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.

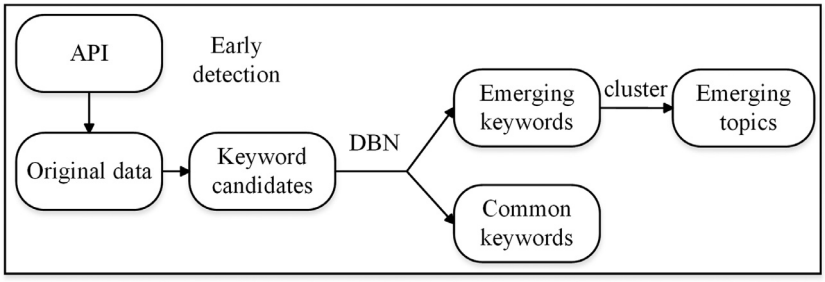

Fig. 1. The overview of our framework.

Graph-based approaches: There are several recent works detecting topics with the graph and graph analytical methods. Ohsawa, Benson, & Yachida (1998) used a *KeyGraph* algorithm to convert terms into a graph based on co-occurrence relations, and employed a community detection approach to partition the graph. Wang & Ohsawa (2013) proposed a systematic approach to turning data into effective human insights. Chen & Neill (2014) considered all potentially useful information in a unified nonparametric statistical framework, and facilitated the early detection of social events. Glavaš & Šnajder (2014) combined machine learning and rule-based models to extract sentence-level event mentions, and built the event graphs for information retrieval and multi-documents. *IdeaGraph* (Zhang, Wang, & Xu 2016) fused semantic relations and co-occurrence relations into a term graph and detected topics from the graph using a graph analytical method.

The majority of previous researches aim to detect topics from a fixed corpus during a long time period, and informative features can be properly selected from the whole timeline for different algorithms. Due to less consideration of timeliness, emerging topic can be detected by many approaches when it reaches a large scale. But early detection is hard to implement due to the poor information in a relatively small scale. There is only a few work on the early detection of emerging topic. Nikolov & Shah (2012) developed a new algorithm that can predict which topics will be trending, with a speed as fast as in average one and a half hours earlier than the algorithm of Twitter. Our work shares the same opinion that emerging topics have latent source signals in the early period. Compared with recent work, the main differences of our method lie in the features selection and the approach of detecting emerging trends. These works commonly extract features based on term frequency (Yang et al. 2015), word distribution (Chen, Buntine, Ding, Xie, & Du, 2015), and geographical location (Chen & Neill, 2014; Unankard, Li, & Sharaf, 2015). The emerging trends are identified by aging theory (Yu, Zhao, Chang, & He, 2014; Bao, Xu, Min, & Hossain, 2015), dynamic model (Du, Wu, He, & Liu, 2012), time window (Hu, Sun, & Li, 2015), and non-negative matrix factorization (Hayashi, Maehara, Toyoda, & Kawarabayashi, 2015). Different from these works, we select features from topology properties to represent the characteristics of topic diffusion, and detect the emerging trends by the dynamic changes of conditional probabilities calculated by DBNs.

3. Dataset collection

In this section, we introduce the experimental dataset which is collected from Sina micro-blogging, and then we present the detail of labeled data.

3.1. Dataset

The dataset is collected by APIs provided by Sina micro-blogging[1]. The dataset has 13, 973, 119 tweets which are posted by 69, 394 users from Sep. 1, 2012 to Oct. 31, 2012, including two types of data: 1) basic user properties, such as user's ID, name, gender, follower; 2) textual data, such as ID of tweet, post-time, tweet content containing the retweet tag '//@' which is used to recognize the retweeting relationship between users.

Table 1
Examples of emerging topics and non-emerging topics.

Topics	Start time	Type	Description
Collision	2012-10-02 07:00	Emerging	A ship collision accident in Hong Kong
Policy	2012-10-01 09:00	Emerging	A new policy that cars could pass the highways for free
Sunset	2012-10-01 16:00	Non-emerging	Discussion about a rarely and beautiful sunset
Teleplay	2012-10-05 15:00	Non-emerging	Discussion about television dramas on the air

3.2. Labeled data

There is no standard dataset for topic detection in Sina micro-blogging, so we manually label the emerging topics with the help of search engine technology. We use the keywords of topic as search words and limit the time period as the time period of topic. If the search result shows some relevant information about the topic, then the topic will be regarded as correct. During these correct topics, we label 54 emerging topics and 50 non-emerging topics during October 2012. Emerging topic refers to the topic which will develop into a large scale and last for hours. In contrast, non-emerging topic means the topic which could not last for a long time since their sudden appearances.

As the lack of space, we only detail two topics for each type in Table 1. The two emerging topics are contents about a ship collision accident in Hong Kong and the policy that cars could pass the highways for free during specific legal holiday. One non-emerging topic is the discussion about a rarely and beautiful sunset appearing suddenly, and another is the discussion about popular television dramas on the air.

4. Technical approach

4.1. Problem description

Emerging topic implies the topic which can cause tremendous attentions in a short time period and last for a long time. In Fig. 2, we present the evolution of an emerging topic, t_s is the appearance time when the topic is starting to burst, t_h is the time when topic becomes emerging near the peak time. The period from t_s to t_h is known as the emerging phase (Chen, Amiri, Li, & Chua, 2013). Emerging topic usually can be detected with a sudden increase over some baselines. Because we expect to detect the emerging topics before t_h as early as possible, which means to detect the emerging topics by predicting the sudden increase in the early period, so our attention is concentrated on the emerging phase. This

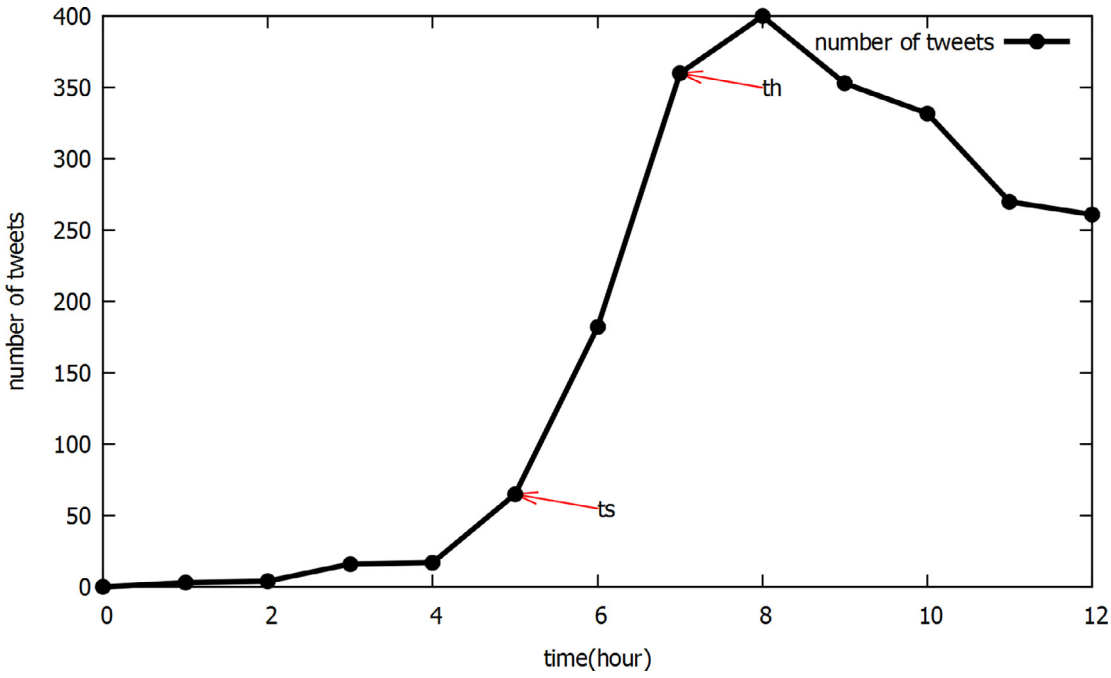

Fig. 2. The evolution of an emerging topic.

sudden increase is actually driven by certain immanent characteristics of emerging topic which cannot be represented by frequency obviously. As a result, we should first analyze these latent characteristics cautiously.

In micro-blogging, it has been found that the vast majority of emerging topics live and die over a timescale of at most a few hours (Nikolov & Shah, 2012). The immanent characteristics of emerging topics are multitudinous due to the varieties of user influences and their behaviors. All these make it difficult to evaluate the topic by static features. By the labeled topics, we compare the differences between emerging topics and non-emerging topics, and dynamically analyze the qualities which emerging topics share in common. We find that emerging topics could usually attract users in the early period, and the number of attracted users increases speedily. In other words, the topic with interesting and significant content will attract the users instinctively. Also, users with large number of followers could make greater influence than the user with less followers if they retweet the same tweet. This is due to a more widely diffuse through their social relationships of following. In summary, the emerging topics in micro-blogging networks have the following characteristics:

- *Attractiveness*: The *attractiveness* of topic means the capacity of topic that could attract users to join in the topic diffusion.
- *Key-node*: The *key-node* means the influential user who can cause vast amount of retweets in the topic diffusion, and these users could greatly promote the development of emerging topic.

4.2. Feature selection

In this subsection, we first introduce two types of network which represent the relationships of retweeting and following respectively, then we select features from these networks to represent the inherent characteristics of the emerging topics. As we discussed in the related work, many features extracted from term frequency, term distribution, and time feature have been applied to the topic detection. Different from previous researches, we select the features that have obviously and straightforward relations

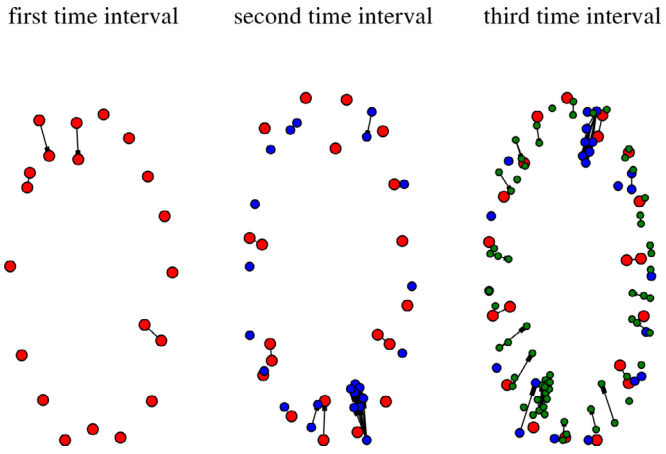

Fig. 3. The dynamic change of retweeting network of an emerging topic.

with emerging topics in order to dynamically describe the topologic structures of topic diffusion.

In this paper, we only consider two kinds of user behavior which are tweeting and retweeting. A user can join in a topic T with tweeting or retweeting the content about the topic in micro-blogging networks. Based on Cascading Model, information cascades are formed by retweeting behaviors (Yang & Leskovec, 2010). Users retweet the same tweet can form a retweeting chain $g_{T,Nc}$, and more retweeting chains about the same topic can make up the retweeting network. The retweeting network of topic T is a directed graph $G_T=(V,E)$, where each node in V represents a user involved in the topic T, and each edge $<V_a, V_b>$ in E represents user V_b retweets a tweet from user V_a about topic T. The retweeting network G_T also can be described as a set consisting of retweeting chains $G_T = [g_{T,1}, g_{T,2}, \cdots, g_{T,Nc}]G_T= [g_{T,1},g_{T,2},\cdots,g_{T,Nc}]$. Fig. 3 illustrates the dynamic change of retweeting network of an emerging topic in three continuous time intervals, and nodes of different time intervals are distinguished by different colors.

Besides, the following network is a directed graph $G_U=(U, F)$ which is generated by the following relationships (Cataldi, Di Caro, & Schifanella, 2010), where U is the user set and F is the set of directed edges. Given two users u_a and u_b, the edge $<u_a, u_b>$ exists only if u_b is a follower of u_a. The out-degree of u_a is the number of users who followed user u_a, and the in-degree of u_a is the number of users who are followed by u_a.

We compare the differences between emerging topics and non-emerging topics in four aspects which are the total number of nodes, the increased number of nodes in each time interval, the number of retweeting chains, and the maximum node number of retweeting chains. We select four typical topics including two emerging topics and two non-emerging topics to illustrate the differences in Fig. 4.

As the *attractiveness* of emerging topic is a nonspecific characteristic, rather than measuring its semantic meaning in Natural Language Processing (NLP) way, we use the total number of nodes and the number of retweeting chains to represent the *attractiveness* by their conditional independence. The total number of nodes in retweeting network implies the number of users who join in the topic diffusion. The number of retweeting chains is the number of distinct tweets which are retweeted by other users. It can be seen from Fig. 4 that emerging topics have more users and retweeting chains than the non-emerging topics. Therefore, the *attractiveness* of topic has direct dependencies with the total number of users and the number of retweeting chains.

Detecting the *key-node* implies that we need to find out if any influential users are involved in the topic, and these users may already or will have a significant impact on the topic diffusion. We analyze the users with large out-degree in the retweeting network and their out-degrees in the following network. The nodes with large out-degree are usually contained in the retweeting chains with large number of nodes, and a large size of retweeting chain reveals that someone has already impacted the topic diffusion through whose tweet has been retweeted for plenty times. Therefore, the users with large out-degree in retweeting network can be recognized as *key-node*. Meanwhile, the number of followers has a direct ratio with the retweet rate (Suh, Hong, Pirolli, & Chi, 2010). This suggests that user with large out-degree in the following network (follower number) has a greater possibility to be a *key-node*. Thus, whether the topic has any *key-node* can be straightly reflected by the maximum node number of retweeting chains and the total number of followers.

4.3. Dynamic Bayesian Networks

We propose a new design of DBNs for emerging keyword, which has three hidden variables and four observable variables. The hidden variable X^1 represents whether the keyword is an *attractiveness* one, hidden variable X^2 represents whether there has any *key-node* in the diffusion of topic containing the keyword, hidden variable X^3 represents whether the keyword is an emerging keyword. The four observable variables O^1, O^2, O^3, and O^4 represent the number of nodes in the retweeting network, the number of retweeting chains, the maximum node number of retweeting chains, and the total number of followers respectively.

Using the first-order Markov assumption we propose a new emerging keyword detection model as shown in Fig. 5, with hidden variables in square nodes and observable variables in circle nodes. The observable variables are assumed to be independent of each other. Each variable is clustered into three states by K-means algorithm, i.e., for variable O^1, we use three discrete states $O^1=i$, for i in (0, 1, 2) to represent different dimensions of the variable.

Like the definition which was employed in Hidden Markov Model (HMM) (Eddy, 1996), parameters can be defined as follows:

$$\begin{aligned}
\lambda &= (\pi^1, \pi^2, A^1, A^2, A^3, B^1, B^2)\\
\pi_i^1 &= P(X_1^1 = i), 0 \le i \le 1\\
\pi_g^2 &= P(X_1^2 = g), 0 \le g \le 1\\
A_{ij}^1 &= P(X_t^1 = i | X_{t-1}^1 = j), 0 \le i \le 1, 0 \le j \le 1\\
A_{gh}^2 &= P(X_t^2 = g | X_{t-1}^2 = h), 0 \le g \le 1, 0 \le h \le 1\\
A_{kig}^3 &= P(X_t^3 = k | X_t^1 = i, X_t^2 = g), 0 \le i \le 1, 0 \le g \le 1,\\
&\quad 0 \le k \le 1\\
B_i^1(y^1, y^2) &= P(O_t^1 = y^1, O_t^2 = y^2 | X_t^1 = i), 0 \le y^1 \le 2, 0 \le y^2 \le 2\\
B_g^2(y^3, y^4) &= P(O_t^3 = y^3, O_t^4 = y^4 | X_t^2 = g), 0 \le y^3 \le 2, 0 \le y^4 \le 2
\end{aligned}$$

λ is the parameter of DBNs model. π^1 and π^2 are the initial state distributions. A^1, A^2, and A^3 are the transition matrices for three hidden variables. B^1 and B^2 are the observation matrices between hidden variables and their observable variables.

4.4. DBNs learning and inference

One of the advantages of using DBNs is that the graphical structure representing the conditional independence among variables allows us to compute the joint probability of a subset of variables very efficiently (Murphy, 2002). The inference of DBNs is to compute the marginal probability $P(X_i|O_{1:t})$ of hidden variables X_i given an observation sequence $O_{1:t}=O_1, O_2,\ldots,O_t$.

The joint probability of variables can be factored into a product of local conditional probabilities one for each variable through conditional independencies or d-separation (Nielsen & Jensen, 2009).The full joint probability for the DBNs (Fig. 5) can be computed as follow:

$$\begin{aligned}
P(X_{1:T}^{1:3}, O_{1:T}^{1:4}) &= P(O_{1:T}^{1:4} | X_{1:T}^{1:3}) P(X_{1:T}^{1:3})\\
&= P(X_1^1) P(X_1^2) P(X_1^3 | X_1^1, X_1^2)\\
&\quad \times \prod_{t=1}^{T} P(O_t^1, O_t^2 | X_t^1) P(O_t^3, O_t^4 | X_t^2)\\
&\quad \times \prod_{t=2}^{T} P(X_t^1 | X_{t-1}^1) P(X_t^2 | X_{t-1}^2) P(X_t^3 | X_t^1, X_t^2)
\end{aligned}$$

where $X_{1:T}^{1:3} = \begin{bmatrix} X_1^1 \\ X_1^2 \\ X_1^3 \end{bmatrix} \cdots \begin{bmatrix} X_T^1 \\ X_T^2 \\ X_T^3 \end{bmatrix}$ and $O_{1:T}^{1:4} = \begin{bmatrix} O_1^1 \\ O_1^2 \\ O_1^3 \\ O_1^4 \end{bmatrix} \cdots \begin{bmatrix} O_T^1 \\ O_T^2 \\ O_T^3 \\ O_T^4 \end{bmatrix}$

In the training stage of DBNs, our task is finding the optimal parameter λ that computes the maximum likelihood over the training data, $\hat{\lambda} = \text{argmax}_\lambda P(O_{1:T}^{1:4} | \lambda)$. Just like the HMM, the parameters of the model include initial state probabilities of X^1 (π^1), initial state probabilities of X^2 (π^2), transition probabilities of X^1 (A^1), transition probabilities of X^2 (A^2), transition probabilities of X^3 (A^3), observation probabilities of X^1 (B^1), and observation probabilities of X^2 (B^2).The three hidden variables of proposed DBNs can be trained by exploiting the EM algorithm (Murphy, 2002). The parameters and their update formulas with the maximum likelihood method are given as follows:

$$\hat{\pi}_i^1 = E[X_1^1 = i] = \frac{P(O_1^1, O_1^2, X_1^1 = i | \lambda)}{P(O_1^1, O_1^2 | \lambda)}$$

$$\hat{\pi}_g^2 = E[X_1^2 = g] = \frac{P(O_1^3, O_1^4, X_1^2 = g | \lambda)}{P(O_1^3, O_1^4 | \lambda)}$$

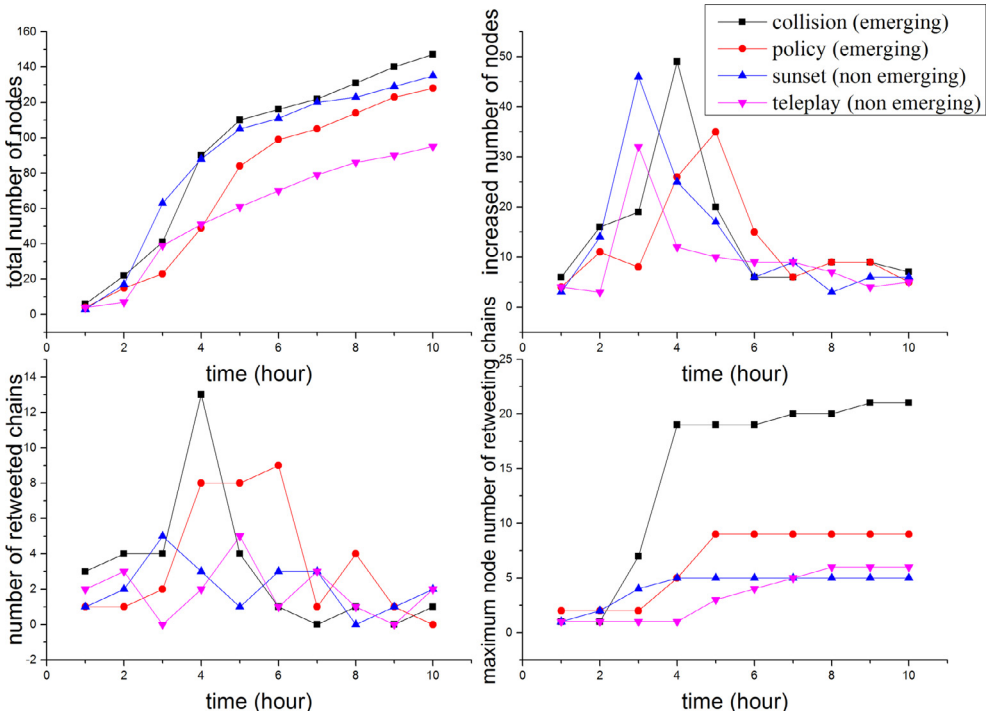

Fig. 4. The comparison between two types of topics.

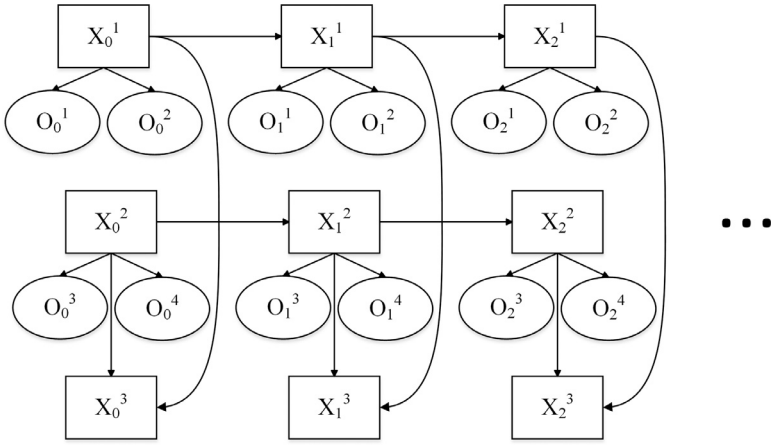

Fig. 5. The proposed Dynamic Bayesian Networks for emerging keyword.

$$\hat{A}^1_{ij} = \frac{E[X^1_{t+1}=j|X^1_t=i]}{E[X^1_t=i]} = \frac{\sum_{t=1}^{T-1}\frac{P(O^1_t,O^2_t,X^1_{t+1}=j,X^1_t=i|\lambda)}{P(O^1_t,O^2_t|\lambda)}}{\sum_{t=1}^{T-1}\frac{P(O^1_t,O^2_t,X^1_t=i|\lambda)}{P(O^1_t,O^2_t|\lambda)}}$$

$$\hat{A}^2_{gh} = \frac{E[X^2_{t+1}=h|X^2_t=g]}{E[X^2_t=g]} = \frac{\sum_{t=1}^{T-1}\frac{P(O^3_t,O^4_t,X^2_{t+1}=h,X^2_t=g|\lambda)}{P(O^3_t,O^4_t|\lambda)}}{\sum_{t=1}^{T-1}\frac{P(O^3_t,O^4_t,X^2_t=g|\lambda)}{P(O^3_t,O^4_t|\lambda)}}$$

$$\hat{A}^3_{kig} = \frac{E[X^3_t=k|X^1_t=i,X^2_t=g]}{E[X^1_t=i,X^2_t=g]} = \frac{\sum_{t=1}^{T-1}\frac{P(O^1_t,O^2_t,O^3_t,O^4_t,X^3_t=k,X^1_t=i,X^2_t=g|\lambda)}{P(O^1_t,O^2_t,O^3_t,O^4_t|\lambda)}}{\sum_{t=1}^{T-1}\frac{P(O^1_t,O^2_t,O^3_t,O^4_t,X^1_t=i,X^2_t=g|\lambda)}{P(O^1_t,O^2_t,O^3_t,O^4_t|\lambda)}}$$

$$\hat{B}^1_i(y^1,y^2) = \frac{E[O^1_t=y^1,O^2_t=y^2|X^1_t=i]}{E[X^1_t=i]} = \frac{\sum_{t=1}^{T}\frac{P(O^1_t=y^1,O^2_t=y^2,X^1_t=i|\lambda)}{P(O^1_t,O^2_t|\lambda)}}{\sum_{t=1}^{T}\frac{P(X^1_t=i|\lambda)}{P(O^1_t,O^2_t|\lambda)}}$$

$$\hat{B}^2_g(y^3,y^4) = \frac{E[O^3_t=y^3,O^4_t=y^4|X^2_t=g]}{E[X^2_t=g]} = \frac{\sum_{t=1}^{T}\frac{P(O^3_t=y^3,O^4_t=y^4,X^2_t=g|\lambda)}{P(O^3_t,O^4_t|\lambda)}}{\sum_{t=1}^{T}\frac{P(X^2_t=g|\lambda)}{P(O^3_t,O^4_t|\lambda)}}$$

In the detecting stage, the Bayesian rule is used to obtain the joint probability between the states of variables, the probability can be defined as:

$$\begin{aligned} &P(X^3_t=k|y^1_{1:t},y^2_{1:t},y^3_{1:t},y^4_{1:t}) \\ &= \sum_{i=1}^{M_1}\sum_{g=1}^{M_2} P(X^1_t=i|y^1_{1:t},y^2_{1:t})P(X^2_t=g|y^3_{1:t},y^4_{1:t}) \\ &\quad \times P(X^3_t=k|X^1_t=i,X^2_t=g) \end{aligned} \tag{1}$$

In this equation, the joint probability $P(X^3_t = k|y^1_{1:t}, y^2_{1:t}, y^3_{1:t}, y^4_{1:t})$ is the probability that variable $X_t{}^3$ belongs to the state k at time interval t given all the observation sequences $y^1_{1:t}$, $y^2_{1:t}$, $y^3_{1:t}$, $y^4_{1:t}$

from time interval 1 to t. $P(X_t^1 = i | y_{1:t}^1, y_{1:t}^2)$ is defined as the probability that hidden variable X^1 belongs to state i at time interval t given observation sequences $y_{1:t}^1$, $y_{1:t}^2$ from interval 1 to t, $P(X_t^2 = g | y_{1:t}^3, y_{1:t}^4)$ is similarly defined. $P(X_t^3 = k | X_t^1 = i, X_t^2 = g)$ is the probability that hidden variable X^3 belongs to state k when the hidden variables $X_t^1 = i$ and $X_t^2 = g$ at time interval t. In other words, this probability is a part of the transition matrix A^3, $P(X_t^3 = k | X_t^1 = i, X_t^2 = g) = A_{kig}^3$. The observation sequences $y_{1:t}^1$, $y_{1:t}^2$, $y_{1:t}^3$, $y_{1:t}^4$ can be obtained for each keyword during a certain time period. By these sequences, the probabilities $P(X_t^1 = i | y_{1:t}^1, y_{1:t}^2)$ and $P(X_t^2 = g | y_{1:t}^3, y_{1:t}^4)$ can be obtained by Viterbi Algorithm based on HMM respectively as follows:

The initial probabilities are acquired by followed equations:

$$P(X_1^1 = i | y_{1:t}^1, y_{1:t}^2) = \pi_i^1 B_i^1(y_1^1, y_1^2) \tag{2}$$

$$P(X_1^2 = g | y_{1:t}^3, y_{1:t}^4) = \pi_g^2 B_g^2(y_1^3, y_1^4) \tag{3}$$

Then, for $2 \le t \le T$, the probability can be iterated from Equation (2) and (3) as:

$$P(X_t^1 = i | y_{1:t}^1, y_{1:t}^2) = \max_{1 \le j \le M_1} [P(X_{t-1}^1 = j | y_{1:t}^1, y_{1:t}^2) A_{ji}^1] B_i^1(y_t^1, y_t^2) \tag{4}$$

$$P(X_t^2 = g | y_{1:t}^3, y_{1:t}^4) = \max_{1 \le j \le M_1} [P(X_{t-1}^2 = h | y_{1:t}^3, y_{1:t}^4) A_{hg}^2] B_g^2(y_t^3, y_t^4) \tag{5}$$

where π_i^1, π_g^2, A_{ji}^1, A_{hg}^2, B_i^1, and B_g^2 are the parameters of DBNs.

The probability $P_t^{w_z}$ of a keyword w_z being an emerging keyword at time interval t is defined as:

$$P_t^{w_z} = P(X_t^3 = 1 | y_{1:t}^1, y_{1:t}^2, y_{1:t}^3, y_{1:t}^4) \tag{6}$$

where $X_t^3 = 1$ represents the state of variable X^3 is 1 at time t, and it means that the keyword w_z is an emerging one. $y_{1:t}^1$, $y_{1:t}^2$, $y_{1:t}^3$, $y_{1:t}^4$ are the feature sequences of keyword w_z from starting time 1 to time t.

To identify the emerging keywords at time t, we need to set a threshold parameter of probability p_{th}, and the emerging keywords can be identified by:

$$\forall w_z \in W_t, P_t^{w_z} > p_{th} \Leftrightarrow w_z \in K_t$$

where K_t is the emerging keyword list at time interval t. The probability parameter p_{th} will be discussed in the experiment.

4.5. Cluster

After we obtain the emerging keyword list K^t at time interval t, we analyze the co-occurrence relations between each pair of keywords, and cluster the emerging keywords into emerging topics. Each emerging topic can be described by a set of emerging keywords which are semantic related. For example, the topic that cars could pass the highways for free during specific legal holiday can be represented as a set consisting of emerging keywords: {highways, cars, holiday, free}.

Compared with huge size of tweet corpus in each time interval, the emerging keyword list is much easier for clustering. We apply the DBSCAN algorithm (Ester, Kriegel, Sander, & Xu, 1996) to cluster emerging keywords into emerging topics by co-occurrence relations. The co-occurrence relation between keywords w_1 and w_2 at time interval t can be briefly defined as the correlation C_{w_1, w_2}^t, and can be formulated as:

$$C_{w_1,w_2}^t = \frac{N_{w_1,w_2}^t}{N_{w_1}^t + N_{w_2}^t - 2N_{w_1,w_2}^t} \tag{7}$$

where $N_{w_1}^t$ is the number of tweets containing the keyword w_1 at time interval t, and $N_{w_2}^t$ is similarly defined. N_{w_1,w_2}^t is the number of tweets containing both keywords w_1 and w_2 at time interval t.

Table 2
Several topics detected by our method.

Keywords	Detected time	Numbers
Earthquake, depth, junction, Yunnan	2012-09-07 12:00	813
Happy, teacher, day	2012-09-10 07:00	4997
Aircraft, carrier, marine, LiaoNing	2012-09-23 19:00	573
Voice, China, final	2012-09-30 21:00	11856

With the Equation (7), we can implement the cluster of keywords with a threshold parameter C^t in supervised or unsupervised way. The supervised way means that the threshold parameter of correlation is determined with an empirical value, and the threshold parameter of unsupervised way is calculated by the average value of all the correlations at each time interval.

At time interval t, the topic can be clustered by:

$$\forall w \in K^t, w \in T\{w_1, \cdots w_z\} \Leftrightarrow C_{w,w_n}^t > C^t, for w_n in T$$

where topic T is defined as a set consisting of emerging keywords $\{w_1, \ldots w_z\}$.

5. Experimental results

In this section, we evaluate the proposed method on real dataset which is collected from Sina micro-blogging. The experimental results are discussed from two aspects which are emerging keywords and emerging topics respectively. The Chinese word segmentation has been done by ICTCLAS[2] tools.

5.1. Emerging keywords

This subsection aims to show the results of emerging keyword detection by Dynamic Bayesian Networks. The probability of each keyword being an emerging keyword can be acquired along the timelines. Fig. 6 shows an example of keyword which turns into an emerging keyword in a few hours. The probability increases gradually with the growth of topic. There is a distinct tradeoff between the probability parameter and detecting time. The smaller probability parameter is, the earlier detecting time will be. For example, the keyword will be detected at 10:00 when the probability parameter is set to 0.6, if we set the probability parameter to 0.4, the detected time will be 9:00 (Fig. 6). In our experiment, the interval of probability is (0.1,0.7) under the DBNs parameter λ, and the p_{th} will be chosen from this range.

In order to evaluate the performance of emerging keyword detection, the correct emerging keywords are crudely decided by the term frequency which can be acquired by experience. Fig. 7 shows the accuracy comparison with different probability parameters. The result shows that the accuracy of emerging keywords increases rapidly with the increasing of probability parameter. The sudden raise of the accuracy of non-emerging keywords demonstrates that the accuracy achieves and maintains 95% when the probability parameter p_{th} is larger than 0.2. As a result, a large probability parameter should be selected for a comprehensive consideration of emerging and non-emerging keywords.

5.2. Emerging topics

In this subsection, experiments are carried out to demonstrate the effectiveness of our approach. By clustering the emerging keywords into emerging topics, we detect 41 emerging topics during one month and Table 2 gives several detected topics specifically. In Table 2, detected time is the earliest time when the keywords of topic are detected, and numbers is the number of tweets which belong to the topic in the next 5 hours since the detected time.

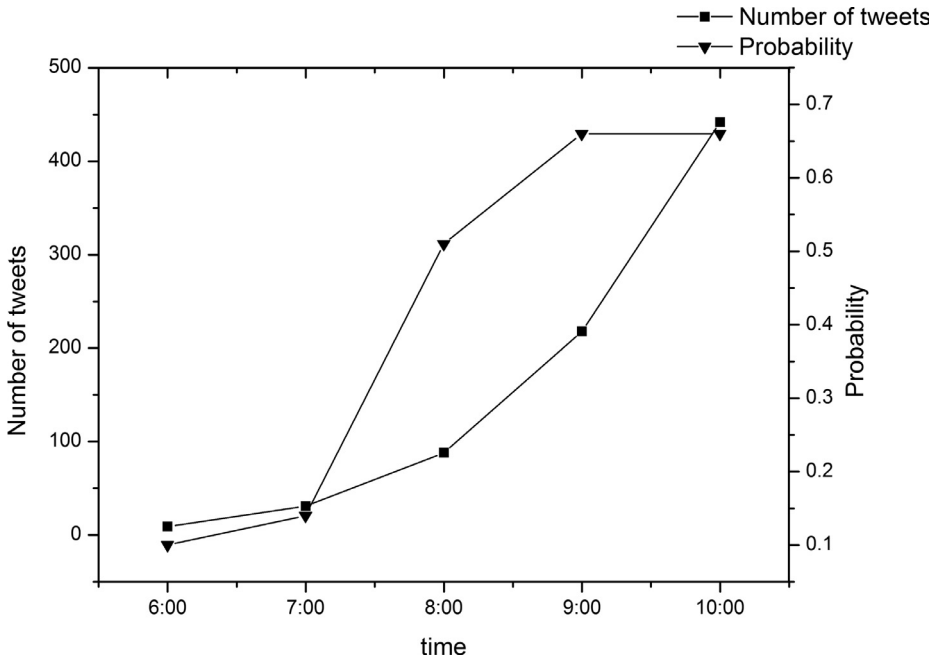

Fig. 6. The dynamic change of probability.

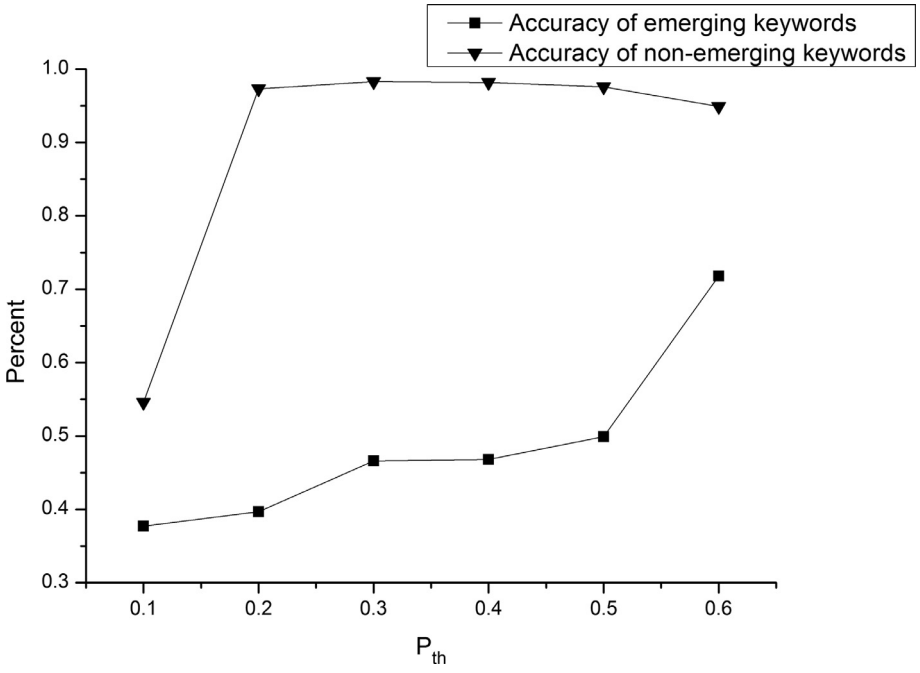

Fig. 7. The comparison of accuracy with different probability parameters.

The evaluation criterions not only refer to the traditional information retrieval precision, recall, and F_1-measure, but also introduce delay time as a new evaluation criterion.

We first compare the performances of our method with different probability parameters p_{th}. Then we compare the performances with several other methods introduced below:

Topic Graph (TG) (Cataldi, Di Caro, & Schifanella, 2010): It is a system that performs emerging topic detection over two core algorithms, which are aging theory and Page Rank. Here we set dumping factor d=0.01, $drop^t$=10.

MACD (Du, Wu, He, & Liu, 2012): This is a novel way based on MACD (Moving Average Convergence/Divergence) model which is widely used in stock market. Here we set α=0.35, β=100, γ=0.5, θ=1, n=5.

TF-IDF+NN: This is a classical model for topic detection. It utilizes *TF-IDF* to detect emerging keywords and cluster the emerging topics by Nearest Neighbor algorithm. Here we set the threshold parameter of *TF-IDF* to 1, the distance between two words refers to Ruthven & Lalmas (2003), and the threshold of distance is set to -1.

Fig. 8 illustrates the performances of our method with different probability parameters. It can be seen that different probability parameters achieve nearly the same performance with recall. Performances of precision and F_1-measure increase steadily with the increasing of probability parameter. The main reason for

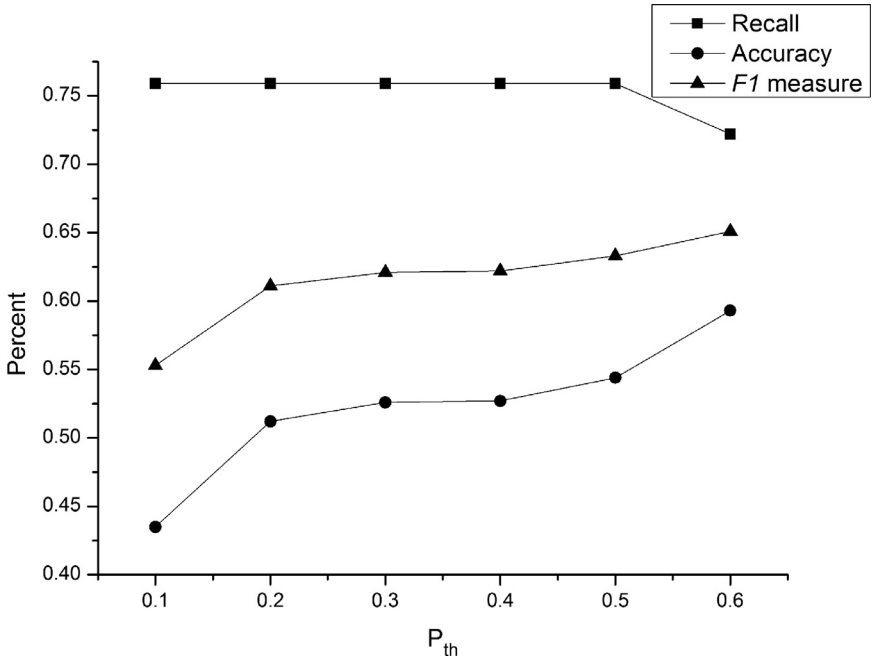

Fig. 8. The comparison of recall, accuracy, and *F1*-measure with different probability parameters.

Table 3
Performance comparison with different methods.

Model	Precision	Recall	F_1-measure	Delay time (hour)
Our method	59.3%	72.2%	65.1%	–
TG	66.7%	53.7%	59.5%	1.33
MACD	93.7%	44.4%	60.3%	2.27
TF-IDF+NN	51.4%	70.4%	59.4%	1.24

this appearance is that the real emerging keywords usually get higher probabilities of becoming emerging keywords with DBNs. With the increase of probability parameter p_{th}, our method could filter the keywords with lower probabilities. These filtered keywords have barely influence the clustering results of emerging topic, which can be seen from the performance of recall. In general, our method achieves the best performance with the probability parameter p_{th}=*0.6*.

As discussed above, we evaluate the proposed method by setting the probability parameter p_{th}=0.6 and time interval t=1 hour. The performance comparisons with different methods are shown in Table 3. Since our method achieves better performance measured by considering the detected time, delay time means the average delay time when detecting the same topic compared with our method.

Table 3 shows the precision, recall, F_1-measure, and delay time of different methods respectively. From the table, it is observed that our method achieves better performance than *TF-IDF*+NN across all evaluation metrics. TG gets a higher precision and a lower recall, and the implement of aging theory improves the accuracy partly. The main reason for the poor performance of *TF-IDF*+NN is that the features of *TF-IDF* with large amount of text documents could cause the deficiency of important emerging terms. MACD gets the best precision and the worst recall, that is because the excessive concern of dynamic changes filters some emerging keywords partially. Meanwhile, we find that our method is able to detect emerging topics in advance of other methods with one to two hours. It is mainly because of the features used in Dynamic Bayesian Networks represent the topology changes of topic diffusion in the early period, from which emerging keywords could be identified effectively.

In summary, this paper emphasizes the timeliness of emerging topic detection, and the experimental results validate the effectiveness of our method. From the experimental results, we can obtain useful insights presented as following:

- The keywords with same term frequency but different diffusion topology reflect the different emerging levels. If more users retweet the tweets with a certain keyword and form more retweeting chains, then the keyword will have a large probability to be an emerging one. In other words, retweeting is more important for information diffusion than tweeting.
- The DBN-based model achieves a good representation for keyword's temporal evolution, and the performances compared with the other methods confirm the effectiveness of proposed model for real-time applications.

These insights are not only useful for researchers to study trends in micro-blogging networks, but also supportive for engineers to enhance their real-time systems.

6. Conclusion

In this paper, we propose a novel method to solve the problem of early detection of emerging topics in micro-blogging networks. Different from previous researches in the feature selection and the approach of detecting emerging trends, we select the features from topology properties of topic diffusion and detect the emerging trends by the dynamic changes of conditional probabilities calculated by DBNs. First, we analyze the main characteristics of emerging topics, which could be characterized by the topology features of the retweeting network and the following network. Then we develop a way to detect emerging keywords based on Dynamic Bayesian networks. The DBN-based model could appropriately represent the temporal evolution of emerging keyword to calculate the probability of a keyword being an emerging one. Finally, based on the co-occurrence relations of keywords, we apply the DBSCAN algorithm to cluster the emerging keywords into emerging topics. With the experimental result on real data, we illustrate

the effectiveness of our proposed method by comparing with the other methods for real-time applications.

We also find one limitation in this work as well. The DBN-based model calculate the state of a certain variable by the conditional probabilities between the variables. If we add more features in the DBN-based model, the states of variables and computational complexity will increase multiply, so a few effective features are necessary for the model.

The future work can be carried out from several directions. The first direction is the improvement of features employed for keyword detection, and the features can be selected from the semantic relations and temporal dimension. Based on our method, the second direction is to extract different characteristics from different types of topics for domain-related systems, such as political, entertainment, sports. While the current work applies the DBN-based model to represent the temporal evolution of keyword, the third direction would extend this model to the dynamic process of individual or group user behavior. Finally, it would be quit useful to consider the applications of our method to other intelligent systems which emphasize on timeliness, e.g., real-time recommendation system, ad-targeting system, marketing strategy system.

Acknowledgements

The research presented in this paper is supported in part by the National Natural Science Foundation (61572397, 61502383, 61375040, 61571360), Fundamental Research Project of Natural Science in Shaanxi Province (2015JM6298, 2015JM6299), Specialized Research Fund for the Doctoral Program of Higher Education (20120201120023) and Specialized Research Plan Funded Project of Shaanxi Province Department of Education (15JK1505).

References

Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study final report.

Allan, J. (2002). Introduction to topic detection and tracking. *Topic detection and tracking* (pp. 1–16). Springer US.

Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM, 40*(3), 66–72.

Bao, B. K., Xu, C., Min, W., & Hossain, M. S. (2015). Cross-platform emerging topic detection and elaboration from multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 11*(4), 54.

Bernabé-Moreno, J., Tejeda-Lorente, A., Porcel, C., & Herrera-Viedma, E. (2015). A new model to quantify the impact of a topic in a location over time with Social Media. *Expert Systems with Applications, 42*(7), 3381–3395.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993–1022.

Bun, K. K., & Ishizuka, M. (2002). Topic extraction from news archive using TF∗ PDF algorithm. In *Web Information Systems Engineering, International Conference on* (p. 73).

Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining* (p. 4).

Chen, C., Buntine, W., Ding, N., Xie, L., & Du, L. (2015). Differential topic models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 37*(2), 230–242.

Chen, C. C., Chen, Y. T., Sun, Y., & Chen, M. C. (2003). Life cycle modeling of news events using aging theory. *Machine Learning: ECML 2003* (pp. 47–59). Springer Berlin Heidelberg.

Chen, F., & Neill, D. B. (2014). Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1166–1175).

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010, April). Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1185–1194).

Chen, K. Y., Luesukprasert, L., & Chou, S. C. (2007). Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *Knowledge and Data Engineering, IEEE Transactions on, 19*(8), 1016–1025.

Chen, Y., Amiri, H., Li, Z., & Chua, T. S. (2013). Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 43–52).

Chen, Z., & Liu, B. (2014). Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1116–1125).

Cieri, C. (2000). Multiple annotations of reusable data resources: Corpora for topic detection and tracking. In *Actes 5ième Journées Internationales d'Analyse Statistique des Données Textuelles (JADT)*.

Du, Y., Wu, W., He, Y., & Liu, N. (2012). Microblog bursty feature detection based on dynamics model. In *Systems and Informatics (ICSAI), 2012 International Conference on* (pp. 2304–2308).

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology, 6*(3), 361–365.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Forney G.D.,Jr (2005). The viterbi algorithm: A personal history. *arXiv preprint cs/0504020*.

Glavaš, G., & Šnajder, J. (2014). Event graphs for information retrieval and multi–document summarization. *Expert systems with applications, 41*(15), 6904–6916.

Hassan, A., Radev, D., Cho, J., & Joshi, A. (2009). Content based recommendation and summarization in the blogosphere. *Ann Arbor, 1001*, 48109.

Hayashi, K., Maehara, T., Toyoda, M., & Kawarabayashi, K. I. (2015). Real-Time Top-R Topic Detection on Twitter with Topic Hijack Filtering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 417–426).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning, 42*(1-2), 177–196.

Hu, J., Sun, X., & Li, B. (2015). Explore the evolution of development topics via online LDA. In *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on* (pp. 555–559).

Kim, Y., & Shim, K. (2014). TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems, 42*, 59–77.

Kim, H., Choo, J., Kim, J., Reddy, C. K., & Park, H. (2015). Simultaneous Discovery of Common and Discriminative Topics via Joint Nonnegative Matrix Factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 567–576).

Kotov, A., Zhai, C., & Sproat, R. (2011). Mining named entities with temporally correlated bursts from multilingual web news streams. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 237–246).

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (pp. 591–600).

Murphy, K. P. (2002). Dynamic bayesian networks. *Probabilistic Graphical Models, M. Jordan, 7*.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Nikolov, S., & Shah, D. (2012). A nonparametric method for early detection of trending topics. In *Proceedings of the Interdisciplinary Workshop on Information and Decision in Social Networks (WIDS 2012)*.

Ohsawa, Y., Benson, N. E., & Yachida, M. (1998). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on* (pp. 12–18).

Popescu, A. M., & Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1873–1876).

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review, 18*(02), 95–145.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real–time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851–860).

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on* (pp. 177–184).

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology, 62*(2), 406–418.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM, 10*, 178–185.

Unankard, S., Li, X., & Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. *World Wide Web, 18*(5), 1393–1417.

Vicient, C., & Moreno, A. (2015). Unsupervised topic discovery in micro-blogging networks. *Expert Systems with Applications, 42*(17), 6472–6485.

Wang, C., Zhang, M., Ru, L., & Ma, S. (2008). Automatic online news topic ranking using media focus and user attention based on aging theory. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1033–1042).

Wang, H., & Ohsawa, Y. (2013). Idea discovery: A scenario-based systematic approach for decision making in market innovation. *Expert Systems with Applications, 40*(2), 429–438.

Wang, Y., Agichtein, E., & Benzi, M. (2012). Tm-lda: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 123–131).

Wu, Y., Ding, Y., Wang, X., & Xu, J. (2010). On-line hot topic recommendation using tolerance rough set based topic clustering. *Journal of Computers, 5*(4), 549–556.

Yang, J., & Leskovec, J. (2010). Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 599–608).
Yang, Z., Wang, C., Zhang, F., Zhang, Y., & Zhang, H. (2015,). Emerging Rumor Identification for Social Media with Hot Topic Detection. In *2015 12th Web Information System and Application Conference (WISA)* (pp. 53–58).
Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., & Ma, W. Y. (2015). Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1351–1361). International World Wide Web Conferences Steering Committee.
Yu, R., Zhao, M., Chang, P., & He, M. (2014). Online hot topic detection from web news archive in short terms. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on* (pp. 919–923).
Zhang, C., Wang, H., Cao, L., Wang, W., & Xu, F. (2016). A hybrid term–term relations analysis approach for topic detection. *Knowledge-Based Systems*, *93*, 109–120.
Zhou, Y., Guan, X., Zheng, Q., Sun, Q., & Zhao, J. (2010). Group dynamics in discussing incidental topics over online social networks. *Network, IEEE, 24*(6), 42–47.