Li *et al. Smart Learning Environments* (2015) 2:12
DOI 10.1186/s40561-015-0019-6

Smart Learning Environments
a SpringerOpen Journal

RESEARCH Open Access

Towards better understanding of hot topics in online learning communities

Yanyan Li*, Yafeng Zheng, Haogang Bao and Yang Liu

* Correspondence: liyy@bnu.edu.cn
R&D Center for Knowledge Engineering, School of Educational Technology, Beijing Normal University, Beijing, China

Abstract

Online learning communities provide open workspaces allowing learners to share information, exchange ideas, address problems and discuss on specific themes. But with the continuously increasing artifacts in online communities, learners feel it difficult to quickly and easily gain an insight into a certain theme. To facilitate and support learners have a better understanding of the communication focus, this paper presents an approach to discover the hot topics and patterns of topics evolutions in online learning communities. Firstly, hot terms are extracted based on three features: the frequency of the terms used in the document collection; the location of the terms within a document; the breadth of terms distribution in the document collection. Then a term association network is constructed by computing the terms co-occurrence and distance between them. Finally, an algorithm is proposed to select the kernel term and its associated terms as term clusters to represent the hot topics with multi-facets expression. Two case studies on real datasets are conducted to demonstrate the effectiveness and usefulness of term cluster in helping users better understand hot topics in online learning communities. Potential applications in learning scenarios are also discussed.

Keywords: Hot topic; Term extraction; Term association network; Online learning communities

Introduction

With the widespread adoption of social media, more and more people prefer to express and share their true thoughts and opinions in diverse communities. Online learning environments are now perceived as a network of knowledge comprised of interconnected individuals and enormous amounts of artifacts. Providing the collaborative means for achieving shared creation and shared understanding, mutual exchange between community members is encouraged to support individual and collective learning (Woolley & Ludwig-Hardman 2000). However, with the increasing growth of artifacts in online learning communities, it is difficult for learners to easily have a global view on a specific topic or know about how the topics evolve over time in such large text datasets. For example, if one wants to know the emerging popular technologies in a discussion forum on advanced learning technologies, it would be a frustrating experience for him or her to read through all discussion transcripts.

Therefore, to better understand what hot topics are discussed in online learning communities becomes an important task. Herein, by saying "hot topic", we refer to the topics that appear frequently in online learning communities during a period of time. Through exploiting the hot topics and tracking the trend of a specific topic, it would help learners keep abreast of popular, new and intertwining topics over time. Especially for a novice, he or she can quickly gain an insight of a certain theme by simply taking a look at the hot topics.

In this paper, we propose a framework-based representation of hot topics composed of focal term and facet term. Based on this framework, an algorithm is presented to discover the key terms and related terms to form the hot topics. The applications on real datasets from an educational blog and a "parent–child" forum have shown that our approach can effectively identify term clusters as hot topics in online learning communities and reveal the topic evolution patterns over time.

The remainder of this paper is organized as follows. Section 2 discusses the related work; Section 3 presents the approach to detect hot topics; Section 4 introduces the details of two case studies; Section 5 discusses the findings and implications for supporting informal learning; and section 6 concludes this paper.

Related work

The task of hot topic detection is to exploit topics that appear frequently during a period of time. Previous studies on topic detection can be classified into three categories: statistics-based approaches, linguistics-based approaches, and topic clustering approaches (Zhang & Li 2011).

Regarding statistics-based methods, different term-weighting schemes are adopted to capture the important or representative terms that feature in the content of a document. Based on hot terms' TF*PDF value, Bun and Ishizuka proposed to extract highest weighted sentences and pages from online news archive (Bun & Ishizuka 2002). Alghamdi H M et al. combined TF*PDF algorithm with improved vector space model to extract hot terms from on arabic dark web forums (Alghamdi and Selamat 2012). Based on the extracted hot terms, key sentences were identified and grouped into clusters to represent hot topics by using multidimensional sentence vectors. Zhang and Li used TF*PDF algorithm and life cycle modeling to extract hot terms from Yahoo! Answers (Zhang & Li 2011).

Linguistics-based approaches employ the linguistics features of words, sentences and documents to analyze the texts. Zheng proposed a document representation methodology to take into account both noun phrases and various semantic relationships, as there were various semantic relationships that could relate a pair of words (Zheng et al. 2009). Ginter proposed an unsupervised method, based on hidden Markov models, which was combined with latent semantic analysis to define topics of interest without necessarily data annotation; this method could also be used to identify short segments (Ginter et al. 2009).

Topic clustering has been studied in topic digital library construction and stock market news analysis. It comprises three steps: topic extraction, document clustering, and clustering description. K-means clustering and support vector machine method are usually adopted to group clusters (Li & Wu 2010; Preethi et al. 2012). Cui et al. used HDP to analyze various evolution patterns that emerge from multiple topics (Cui et al. 2011).

Most works related to topic detection focus on the topic extraction and general evolution, yet few researches has been conducted to examine how topic merges, fades and splits (Cui et al. 2011). On the other hand, the problem of topic detection has been widely studied in news (Zhang and Li, 2011; Jahnavi and Radhika 2013; Chen et al. 2007) and emails (Kleinberg 2003), and a few work has been done on blogs topic detection (Wang 2014; Kien-Weng Tan et al. 2011; Nagano et al. 2008; Bhadoria et al. 2012) in recent years. On the basis of statistics method, this study focuses on extracting the hot topics with different facets to represent the detailed discussion points and unveiling topic evolution patterns in online learning communities.

Hot topic detection approach

Following works in hot topic detection in online communities (Zhang & Li 2011), a "hot topic" is defined as a topic that is discussed frequently during a period of time. The usual ways use a term cluster to represent a topic, which has the potential defects that all the terms are delivered to the users with no differences or associated with a value of weight. This display manner of a set of terms often makes the users feel confused and difficult to understand the meaning of the topics. By contrast, by computing the relationship between terms, we present a straightforward way to convey the meaning of the topics via the combination of "focal term" and "facet terms". Each hot topic is composed of different topic facets, and each topic facet represents a certain aspect of the topic. Based on our previous work on automatic extraction of interpretable topics from online discourse (Zhang et al. 2012), we define a hot topic as $HT = \{focal\text{: } facet_1, facet_2, \ldots, facet_n\}$, where HT represents a hot topic. Each HT is composed of a "focal term" and several "facet terms". The focal term represents the core meaning, while the facet terms express the different aspects of the topic. The facet terms provide an in-depth explanation for the focal terms. In this way, the hot topics can be easily understood in a meaningful manner.

Key term extraction

Since the terms are the basic elements of topics, the first step to detect hot topics is to extract key terms. Figure 1 illustrates the procedure of hot terms extraction. As the figure shows, the process comprises three steps: term extraction, term filtering, and term selection.

To extract key terms from a text document, the basic language elements need to be considered. In the process of the key terms extraction, the step of "word segmentation

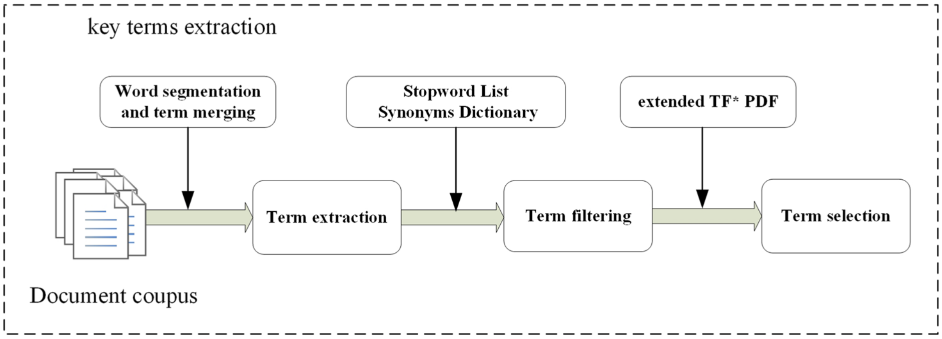


Fig. 1 The process of the key terms extraction

and term merging" is related to languages. In Chinese, the basic element is a character, encoded in a two-byte code. As Chinese is a hieroglyph language, each Chinese character has its own meanings and can be written in sequence without any spaces in-between. So, the first step is to split the documents into a list of separate term via tokenization and part of speech analysis. The terms in the list are sequenced according to their occurrence frequency. We assume that a longest repeated string is often a correct key term as its repetition provides sufficient evidence for the decision on its left and right boundaries. Thus, two adjacent terms in the list are merged into a longer term if their occurrence frequencies exceed a predefined threshold. The first term of the pair is accepted as a key term candidate if its frequency is greater than the threshold and did not merge with its preceding and following term. The process is executed iteratively until no keywords remained for processing in the list.

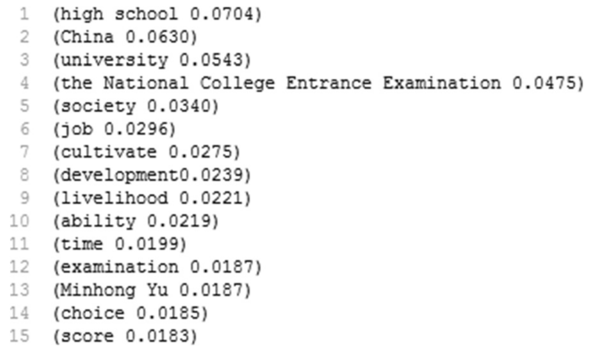
After obtaining the initial term list, the second step is to conduct term filtering according to the stopword list, whilst the common unmeaningful terms, such as prepositions, conjunctions and etc., are removed from the list. People prefer to use abbreviations and acronyms in online learning communities. Considering such informal characteristics of the network language, a synonym dictionary composed of synonym and acronym is also used for the term filtering. In this way, the different terms expressing the same meaning are treated as the one in the extracted key terms, and their occurrence frequencies are accumulated for the accepted one term. An acronym is a specific form of an abbreviation created from the capitalized initial letters or parts of a series of words (Rowe 2003). By referring to the acronym, the terms in different forms are unified and their occurrence frequencies are accumulated too.

Following that, in the third phase, key terms are selected based on three characteristics of a term: 1) the frequency of the term used in the document collection; 2) the location of terms in a documents; and 3) the breadth of terms distribution in the document collection. Traditional TF*IDF considers the term frequency, Bun and Ishizuka (Bun & Ishizuka 2002) proposed the TF*PDF method by taking the term breadth into account. The documents in online learning communities are usually composed of title and content, and title usually reflect the main theme of the document, so we attach more importance to the keywords in titles. Therefore, we propose the extended TF * PDF scheme to measure the significance of a term. The basic idea of this scheme is that terms are assigned greater weights that occur frequently in many documents or appeared in the important location of the documents, and lower weights to those that are rarely mentioned. The significance of a term i denoted as Sig_i is calculated in (1).

$$\mathrm{Sig}_i = \frac{f_i}{\sum_{t-1}^{F} f_t^2} * \exp\left(\frac{|N_i|}{N}\right) * \frac{\sum_{i-1}^{N} L(i)}{N} \tag{1}$$

Where f_i is the frequency of term i in the given subset. N_i is set of documents in the given subset where term i appears. F is the total number of terms and N is the total number of documents in the given subset. Herein we define that when term i appears in the title of a document, $L(i)$ is assigned the value 2; otherwise $L(k)$ is assigned the value 1.

By sorting the terms in the list with the weight value, the top-ranked k terms are chosen as key terms. For example, Fig. 2 illustrates the list of selected key terms with weight from the document subset with timestamp June-July 2010.



```
(high school 0.0704)
(China 0.0630)
(university 0.0543)
(the National College Entrance Examination 0.0475)
(society 0.0340)
(job 0.0296)
(cultivate 0.0275)
(development0.0239)
(livelihood 0.0221)
(ability 0.0219)
(time 0.0199)
(examination 0.0187)
(Minhong Yu 0.0187)
(choice 0.0185)
(score 0.0183)
```

Fig. 2 A list of selected key terms with weight

Term association analysis

Various measures have been proposed to calculate association between two terms (He 1999 & Salton 1989). The kernel idea to the different measures is based on the number of times for two terms co-occurrence at the same documents. Chen et al. proposed three assumptions to define relations among keywords (Cui et al. 2011): 1) If two keywords appear in one article, it implies that a certain relation exists between these two keywords; 2) The higher the frequency of occurrences of two keywords appearing in one sentence, the higher the relation between them; 3) The shorter the "distance" between two keywords in one sentence, the higher the relation is between them. Moreover, Lee et al. deemed that as the number of words in a sentence increase, the relation between two keywords decrease (Lee & Segev 2012). In other words, the score of a relation in a shorter sentence is higher than the score of a relation in longer sentences.

In this study, we follow these criteria and propose a modified Cosine Similarity to compute the association weight between two terms (Liu et al. 2013). The weight of association between two terms is calculated in the Eq. (2).

$$\mathrm{wgt}\ (t_i,\ t_j)\ =\ \frac{s_{ij}}{\sqrt{s_i\ \cdot}}\ s_j\ \cdot\ \left(\log\frac{\max\ \left(avg_{numij}\right)}{avg_{numij}}\right)^{\alpha}\ \cdot\ \left(\log\frac{\max\ \left(avg_{dij}\right)}{avg_{dij}}\right)^{\beta},\ \ i\ \neq\ j. \tag{2}$$

Where *wgt(t$_i$, t$_j$)* is the degree of correlation between the term *i* and *j*. The variable s_{ij} is the number of sentences in which term *i* and *j* co-occur. And s_i (s_j) denotes the number of sentences in which T_i (T_j) occurs regardless of T_j (T_i). α and β are regularization parameters.

$$avg_{d_{ij}}\ =\ \frac{\sum_{m=1}^{s_{ij}} d_m}{s_{ij}},\ \ i\ \neq\ j. \tag{3}$$

Where d_m denotes the number of characters between the term *i* and *j* in the sentence in which both terms occur in the Eq. (3).

$$avg_{num_{ij}}\ =\ \frac{\sum_{m=1}^{s_{ij}} num_m}{s_{ij}},\ \ i\ \neq\ j. \tag{4}$$

Where num_m denotes the number of characters in the sentence where term i and j co-occur in the Eq. (4).

Therefore, the extracted key terms comprise a term association network in the form of a weighted undirected graph $G = (V,E), V = \{t_1,\ t_2,\ t_3,\ \ldots,\ t_n\}$, $E(G) = ((t_1,\ t_2),\ (t_1,t_3),\ \ldots,\ (t_{n\text{-»}},\ t_n))$. Where V is a set of nodes indicating the terms, and E is a set of edges indicating the association between terms. Each edge is associated with a weight $Wgt(t_i,\ t_j)$ to express the association intensity between two key terms.

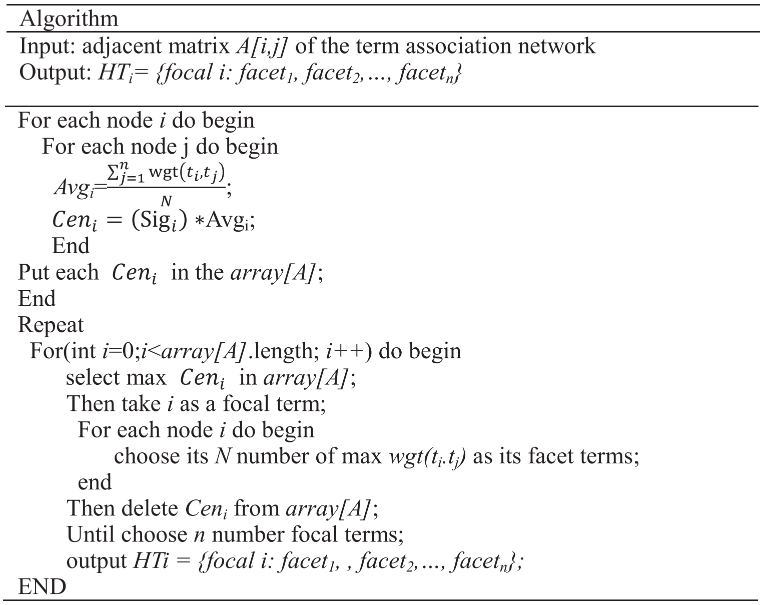
Hot term clustering

The key to discover hot topics from the term association network is to select the "focal term" of each hot topic. Given a term in the network, two factors are considered to select the focal term with higher centrality. One is the importance of the term itself, and the other is its association intensity with other adjacent terms. Equation (5) is used to compute the centrality of each term.

$$Cen_i = Sig_i * \frac{\sum_{j=1}^{n} wgt\ (t_i,\ t_j)}{N} \tag{5}$$

Where Sig_i measures the importance of term i in terms of frequency, breadth, and location of a term. $wgt(t_i,\ t_j)$ is the degree of correlation between the term i and j. The value of $\mathrm{Sig_i}$ and $\mathrm{wgt}(t_i,\ t_j)$ in Eq. (5) can be calculated by using Eq. (1) and Eq. (2) respectively. N is the number of the nodes associated with term i.

Then we propose the following algorithm to extract the focal terms and its facet terms.



```
Algorithm
Input: adjacent matrix $A[i,j]$ of the term association network
Output: $HT_i = \{focal\ i: facet_1, facet_2, \ldots, facet_n\}$

For each node $i$ do begin
    For each node j do begin
      $Avg_i = \frac{\sum_{j=1}^{n} \mathrm{wgt}(t_i,t_j)}{N}$;
      $Cen_i = (\mathrm{Sig}_i) * \mathrm{Avg_i}$;
      End
Put each $Cen_i$ in the $array[A]$;
End
Repeat
  For(int $i$=0;$i$<$array[A]$.length; $i$++) do begin
        select max $Cen_i$ in $array[A]$;
        Then take $i$ as a focal term;
          For each node $i$ do begin
                choose its $N$ number of max $wgt(t_i.t_j)$ as its facet terms;
          end
        Then delete $Cen_i$ from $array[A]$;
        Until choose $n$ number focal terms;
        output $HTi = \{focal\ i: facet_1, , facet_2, \ldots, facet_n\}$;
END
```

Since too many "facet terms" would hinder the comprehension of hot topic, after sorting the associations by the weights, we choose the top-ranked *k* hot terms as the "facet terms" of the given "focal term". Then a cluster, composed of a "focal term" and several "facet terms", is extracted from the hot term network. The procedure of hot topic extraction is an iterative process. After a "focal term" and its "facet terms" are extracted, these terms and the associations between them are removed from the hot terms map. This procedure is repeated until the number of hot topics reached a predefined value or there is no unprocessed term in the network.

Applications

To demonstrate the usefulness and effectiveness of our proposed approach, we have applied our approach to two real datasets. In this section, we present the processing details and describe the results from our explorations.

Data collection

Different online learning communities have different characteristics. Discussion forum is a typical learning community where users can interact flexibly, and the discussion transcripts are short and informal with rich oral language. Different form the forums, blogs enable people to convey their opinions freely and other people can interact with them by replying to or commenting on their blogs. This kind of interaction can also enable people to exchange and share ideas, and herein we treat such form as a learning community. The language in blog is fairly lengthy. In order to investigate whether our approach is effective to discover the hot topics and topic evolution in different types of online learning communities, we intentionally chose two real datasets (i.e., educational blog and parent–child forum) from Sina, which is one of the largest portal websites with more than three hundred million users and provides the most popular blog and discussion channel in China.

Dataset A: Educational blog. We crawled blogs and their properties such as publication time and title, from Sina educational blog to launch our research. This dataset contained educational blogs from August 2008 to August 2011 and we collected 2269 blogs in total. The average number of words for a blog is about 500 and the average number of sentences for a blog is about 12.

Dataset B: "parent–child" forum. We crawled the posts and their properties such as publication time and title, from "parent–child" forum. This dataset contained posts from March 2003 to March 2012 and we collected 14,018 posts in total. The average number of words in a post is about 15 and the average number of sentences for a post is about 6.

Process and results

All the documents are divided into several subsets based on their time stamp. In our study, we set every two months as a time slot, which can be adjusted to cater to different needs. Then, the datasets are preprocessed to extract key terms by following three steps. The first step is to adopt Chinese open-source software ICTCLAS (Zhang et al. 2003) to split Chinese words and merge the terms. Then, we remove stop-words, punctuations and numbers and use a constructed synonym list to filter terms. Thirdly, each term's significance is calculated with the extended TF*PDF scheme. Table 1 shows a

Table 1 Top-ranked 15 key terms extracted on different time slots from educational blog

Time	Hot terms
2010.6 ~ 2010.7	(cultivate 0.0275), (high school 0.0704), (score 0.0183), (examination 0.0187), (choice 0.0185), (ability 0.0219), (development 0.0239), (Minhong Yu 0.0187), (time 0.0199), (livelihood 0.0221), (university 0.0543), (China 0.0630), (society 0.0340), (the National College Entrance Examination 0.0475), (job 0.0296)
2010.8 ~ 2010.9	(job 0.0258), (family 0.0231), (habit 0.0215), (grow up 0.0245), (USA 0.0294), (friend 0.0178), (influence 0.0227), (society 0.0388), (development 0.0292), (entrepreneurship 0.0315), (cultivate 0.0445), (university 0.0499), (ability 0.0364), (China 0.057), (livelihood 0.0305)
2011.6 ~ 2011.7	(graduates 0.115), (quality ranking list 0.039), (Shulian Wu 0.0383), (livelihood 0.0373), (ranking list 0.0576), (university 0.0293), (Peking University 0.0464), (the National College Entrance Examination 0.0490), (quality 0.0718), (leader 0.0516), (undergraduate 0.0641), (reform 0.0421), (university graduate quality 0.0413)
2011.8 ~ 2011.9	(Student 0.077), (classmate 0.0264), (content 0.0269), (society 0.0294), (score 0.0454), (development 0.0310), (knowledge 0.0364), (examinee 0.0451), (job 0.0377), (examination 0.0801), (entrepreneurship 0.0740), (university 0.0538), (time 0.0411), (ability 0.0427), (choice 0.0441)

part of extracted key terms from educational blog and Table 2 shows a part of key terms of “parent–child” forum.

After extracting the key terms from each subset, we calculate the weight of association between these key terms. By conducting experiment for several times, we finally set the association threshold as 1.0. Only those related terms with association weight bigger than 1.0 can be retained. Figures 3 and 4 shows the constructed term association network for educational blog and “parent–child” forum at a time slot, respectively. As the figure shows, the thickness of the edge represents the weight of association. The thicker of the edge between two terms is, the stronger association between two terms is, and vice versa. For example, the “The National College Entrance Examination (NCEE)” has strong association with “policy” and “score” in Fig. 3, and “child” has strong association with “English” in Fig. 4.

Based on the term association network, the algorithm is conducted to obtain hot topics by selecting the top-*K* terms as focal terms with related facet terms. Tables 3 and 4 illustrate the extracted hot topics for educational blog and “parent–child” forum, respectively. As shown in the tables, only top-rank 3 focal terms associated with top-rank 4 facet terms are chosen in each time slot.

With these hot topics extracted from every document subset, we can analyze the changing trend of these hot topics over time. Figures 5 and 6 show a part of global view

Table 2 Top-ranked 15 key terms extracted on different time slots from “parent–child” forum

Time	Hot terms
2010.6 ~ 2010.7	(child 0.0327), (activity 0.0766), (dance 0.078), (method 0.0761), (pupil 0.0778), (primary school 0.0793), (examination 0.0820), (training 0.0901), (reading 0. 1503), (family 0.0381), (environment 0.0846), (english 0.2778), (livelihood 0.0983), (student 0.0975), (parent 0.1099)
2010.8 ~ 2010.9	(interest 0.1327), (livelihood 0.0860), (parent 0.2040), (classmate 0.0896), (infant 0.0901), (famous school 0.0875), (psychology 0.0863), (time 0.0942), (homework 0.1084), (attend school 0.1033), (small class 0.1217), (child 0.0991), (attend class 0.0990), (english 0.3275), (class size reduction 0.0938)
2011.6 ~ 2011.7	(company 0.1097), (interest 0.1145), (student 0.1174), (art 0.1225), (subject 0.1173), (train 0.1201), (international 0.1219), (child 0.1308), (Beijing 0.1601), (activity 0.1405), (cultivate 0.1407), (method 0.1203), (family 0.1965), (education 0.1262), (english 0.5619)
2011.8 ~ 2011.9	(reading 0.0691), (interest 0.0683), (child 0.0704), (knowledge 0.0877), (chinese 0.2863), (english 0.3812), (time 0.1091), (like 0.0105), (tutor 0.1344), (attend class 0.0893), (author 0.1108), (student 0.1885), (learning 0.0731), (cultivate 0.0526), (cartoon 0.0309)

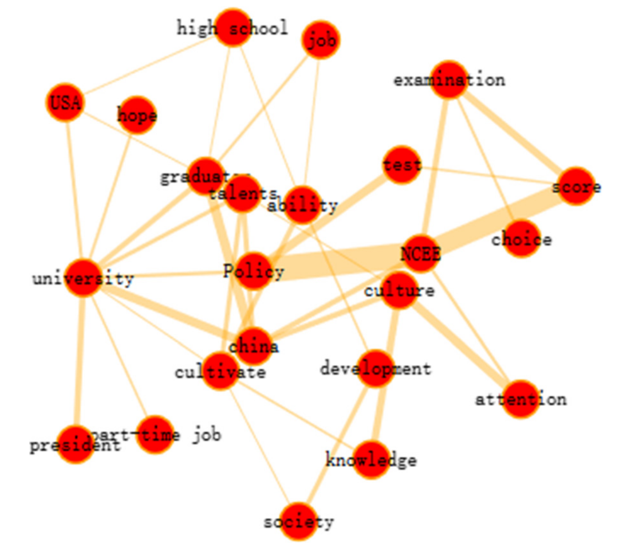


Fig. 3 The term association network for educational blog at June-July 2010

of topic evolution for educational blog and "parent–child" forum, respectively. A hot topic is represented as a hot terms cluster, with the "focal term" in the middle and "facet terms" around it. If two topics have the same "focal term" or one topic's "focal term" is the same as one of the other's "facet terms", we suppose that there is a direct link between the two topics and connect them with a line.

As shown in Fig. 5, we can found that there are three major topics in educational blog, including "NCEE", "University", and "China". The hot topic "NCEE" grows from the facet term to focal term and then declines to facet term from April, 2010 to Sep. 2010, which implies that the discussion focus reveal the key event (i.e., the national college entrance examination in June each year in China) in real world scenario. It is obvious that related topics, such as score, high school, and policy are mostly discussed by people during that period. Notably, there is an interesting finding that "The National

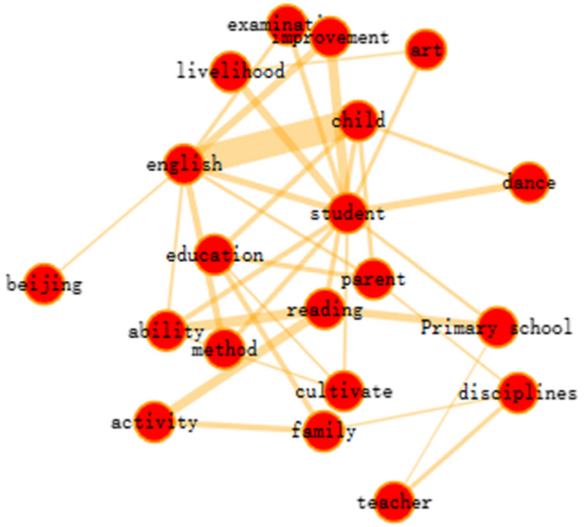


Fig. 4 The term association network for "parent–child" forum at June-July 2010

Table 3 Hot topics extracted from different time slots from educational blog

Time	Topic	
	Focal term	*Facet term*
2010.4 ~ 2010.5	Examination	The National College Entrance Examination, score, high school, time
	China	University, entrepreneurship, family, development
	Job	Talent, situation, teaching, ability
2010.6 ~ 2010.7	China	University, talent, graduates
	The National College Entrance Examination	Policy, score, examination
	Culture	Knowledge, attention
2010.8 ~ 2010.9	China	University, The National College Entrance Examination, entrepreneurship, grow up
	Cultivate	Ability, habit, inference, livelihood
	Society	Hope, attention, development, influence
2010.10 ~ 2010.11	Learning	University, teacher, development, time
	Education	Personality, teaching, optimization, motivation
	Child	Parent, livelihood, education
2010.12 ~ 2011.1	Child	China, school, education, parent
	Parent	Student, learning, teacher, livelihood
	Examination	Score
2011.2 ~ 2011.3	Comprehensive	Major, ranking, evaluation
	Child	Parent, education, learning
	Education	China, child, development, student
2011.4 ~ 2011.5	University	China, major, undergraduate, school
	Child	Parent, learning, education, teacher
	Student	Development, society
2011.6 ~ 2011.7	university	University graduates quality, undergraduate, the National College Entrance Examination, quality ranking list
	China	Ranking list, graduates, society, development
	Quality	Graduates, ranking list
2011.8 ~ 2011.9	Examination	Examinee, choice, classmate, performance
	Entrepreneurship	
	University	Development, society, knowledge

College Entrance Examination” disappears from the hot topic list during the same period in 2011, while “university” is the hot topic instead. In addition, the topic “University” on April, 2011 is split into two topics “China” and “University” with facet terms as “ranking list” and “quality ranking list” on June, 2011. The reason behind that is Shulian Wu, a famous scholar in China, published a list of Chinese university rankings every year around this time. But in 2011, Zhejiang University was ranked the No.1 University over Peking University and Tsinghua University, which are generally recognized as the best two universities in China. Therefore, it sparked a heated debate among the Internet users in China. Another hot topic “Entrepreneurship” emerges on August with no facet terms, which indicates that there is no common discussion points about it.

Regarding the topic evolution pattern in “parent–child” forum as shown in Fig. 6, we can clearly see that there is one major flow that represents the prevalent topic

Table 4 Hot topics extracted from different time slots from “parent–child” forum

Time	Topic	
	Focal term	*Facet term*
2010.4 ~ 2010.5	Time	Activity, tutor, attend class, cultivate
	English	Course, family, interest, Beijing
	Habit	Reading, score, knowledge, ability
2010.6 ~ 2010.7	English	Method, improvement, student, child
	Reading	Activity, ability, primary school
	student	;ivelihood, dance, improvement
2010.8 ~ 2010.9	English	Child, interest, attend class
	Small class	Class size reduction, teaching, development, famous school
	Homework	Attend school, classmate, parent
2010.10 ~ 2010.11	English	Cartoon, training, interest, ability
	Reading	Time, tutor
	Child	Grow up
2010.12 ~ 2011.1	Habit	Child, learning, education, primary school
	Child	Score, teacher, primary school
	Homework	Time, examination, student, mathematics
2011.2 ~ 2011.3	English	Winter vacation, activity, child, homework
	Habit	Method, reading, cultivate, like
	Child	Parent, homework, teacher, learning
2011.4 ~ 2011.5	Student	Attention, activity, child, ability
	School choice	Beijing, haidian district
	Child	Parent, education, learning, art
2011.6 ~ 2011.7	English	Child, subject, Beijing, student
	family	Method, activity, education
	Cultivate	Training, interest, art
2011.8 ~ 2011.9	Student	Attend class, knowledge, time
	English	Cartoon, interest, child, tutorial
	Reading	Chinese language, cultivate, like, composition

“English”. By looking into the facet terms associated with “English”, we can find that the discussion focus is mainly about children’s English interest. This indicates that this topic draws a lot of attention for a long time in “parent–child” forum. Also, we can see that the term “Habit” is extracted and recognized as the main topic in several months, yet the facet terms along with it differs in different months. This could convey comprehensive information concerning the change of hot topics. Furthermore, the hot topic

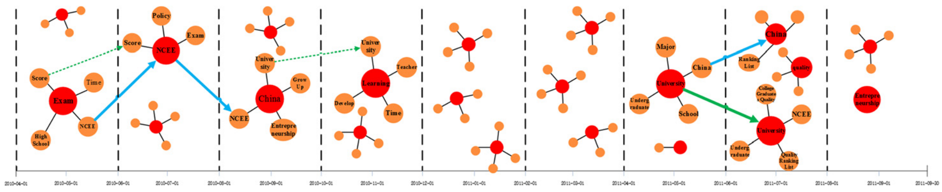


Fig. 5 The topic evolution for educational blog

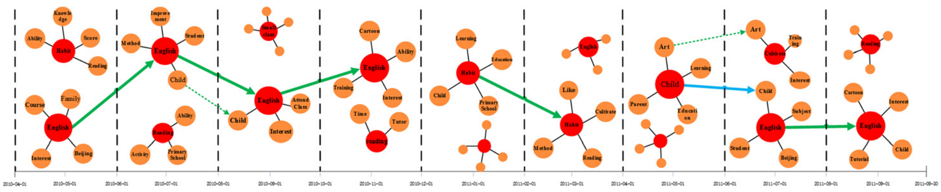


Fig. 6 The topic evolution for "parent–child" forum

"reading" emerges and fades out quickly at different period. Further observation of its facet terms indicates that the discussion focus on "reading" is quite different.

Discussion

The purpose of this study is to propose an approach to automatically extract the hot topics and reveal the topic evolution patterns. Different from other studies, we extract the hot topics composed of focal terms and several facet terms that represent the different aspects of the topic. With the trend analysis, we found that hot topics evolve differently; that is, some hot topics last with the same focus for a long time, some hot topics last but with changing focus; some hot topics emerge and fade out quickly; and some hot topics are split into several hot topics. The evolution pattern makes it easier for users to identify patterns that correlate with merging and branching, yet more in-depth analysis of the internal relationship between hot tops remains to be further examined.

In the previous studies, news, emails and blogs are usually taken as data source for topic detection. Compared with normative text, documents in social networks, such as forums and micro-blogs, exhibit more informal oral language. So we conducted two case studies to verify our approach. The results show that our proposed approach performs satisfactory to extract hot topics for datasets of both the educational blog and the "parent–child" forum. Though the topic evolution pattern of the two datasets differs, the common finding is that the hot topics discovered in the two different types of communities could virtually reflect the situations in the physical world, and thus the visualization of such discovery will help users to easily find out the focus of discussions and better understand of the topic evolution. For instance, regarding the educational blog, we can see a growing and waning interest in NCEE in a certain period, which indicates that the external factors like dates of exams and news stories lead to the emergence of new hot topics, and correspondingly the fading of interests in physical world will cause the decline of the hot topics. As for the "parent–child" forum, the discovered hot topics indicate the lasting interests that parents hold. For example, the hot topics such as "English" and "Habit" remain relatively stable, which implies that the main issues of concern to many Chinese parents are children's early English learning and habit fostering. Further exploration of this observation reveals that the participants in the forum are mainly parents of primary students while the blog authors are some authorized education experts or the ones who tend to express their opinions in education.

One limitation of our approach is that the statistic-based text mining method depends on the number of documents in datasets and the pre-setting of time slot would affect the output of the discovered hot topics. The second weakness is that the selecting threshold is set via several rounds of experiments, which requires adjusting the parameters in different datasets. Moreover, the results are limited by the nature of the sample, since the datasets included only Chinese language content. Further studies are needed to ascertain the applicability of the results for other languages.

Our method has many potential applications in the context of learning. It can help learners to keep track of popular, new, and intertwining topics in online learning communities. With the topic evolution analysis, learners can have a global understanding of the communication focus over time. Furthermore, for the new comers of the online learning communities, it is useful for them to quickly obtain an insight on a certain theme by looking into the hot topics. This would facilitate their involvement in the communities easily and quickly. In practical applications, this method can be integrated to customized recommendation systems that can recommend interesting hot topics associated with learning resources to cater for learners' individual learning demands.

Conclusion

This paper presents a statistic-based text mining method for discovering hot topics and topic evolution in online learning communities. The proposed framework-based representation of hot topics consists of focal term and facet terms that can represent different aspects of the hot topic. By following three steps, i.e., key term extraction, term association analysis, and hot term clustering, hot topics are detected and corresponding evolution patterns over time are discovered by investigating how hot topics emerge, fade and intertwine. Through two case studies, we have demonstrated that our method is effective in extracting hot topics in different types of online learning communities, and the topic evolution can unveil the changing discussion focus and possible key events behind that.

Future work is suggested to improve the approach to address the problem of incremental data analyzing, further explore the topic evolution patterns enriched with critical events, and develop a visualization tool for displaying the topic evolution at different granularities.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors read and approved the final manuscript.

Acknowledgements

This research work is supported by the National Natural Science Foundation of China (NSFC: 61075048) and Special Found for Beijing Common Construction Project.

Received: 15 March 2015 Accepted: 8 July 2015
Published online: 12 August 2015

References

HM Alghamdi, A Selamat, Topic detections in Arabic dark websites using improved vector space model. Proceedings 4th of Conference on Data Mining and Optimization (DMO), 2012,IEEE Computer Society, Langkawi, Malaysia. 6–12 (2012)

RS Bhadoria, M Dixit, R Bansal, AS Chauhan, Detecting and Searching System for Event on Internet Blog Data Using Cluster Mining Algorithm, in *In Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012)* (Springer Berlin Heidelberg, Visakhapatnam, India, 2012), pp. 83–91

KK Bun, M Ishizuka, Topic extraction from news archive using TF* PDF algorithm. In Web Information Systems Engineering, International Conference on. IEEE Computer Society. 73–73 (2002)

KY Chen, L Luesukprasert, SC Chou, Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE Trans on Knowl Data Eng **19**(8), 1016–1025 (2007)

W Cui, S Liu, L Tan, C Shi, Y Song, Z Gao, X Tong, Textflow: Towards better understanding of evolving topics in text. IEEE Trans Vis Comput Graph **17**(12), 2412–2421 (2011)

F Ginter, H Suominen, S Pyysalo, T Salakoski, Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. International journal of medical informatics. **78**(12), e1-e6. (2009)

Q He, Knowledge Discovery through Co-Word Analysis. Library trends. **48**(1), 133–159 (1999)

Y Jahnavi, Y Radhika, Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents. In Advanced Computing Technologies (ICACT), 2013 15th International Conference on. IEEE. 1–6 (2013)

L Kien-Weng Tan, JC Na, YL Theng, Influence detection between blog posts through blog features, content analysis, and community identity. Online Inf Rev **35**(3), 425–442 (2011)

J Kleinberg, Bursty and hierarchical structure in streams. Data Min Knowl Disc **7**(4), 373–397 (2003)

JH Lee, A Segev, Knowledge maps for e-learning. Comput Educ **59**(2), 353–364 (2012)

N Li, DD Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis Support Syst **48**(2), 354–368 (2010)

Y Liu, Y Li, Z Zhang, Designing an Intelligent Interactive Tool for Scaffolding Concept Map Construction Hybrid Learning and Continuing Education: Springer Berlin Heidelberg. 280–289 (2013)

S Nagano, M Inaba, Y Mizoguchi, T Iida, T Kawamura, Ontology-based topic extraction service from weblogs. Conference: Proceedings of the 20th IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA. 468–475 (2008)

T Preethi, K Nirmala Devi, V Murali Bhaskaran, A semantic enhanced approach for online hotspot forums detection. In Recent Trends In Information Technology, 2012 International Conference on. IEEE. 497–501 (2012)

RC Rowe, Abbreviation mania and acronymical madness: private prescription: a thought-provoking tonic on the lighter side. Drug discovery today **8**(16), 732–733 (2003)

Salton, G. Automatic text processing: The transformation, analysis, and retrieval of Reading Addison-Wesley. Addison-Wesley Longman Publishing Co., Inc., Boston, MA. (1989)

XY Wang, Hot Topic Detection in News Blog. Applied Mechanics and Materials **513**, 1114–1118 (2014)

S Woolley, S Ludwig-Hardman. Online learning communities: Vehicles for collaboration and learning in online learning environments. In World Conference on Educational Multimedia, Hypermedia and Telecommunications. World Conference on Educational Media and Technology, 2000 in Montreal, Canada. 1556–1558 (2000)

Y Zhang, N Law, Y Li, R Huang, Automatically extract interpretable topics from online discussion. The 10th International Conference of the Learning Sciences (ICLS 2012), Sydney, Australia, 2-6 July 2012. In ICLS 2012 Proceedings, 2012, v. 1, p. 443-450. (2012)

Z Zhang, Q Li, QuestionHolic: Hot topic discovery and trend analysis in community question answering systems. Expert Systems with Applications **38**(6), 6848–6855 (2011)

D Zhang, S Li, Topic detection based on K-means. IEEE 2011 International Conference In Electronics on Communications and Control (ICECC). Ningbo, China. 2983–2985 (2011)

HP Zhang, Q Liu, ZQ Cheng, H Zhang, HK Yu, Chinese lexical analysis using hierarchical hidden markov model. SIGHAN '03 Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17 Pages 63-70 (2003)

HT Zheng, BY Kang, HG Kim, Exploiting noun phrases and semantic relationships for text document clustering. Inform Sci **179**(13), 2249–2262 (2009)