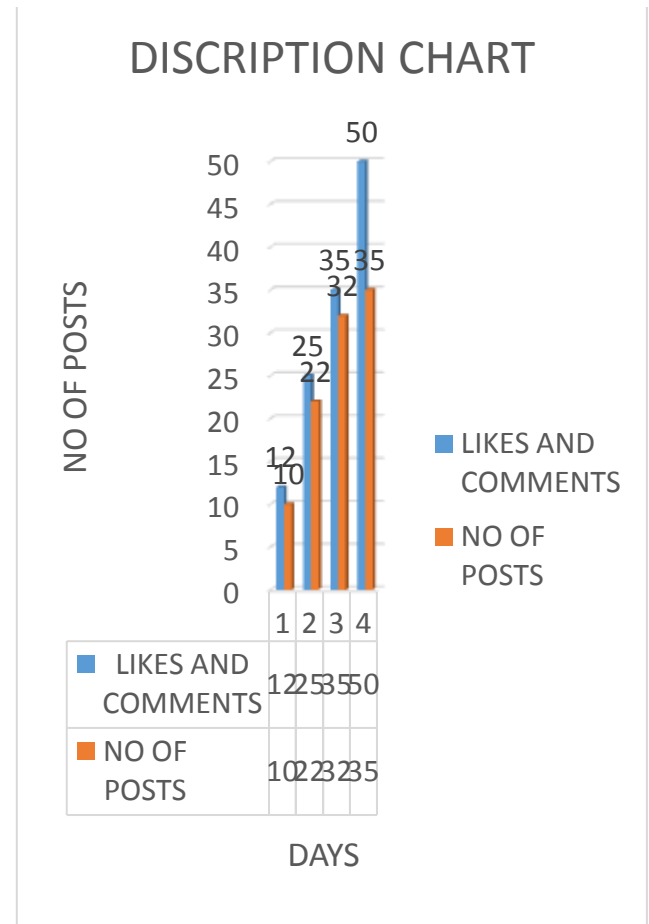


Fig:3 Emerging N



8. CONCLUSION

A new approach was proposed to the emergence of topics in a social network stream. The basic idea of approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. The existing probability model that captures both the number of mentions per post and the frequency of mentioned. The existing method does not rely on the textual contents of social network posts, it is robust to rephrasing and it can be applied to the case where topics are concerned with information other than texts, such as images, video, audio, and so on. All the analysis presented in this paper was conducted offline, but the framework itself can be applied online by implementing the web application on a created online website. It would also be interesting to combine the proposed link-anomaly model with text-based approaches, because the proposed link-anomaly model does not immediately tell what the anomaly is. Combination of the word-based approach with the link anomaly model would benefit both from the performance of the mention model and the intuitiveness of the word-based approach.

9. REFERENCE

- [1] Toshimitsu Takahashi, RyotaTomioka, and Kenji Yamanishi, Member, IEEE "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection" IEEE Transaction on Knowledge and data engineering Vol. 26, NO. 1, Jan 2014.
- [2] Allan.J et al., "Topic Detection and Tracking Pilot Study: Final Report," Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

7. DESCRIPTIONS

Step 1: The mentions are extracted from a social network service in a sequential manner through some API. For each new post, use samples within the past time interval of length T for the corresponding user for training them.

Step 2: Assign an anomaly score to each post based on the learned probability distribution.

Step 3: The score is then aggregated over users.

Step 4: Further it fed into SDNML-based change point analysis.

Step 5: The topics has been detected based on the aggregated value.

8. OUTPUT DISCRIPTION CHART

- [3] Aldous .D, "Exchangeability and Related Topics," Ecole d'Ete' de Probabilities' de Saint-Flour XIII—1983, pp. 1-198, Springer, 1985.
- [4] Chandola.V, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 15:1- 15:58, 2009.
- [5] Urabe. Y, K. Yamanishi, R. Tomioka, and H. Iwai, "Real-Time Change-Point Detection Using Sequentially Discounting Normal- ized Maximum Likelihood Coding," Proc. 15th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD' 11), 2011.
- [6] A. 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.