www.ijiarec.com

ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Determining frequent data in social streams using link-anomaly detection

Dr K.Chandraprabha[1], S.Santhosh Kumar[2],M.Srinivasan[3,]V.Vanitha[4.]

[1]Professor Dept of CSE,K.S.Rangasamy College of Technology,Tiruchengode,Tamilnadu,India.

[2,3,4]StudentsDept of CSE,K.S.Rangasamy College of Technology,Tiruchengode,Tamilnadu,India.

chandraprabhak@ksrct.ac.in, vanithavelu1995@gmail.com, santeee27@gmail.com, srini1954@gmail.com

Abstract-Detection of frequent data is now receiving renewed interest motivated by the rapid growth of social networks.Conventional-term-frequency-based approaches may not be appropriate in this context, because the information exchanged in social network posts include not only text but also images, URLs, and videos. This system focus on frequent data signalled by social aspects of these networks. Specifically, this system focus on mentions of users links between users that are generated dynamically (intentionally orunintentionally) through replies, mentions, and retweets. Propose a probability model of the mentioning behaviour of a socialnetwork user, and propose to detect the frequent of a data from the anomalies measured through the model. Aggregatinganomaly scores from hundreds of users, shows that can detect frequent data only based on the Scalability of the proposed algorithm, in social-network posts. Demonstrate our technique in several real data sets that gathered from Twitter. The experiments show thatthe proposed mention-anomaly-based approaches can detect new topics at least as early as text-anomaly-based approaches, and insome cases much earlier when the topic is poorly identified by the textual contents in posts.

Key terms-Topic detection, anomaly detection, social networks, sequentially normalized maximum likelihood coding, burst detection, Dynamic threshold optimization.

INTRODUCTION

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. Here, we mean by mentions links to other users of the same social network in the form of message-to, reply-to,retweets-of, or explicitly in the text. One post may contain 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. We are interested in detecting emerging topics from social network streams based on monitoring the mentioningbehavior of users. Our basic assumption is that a new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words. A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. Moreover, it cannot be applied when the contents of the messages are mostly nontextual information. On the other hand, the "words" formed by mentions are unique, require little preprocessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents. Fig. 1 shows an example of the emergence of a topic through posts on social networks. The first post by Bob contains mentions to Alice and John, which are both probably friends of Bob, so there is nothing unusual here. The second post by John is a reply to Bob but it is also visible to many friends of John that are not direct friends of Bob. Then in the third post, Dave, one of John's friends, forwards (called retweets in Twitter) the information further down to his own friends. It is worth mentioning that it is not clear what the topic of this conversation is about from the textual information, because they are talking about something (a new gadget, car, or jewellery) that is shown as a link in the text. In this paper, we propose a probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Then this model is used to measure the anomaly of future user behavior. Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behavior of the user. We aggregate the anomaly scores obtained in this way overhundreds of users and apply a recently proposed change point detection technique based on the sequentially discounting normalized maximum-likelihood (SDNML) coding. This technique can detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and pinpoint where the topic emergence. The effectiveness of the proposed approach is demonstrated on four data sets we have collected from Twitter. We show that our mention-anomaly-based approaches can detect the emergence of a new topic at least as fast as text-anomaly-based counterparts. Furthermore, we show that in three out of four data sets, the proposed mention-anomaly-based methods can detect the emergence of topics much earlier than the text-anomaly-based methods, which can be explained by the keyword ambiguity we mentioned above.

RELATED WORKS

Detection and tracking of topics have been studied extensively in the area of topic detection and tracking (TDT). In this context, the main task is to either classify a new document into one of the known topics (tracking) or to detect that it belongs to none of the known categories. Subsequently, temporal structure of topics has been modeled and analyzed through dynamic model selection, temporal text mining, and factorial hidden Markov model. Another line of research is concerned with formalizing the notion of "bursts" in a stream of documents. In his seminal paper, Kleinberg modeled bursts using the time- varying Poisson process with a hidden discrete process that controls the firing rate. Recently, He and Parker developed a physics-inspired model of bursts based on the change in the momentum of topic. All the above-mentioned studies make use of textual content of the documents, but not the social content of the documents. The social content (links) has been utilized in the study of citation networks. However, citation networks are often analyzed in a stationary setting. The novelty of the current paper lies in focusing on the social content of the documents (posts) and in combining this with a change-point analysis.

PROPOSED METHOD

In this project, to proposed a new

approach to detect the frequent data in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. They have proposed a probability model that captures both the number of mentions per post and the frequency of mentioned. Assume that the data arrives from a social network service in a sequential manner through some API. For each new post we use samples within the past T time interval for the corresponding user for training the mention model we propose below. Assign anomaly score to each post based on the learned probability distribution. The score is then aggregated over users and further fed into a change point analysis.

Advantages:

- ✓ The proposed 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 data's are concerned with information other than texts, such as images, video, audio, and so on.
- ✓ The proposed link-anomaly-based methods performed even better than the keyword-based methods on "NASA" and "BBC" data sets.'
- ✓ High in accuracy.
- ✓ Minimum computation time

MODULE DESCRIPTION

Data preprocessing:

In this module, we preprocess the probability model that we used to capture the normal mentioning behavior of a user and how to train the model. We characterize 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 we have to take 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, we would like to avoid putting an artificial limit on the number of users mentioned in a post. Instead, we will assume a geometric distribution and integrate out the parameter to avoid even an implicit limitation through the parameter. The second type of infinity is the number of users one can possibly mention. To avoid limiting the number of possible mentioned, we use Chinese Restaurant Process (CRP) based estimation; who use CRP for infinite vocabulary.

Computing the link-anomaly score:

In this module, we describe 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 that is $\boldsymbol{x} = (t,\ u,\ k,\ V\ )$ by user u at time t containing k mentions to users V, we 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 project). Accordingly the link-anomaly score is defined .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 mentioned.

Change point analysis and DTO:

This technique is an extension of Change Finder proposed, that detects a change in the statistical dependence structure of a time series by monitoring the compressibility of a new piece of data. This module is to used a sequential version of normalized maximum-likelihood (NML) coding called SDNML coding as a coding criterion instead of the plug-in predictive distribution used. Specifically, a change point is detected through two layers of scoring processes. The first layer detects outliers and the second layer detects change-points.

In each layer, predictive loss based on the SDNML coding distribution for an autoregressive (AR) model is used as a criterion for scoring. Although the NML code length is known to be optimal, it is often hard to compute. The SNML proposed is an approximation to the NML code length that can be computed in a sequential manner. The SDNML proposed further employs discounting in the learning of the AR models.As a final step in our method, we need to convert the change-point scores into binary alarms by thresholding. Since the distribution of change-point scores may change over time, we need to dynamically adjust the threshold to analyze a sequence over a long period of time. In this subsection, we describe how to dynamically optimize the threshold using the method of dynamic threshold optimization proposed. In DTO, we use a one-dimensional histogram for the representation of the score distribution. We learn it in a sequential and discounting way.

Burst detection method:

In this module that the change-point detection based on SDNML followed by DTO described in previous sections, we also test the combination of our method with Kleinberg's burst-detection method. More specifically, we implemented a two-state version of Kleinberg's burst-detection model. The reason we chose the two-state version was because in this experiment we expect non-hierarchical structure. The burst-detection method is based on a probabilistic automaton model with two states, burst state and non-burst state.

CONCLUSION

In this project, these systems have proposed a new approach to detect the frequent data in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. This system have proposed a probability model that captures both the number of mentions per post and the frequency of mentioned.Thus the planning to scale up the proposed approach is, to handle social streams in real time. 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

REFERENCES

[1]. Ivax Fellegi.P and Alan sundar.B et al, "A Theory For Record Linkage" *conf. linkage data mining,2013.*

[2] .Bin Yao,Feifei Li et al, "Approximate String Search In Spatial Databases" *conf. Data Mining Knowledge Discovery,2012.*

[3]. Dingming Wu ,Byron Choi et al, "Authentication Of Moving Top-K Spatial Keyword Queries "*ConfKnowlegde Discovery in Data Mining*, pp. 812-821, 2012.

[4]. Kleinberg.J, "Bursty and Hierarchical Structure in Streams," *Data Mining Knowledge Discovery,* Vol. 7, No. 4, pp. 373-397, 2012.

[5]. Parag Singla, Pedro Domingos et al, "Collective Object Identification" *Knowledge Discovery and Data Mining*, 2014.

[6]. Mei.Q and Zhai.C, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining,"*11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining,* pp. 198-207, 2013.

[7]. Krause.A, Leskovec.J, and Guestrin, "Data Association for Topic Intensity Tracking,"*23rd Int'l Conf. Machine Learning (ICML' 06),* pp. 497-504, 2011.

[8]. Yamanishi.K and Maruyama.Y, "Dynamic Syslog Mining for Network Failure Monitoring," *Int'l Conf. Knowledge Discovery in Data Mining,* pp. 499-508, 2013.

[9]. Ahmed k Elmagarmid, Pangiotis et al, "Duplicate Record Detection" *Discovering topics in Data Mining,2012.*

[10]. Rohit Ananthakrishna, Surajit et al, "Eliminating Fuzzy Duplicates In Data Warehouses" *Knowledge Discovery and Data Warehousing, 2013.*