Expert Systems With Applications 57 (2016) 285–295

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks

Qi Dang[a], Feng Gao[b], Yadong Zhou[a,*]

[a] Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, P.R. China
[b] Institute of Systems Engineering, Xi'an Jiaotong University, P.R.China

ARTICLE INFO

Article history:
Received 13 February 2015
Revised 29 March 2016
Accepted 30 March 2016
Available online 1 April 2016

Keywords:
Micro-blogging networks
Emerging topics
Early detection
DBNs

ABSTRACT

Micro-blogging networks have become the most influential online social networks in recent years, more and more people are used to obtain and diffuse information in them. Detecting topics from a great number of tweets in micro-blogging is important for information propagation and business marketing, especially detecting emerging topics in the early period could strongly support these real-time intelligent systems, such as real-time recommendation, ad-targeting, marketing strategy. However, most of previous researches are useful to detect emerging topic on a large scale, but they are not so effective for the early detection due to less informative properties in a relatively small size. To solve this problem, we propose a new early detection method for emerging topics based on Dynamic Bayesian Networks in micro-blogging networks. We first analyze the topic diffusion process and find two main characteristics of emerging topic which are *attractiveness* and *key-node*. Then based on this finding, we select features from the topology properties of topic diffusion, and build a DBN-based model by the conditional dependencies between features to identify the emerging keywords. An emerging keyword not only occurs in a given time period with frequency properties, but also diffuses with specific topology properties. Finally, we cluster the emerging keywords into emerging topics by the co-occurrence relations between keywords. Based on the real data of Sina micro-blogging, the experimental results demonstrate that our method is effective and capable of detecting the emerging topics one to two hours earlier than the other methods.

1. Introduction

With the rapid development of social network, micro-blogging networks (i.e. Twitter, Sina micro-blogging) have been the most important way to obtain information. In micro-blogging networks, people can report their current views and thoughts, comment on the breaking news and events, share the interesting messages. Through these online behaviors, plentiful and valuable information diffuses in the ways of tweeting, retweeting, and commenting to form various topics. Topic is defined as something that happens at a specific time and place, along with all the necessary preconditions and unavoidable consequences in Cieri (2000). Detecting topics from a great number of tweet contents is important for information propagation and business marketing.

Emerging topic usually refers to the content that can attract tremendous attentions in a short period, and the related discussions influence public opinions much more significantly than they do on common topics (Zhou, Guan, Zheng, Sun, & Zhao, 2010). In general, emerging topics are concerned with content of influential emerging events, such as traffic accident, natural disaster, election campaign, and regulation enforcement (Chen, Luesukprasert, & Chou, 2007). Due to the importance and burst of these emerging topics, people expect to know the emerging topics as early as possible to design crisis control strategies, discover business opportunities, and find important information. The early detection could also support the real-time intelligent systems strongly, such as real-time recommendation, ad-targeting, marketing strategy. However, few people are aware of emerging topics before they attract a large number of users in the present micro-blogging networks.

In the related work of topic detection, many researchers have obtained a lot of achievements. Most of them utilize the keywords based approach and the extensions with traditional features, including term frequency, term distribution, and time feature (Bun and Ishizuka, 2002; Blei, Ng, & Jordan, 2003). These methods are useful for detecting topics from a fixed corpus with the whole timeline, or detecting emerging topics when the number of joined users increases into a large scale. Nevertheless, they are less effective for detecting emerging topic in the early period timely.

* Corresponding author. Tel: +8618509231436.
E-mail addresses: qdang@sei.xjtu.edu.cn (Q. Dang), fgao@sei.xjtu.edu.cn (F. Gao), ydzhou@xjtu.edu.cn (Y. Zhou).

http://dx.doi.org/10.1016/j.eswa.2016.03.050
0957-4174/

In recent years, a few researchers explore the early detection of emerging topics with aging theory (Cataldi, Di Caro, & Schifanella, 2010; Yu, Zhao, Chang, & He, 2014), dynamic model (Du, Wu, He, & Liu, 2012), and latent source signals (Nikolov & Shah 2012). These works propose several preliminary methods and provide useful knowledge for our work, but there are still two challenges as follows:

- In the early period of topic diffusion, the differences between emerging topics and non-emerging topics are inconspicuous to quantify. The burst features of previous researches are commonly extracted by the temporal evolution based on term frequency, which are less effective for the early detection.
- In the diffusion of emerging topic, the interval between the appearance time and peak time is very short, half of the retweets occur within an hour of the source tweet (Kwak, Lee, Park, & Moon, 2010). To detect these short-term topics, timeliness is a primary factor to be considered.

In this paper, we propose an early detection method for the emerging topics based on Dynamic Bayesian Networks (Murphy, 2002). First, we analyze the characteristics of topic diffusion in the early period, and several topology features are selected by comparison analysis between emerging topics and non-emerging topics. Then we propose a Dynamic Bayesian Network (DBN) model for emerging keyword detection. We initially create a term list of emerging keyword candidates by term frequency in a given time interval. For each candidate, we build a DBN-based model by the joint conditional probabilities between the selected features, and calculate the probability of a candidate being an emerging keyword. The learning and interface of DBNs are implemented by EM algorithm (Murphy, 2002) and Viterbi Algorithm (Forney, 2005) respectively. The emerging keywords are obtained by ranking their probabilities in each time interval. Finally, we calculate the co-occurrence relations between keywords and cluster emerging keywords into emerging topics. The framework of our method is shown in Fig. 1.

The main contributions of this paper are the following:

- We propose a new method for detecting emerging topics during the early period in micro-blogging networks. Different from earlier work, we select features from topology properties of topic diffusion and detect emerging trends by dynamic changes of conditional probabilities calculated by DBNs.
- We find two characteristics of emerging topic which are *attractiveness* and *key-node*, and analyze their dependencies with the features selected from the retweeting network and the following network.
- We build a new DBN-based model to represent the temporal evolution of keyword. The model can discover emerging keywords by calculating the probability of a keyword being an emerging one.

The main structure of the content is organized as follows: Section 2 provides an overview of related works. Section 3 induces the dataset and labeling. Section 4 presents our method of emerging topic detection by DBN-based model. The result analysis and comparison experiment are introduced in Section 5. We conclude this work and expose the future work in Section 6.

2. Related work

There are a great deal of researches for Topic Detection and Tracking (Allan, Carbonell, Doddington, Yamron, & Yang, 1998; Allan, 2002), which mainly have two direction: topic detection and topic tracking. Topic detection aims at detecting emerging topics from text data, and topic tracking focusses on tracking the evolutions of topics over time series. In recent years, topic detection has been applied to many extensive applications, such as detecting large scale events like earthquakes (Sakaki, Okazaki, & Matsuo, 2010), predicting political election outcomes (Tumasjan, Sprenger, Sandner, & Welpe, 2010), recommending interesting topic or URL for user (Balabanović & Shoham, 1997; Hassan, Radev, Cho, & Joshi, 2009; Chen, Nairn, Nelson, Bernstein, & Chi, 2010), finding controversial topics (Popescu and Pennacchiotti, 2010), extracting meaningful topics by filtering the hijacked topics (Hayashi, Maehara, Toyoda, & Kawarabayashi, 2015), quantifying the impact of a topic during a given period (Bernabé-Moreno, Tejeda-Lorente, Porcel, & Herrera-Viedma, 2015). Specifically, topic detection can be classified into four categories by their main algorithms as follows.

Keyword-based approaches: Many topic detection approaches have been developed by measuring the importance/burst of keywords, and identifying the topic by the co-occurrence relations between keywords. Bun and Ishizuka (2002) presented the TF∗PDF algorithm which extends the well-known TF-IDF algorithm. Kotov, Zhai, & Sproat (2011) mined the named entities with temporally correlated bursts from multilingual web news streams. Wu, Ding, Wang, & Xu (2010) used the tolerance rough set model to enrich the set of feature words into an approximated latent semantic space from which they extracted hot topics by a complete-link clustering. Thelwall, Buckley, & Paltoglou (2011) combined sentiment analysis methods to detect burst events. Du, Wu, He, & Liu (2012) extracted burst feature by computing the term frequency and tweet weigh in a given time interval. By the semantic information between a term and its meaning, Vicient & Moreno (2015) applied a semantic similarity measure to group related terms into new topics. Yang et al. (2015) introduced a hot topic detection method combining bursty term identification and multi-dimension sentence modeling to automatically detect emerging topics for rumor identification.

Probabilistic topic models: A number of probabilistic topic models have been investigated, such as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan 2003) and probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 2001). Many variants on the basis of LDA and pLSA were proposed for dynamic topic modeling (Blei & Lafferty, 2006). Wang, Agichtein, & Benzi (2012) proposed TM-LDA to estimate transition probabilities of topics and predicted future topics from past observations. Chen & Liu (2014) proposed a topic model AMC which could mine prior knowledge from the past results for the future modeling. Based on LDA and matrix factorization, Kim & Shim (2014) introduced a recommendation system to recommend top-K users and tweets. Kim, Choo, Reddy, & Park (2015) presented a topic model based on joint nonnegative matrix factorization to identify common and discriminative topics. Yuan et al. (2015) implemented the lightLDA which enable very large data sizes and models to be processed on a small computer cluster. Hu, Sun, & Li (2015) proposed a novel approach to capture both strength and content evolution simulaneously via On-Line LDA.

Aging theory: Aging Theory was first presented in information retrieval by Chen et al. (2003) based on a biological metaphor. Wang, Zhang, Ru, & Ma (2008) ranked topics from news through the concept of *burstiness*. Cataldi, Di Caro, & Schifanella (2010) used timelines to represent temporal documents, transforming the issue into a topic visualization problem that can be solved by assessing the birth and death of topics. Chen, Amiri, Li, & Chua (2013) proposed a topic detection technique that permit to retrieve the most emerging topics expressed by the community in real-time. Yu, Zhao, Chang, & He (2014) adopted aging theory to build the life cycle model of events, and detected topics by ranking the hotness of topic. Bao, Xu, Min, & Hossain (2015) provided a method on emerging topic detection and elaboration using multimedia streams cross different online platforms, which are microblogging, news portal, and imaging sharing platforms.