Expert Systems with Applications 38 (2011) 6848–6855

Contents lists available at ScienceDirect

Expert Systems with Applications

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

QuestionHolic: Hot topic discovery and trend analysis in community question answering systems

Zhongfeng Zhang, Qiudan Li *

Key Laboratory of Complex System and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Keywords:
Community question answering (CQA)
Hot topic
Trend analysis

ABSTRACT

Community question answering (CQA) has recently become a popular social media where users can post questions on any topic of interest and get answers from enthusiasts. The variation of topics in questions and answers indicate the change of users' interests over time. It can help users focus on the most popular products or events and track their changes by exploiting hot topics and analyzing the trend of a specific topic. In this paper, we present a hot topic detection and trend analysis system to capture hot topics in a CQA system and track their evolutions over time. Our system consists of hot term extraction, question clustering and trend analysis. Experimental results using datasets from Yahoo! Answers show that our system can discover meaningful hot topics. We also show that the evolution of topics over time can be accurately exploited by trend graphing.

1. Introduction

With the rapid growth of Web 2.0, community question answering (CQA) has emerged as an alternative information seeking channel where users can post their questions and have their questions answered by other users. Several CQA services, including Yahoo! Answer, Baidu Zhidao, and Naver, have been popular on the internet. For instance, Yahoo! Answers has already attracted millions of users, and stored hundreds of millions of answers to previously asked questions. The rapid growth of user participation in CQA has made it necessary to better understand user behavior and provide better services for users.

In Cao, Duan, Lin, Yu, and Hon (2008) and Duan, Cao, and Yu (2008), a question is considered to be a combination of question topic and question focus. Question topic usually presents major context/constraint of the question characterizing the users' interest. By posting a question on certain topic, the asker shows an interest in it. Question focus presents certain aspect of the question topic. And the topic focus of the question presents certain aspect of the topic. The changes of users' interests in CQA can be reflected by the variation of topics in questions and answers. By exploiting the hot topics and tracking the trend of a specific topic, it would help users focus on the most popular topics and their evolutions. Here, by saying "hot topic", we refer to topics that appear frequently in asked questions during a period of time.

In this paper, we propose a hot topic detection and trend analysis system, which provides a mechanism to help users capture hot topics and track their trends over time in CQA. The system consists of three major components, including hot term extraction, question clustering and trend analysis. Hot term extraction is to extract keywords that appear frequently in questions and can be used to represent topics. A hot term typically have two characteristics: it is used in many questions; the frequency of its usage varies over time. Question clustering is to cluster questions related to a topic into different groups, with a cluster label describing the topic focus of each cluster. With question clustering, users can better understand the topic focus distribution about certain topic, grasping a bigger picture. Moreover, the selected cluster labels provide better description of each cluster, and can help users understand the topic comprehensively and accurately. Finally, the trend analysis tool is developed, and a trend graph indicating question variation of the topics over time is generated. With trend graphing, users can gain an overall view of the topic evolution.

Our specific contributions include:

(1) We propose a novel mechanism which can help users capture hot topics and their evolution trends in CQA.
(2) Based on excellent work in hot topic discovery of news and blogs (Glance, Hurst, and Tomokiyo, 2004; Chen, Luesukprasert, & Chou, 2007), we implement a hot topic discovery and trend analysis system in CQA and evaluate our system with datasets from Yahoo! Answers.

The rest of this paper is organized as follows. In Section 2, we briefly review previous work on CQA and hot topic detection. In Section 3, our system framework and the detailed hot topic

* Corresponding author. Tel./fax: +86 10 62558794.
E-mail address: qiudan.li@ia.ac.cn (Q. Li).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.12.052

discovery algorithm are presented. In Section 4, a system is implemented and evaluated using real-world datasets. Finally, Section 5 concludes the whole paper.

2. Related work

This section reviews works related to this study on CQA, hot topic detection and trend analysis.

2.1. Community question answering

Due to the popularity of Yahoo! Answers and other similar sites, CQA has inspired active research interests in literature. In order to find high-quality answers, questions and users, Bian, Liu, Zhou, Agichtein, and Zha (2009) presented a semi-supervised coupled mutual reinforcement framework for calculating content quality and user reputation. Jeon, Croft, and Lee (2005) presented a machine translation model to find similar questions using the similarity between answers from a CQA service. Bian, Liu, Agichtein, and Zha (2008) developed a ranking system to retrieve relevant and high-quality answers. Jurczyk et al. (2007) showed that user feedback for many topics is sparse, thus is insufficient to reliably identify good answers from the bad ones. They studied the problem of discovering authoritative users by analyzing the link-structure of a general-purpose question answering community. Liu, Bian, and Agichtein (2008) introduced the problem of predicting information seeker satisfaction in CQA communities, which is to predict whether an asker will be satisfied with the answers provided by community participants. They presented a general prediction model and developed a variety of content, structure, and community-focused features for this task. Cao, Duan, Lin, Yu, and Hon (2008) studied the problem of question recommendation. Given a question as a query, they attempted to retrieve and rank other questions according to their likelihood of being good recommendations for the queried question.

The problem of detecting hot topic in a CQA service and tracking their evolution trends has not been cast much attention. In this paper, we focus on this problem and develop a hot topic detection and trend analysis system.

2.2. Hot topic detection and trend analysis

A topic is defined as a seminal event or activity, along with all directly related events and activities (The 2004 Topic Detection & Tracking (TDT2004)). The task of hot topic detection is to exploit topics that appear frequently during a period of time". The problem of topic detection has been studied previously in news (Allan, Carbonell, & Doddington, 1998; Bun & Ishizuka, 2002; Chen et al., 2007), blogs (Glance et al., 2004; Sekiguchi, Kawashima, Okuda, & Oku, 2006; Nagano, Inaba, Mizoguchi, & Iida, 2008) and emails (Kleinberg et al., 2002) etc. Three topic detection methods have been explored, including statistics, linguistics and topic clustering.

In statistics methods, a term-weighting scheme is used to capture the important or representative terms that feature in the content of a document (Chen, Luesukprasert, & Chou, 2007). The most commonly used term-weighting schemes include word frequency, TF * IDF and TF * PDF. As discussed in Bun and Ishizuka (2002), Chen et al. (2007), the TF * IDF scheme emphasizes the importance or uniqueness of each term. It always tries to give significant weight to terms that appear in few documents. However, for hot topic extraction, terms representing hot topics should appear frequently in a large number of documents. To address this problem, Bun and Ishizuka (2002) proposed TF * PDF scheme, which assigns greater weights to terms that occur frequently in many documents and lower weights to those that are rarely mentioned. In this paper, we adopt the TF * PDF scheme for hot term extraction.

Linguistics approaches use the linguistics features of words, sentences and documents. These approaches include the lexical analysis, syntactic analysis, discourse analysis (Zhang et al., 2008). Topic clustering has been studied in topic digital library (TDL) construction and stock market news analysis. It is consisted of three steps, including topic extraction, document clustering and clustering description.

Glance et al. implemented a trend searching system, BlogPulse, for discovering trends across weblogs (Glance, Hurst, & Tomokiyo, 2004). Rajaraman et al. (2001) analyzed the trends of topics being tracked from a stream of text documents. The trends of events in market news were tracked to help understand the influence of these events over the market in Dey, Mahajan, and Haque (2009).

Several online services have been developed for people to understand the hot topics and their trends. With Google Insights or Baidu Index, users can see how frequently the topics have been searched via search engines over time. Twopular (2008) provides users the hot keywords used on Twitter. Through Twopular, users may also compare the usage of different keywords varying over time on Twitter. Microplaza (2009) analyzes a user's network on Twitter, and displays the hot links shared by people he follows.

In this paper, we focus on providing information services for users in CQA by mining hot topics and trend evolution.

3. System architecture

3.1. The overall description

CQA systems are centered on three entities and their relationships, including users ($\mathcal{U}$), questions ($\mathcal{Q}$) and answers ($\mathcal{A}$). Thus, a CQA space can be represented as a tuple: $S = (U,Q,A,Y)$, where Y is a ternary relationship among users, questions and answers, i.e. $Y \subseteq U \times Q \times A$. Each element of Y is called a question thread, which is defined as the combination of a question asked by a user and the answers related to the question.

Recall that a user can post only one question per question thread, while the question can be answered by several users. The asker can select one as the best answer from the answer set. A question thread can be defined as follows:

Definition 1 (*Question thread*). A question thread is represented by a tetrad: (u,q,B_{ua},T_{ua}), where

$u \in U$: the asker of the question;
$q \in Q$: the question, including question title, content, category, and posted date, etc.
$B_{ua} = (u,a)$: the best answer to question q;
$T_{ua} = \{u \in U, a \in A | (u,q,a) \in Y\}$: the list of answers to question q posted by other users.

In our system, the question text is the combination of question title and question content. A question is represented by a classic vector space model, which transforms the question text into a vector space.

Definition 2 (*Question*). A question q is represented as: $q = [w_{1,q}, w_{2,q}, \ldots, w_{N,q}]^{\mathrm{T}}$, where $w_{t,q}$ is the weight of term t in q.

Our system is to extract hot topics and track their trends from question collection in CQA. The architecture of our system is shown in Fig. 1.

First, questions are downloaded from CQA sites. In this paper, we collect questions from Yahoo! Answers. Then, the downloaded questions are extracted and indexed for later process using apache lucene – an open source indexing engine.

Three preprocessing procedures are conducted. First, through tokenization, the question text is split into individual words. Then, a stopword list is used to remove stop words. Third, a stemming process based on Porter's stemming algorithm is followed, to find a semantic representation of an inflected word.

The hot term extraction module extracts hot terms among questions during a certain period, which is the first step of hot topic discovery. As shown in Chen et al. (2007), simply considering the term frequency is insufficient for hot term extraction. Thus, the variant usage of terms is also taken into consideration with time analysis.

Then, the user chooses the interested hot terms which can be used to represent a hot topic. Questions related to the topic are retrieved and clustered. A cluster label is assigned to each cluster describing its topic focus. Finally, trend analysis is performed to compare the variation of the community's interests in related topics.

3.2. Hot topic discovery and trend analysis

In this section, we describe the detailed algorithms used in our system. Following works in hot topic detection and tracking in news stories and blogs (Glance et al., 2004; Chen et al., 2007), a "hot topic" is defined as a topic that is discussed frequently during a period of time. Each hot topic is composed of different topic focuses, with each topic focus presenting a certain aspect of the topic. Intuitively, we expect "Global recession", "caused the recession", "recession affected" represent different topic focuses relating to the topic "recession".

The hot term extraction step makes sure that the extracted hot terms are meaningful, and can be potentially used to represent hot topics. Then, the clustering algorithm is used to group topic related questions into clusters, and for each cluster a label is generated to represent the topic focus of the cluster.

3.2.1. Hot term extraction

Since terms or words are the basic elements of questions, changes in the question contents will be reflected by the variation of the terms' usage. Since a topic is composed of many related events, changes in a topic's popularity are accompanied by variance in the usage of "hot terms". Thus, by identifying "hot terms" that frequently appear in different questions, corresponding hot topics can be accurately identified.

The procedure of hot term extraction is shown in Fig. 2. Two characteristics of a term are considered: the frequency of the term used in the question collection; the variation of its usage over time.

The former characteristic is measured by TF * PDF (Bun & Ishizuka, 2002; Chen et al., 2007), which has been used for topic detection in news stories. The definition is:

Definition 3 (*TF * PDF*). Given a term *j* in a question collection, the TF * PDF is calculated as Bun and Ishizuka (2002):

$$TFPDF_j = \sum_{c=1}^{|C|} |F_{jc}| \exp\left(\frac{n_{jc}}{N_c}\right)$$

$$|F_{jc}| = \frac{F_{jc}}{\sqrt{\sum_{k=1}^{K} F_{kc}^2}}$$

where $TFPDF_j$ is the TF * PDF value for term *j*. F_{jc} is the frequency of term *j* in category *c*. n_{jc} is the number of questions in category *c* in which term *j* appears. *K* is the total number of terms in category *c*; $|C|$ is the number of different categories.

The latter characteristic is calculated by tracking the life cycle of the term. Since tracking all the terms that appear in the question

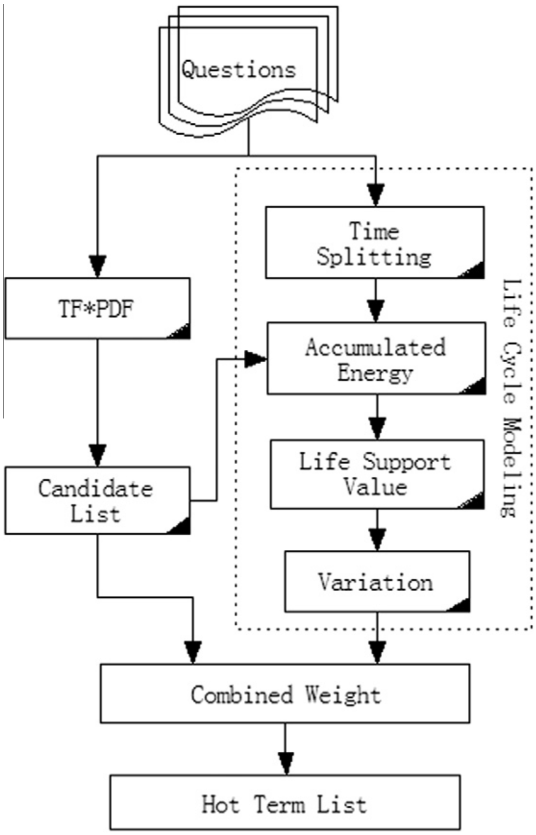


Fig. 2. Hot term extraction procedure from questions.

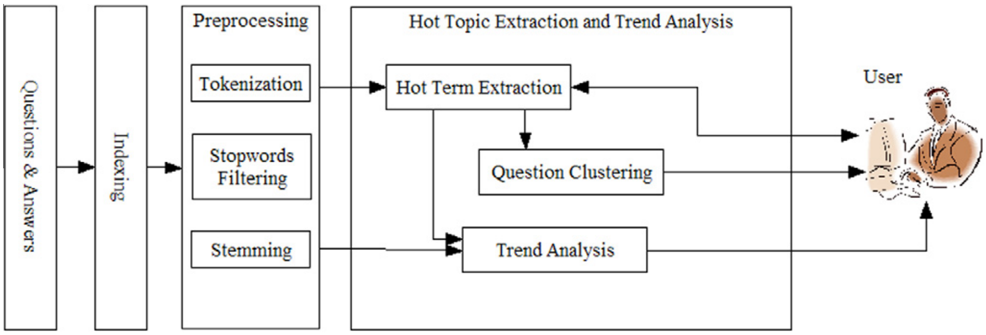


Fig. 1. The architecture of our system.

collection would be computationally expensive and unnecessary, we first construct a candidate list based on the TF * PDF value of terms. We only build life cycle model for terms in the candidate list.

According to the life cycle model (Chen et al., 2007), the accumulated energy $E_{t,s}$ of term t measures the frequency of t appearing in a specified time slot s. The life support value of t at time slot s is calculated as the logarithm of $E_{t,s}$, represented as $lifeSupport_{t,s}$. The variation of its usage over time in the question collection can be computed as:

$V_t = \sqrt{\frac{1}{N}\sum_s(lifeSupport_{t,s} - \overline{lifeSupport})^2}$, where N is the number of time slots, $\overline{lifeSupport}$ is the average of life support value for term t.

The overall weight of term t is measured by combining the two characteristic together with a linear weighted model:

$$weight_t = \lambda^{*}TFPDF_t + (1-\lambda)^{*}V_t(1+\chi_t)$$

where λ is the adjustable weight factor; χ_t is the disagreement of the two characteristics. χ_t can be calculated as: $\chi_t = \frac{FO_t - VO_t}{T}$, where FO_t is the rank of term t by sorting the terms with TF * PDF, VO_t is the rank of term t by sorting the terms with V_t, and T is the number of terms in the candidate list.

By sorting the terms in the candidate list with the combined weight, the top-ranked k terms can be chosen as hot terms. These hot terms reflects the hot topics that people care about most.

3.2.2. Question clustering

With the extracted hot terms, users can choose the interested ones for question retrieval. We then use a clustering algorithm to group these questions. Questions on each topic are grouped into several semantically well structured clusters, with each cluster revealing one aspect of this topic. By carefully selecting cluster labels, users can easily grasp the topic focus of the topic at a higher level without looking through the question list. Thus, performing clustering on topic related questions would help users interact with a CQA community.

In Osiński and Weiss (2005), a concept driven clustering algorithm—Lingo was proposed for search result clustering. This algorithm has been adopted in several clustering search engines (Carrot2, 2002) and showed good performance.

We combine our hot topic discovery module with lingo to cluster topic related questions. Fig. 3 shows the main procedure of the question clustering model:

(1) Question cluster label induction: this step aims to identify the topic keywords from topic related questions. It works as follows: Firstly, identify candidate topic keywords by extracting frequency phrases that appear in the question set. Secondly, the question set is represented by a term-question matrix $\mathcal{Q}$, and cluster-label-candidate matrix P. Thirdly, perform SVD decomposition on $\mathcal{Q}$, such that $\mathbf{Q} = USV'$. Then, perform pruning by $M = U'_k P$, where U_k is the first k columns of $\mathcal{U}$. Finally, compute the cluster label scores based on $\mathcal{M}$, and return labels whose scores are greater than threshold.
(2) Question clustering: in this step, each question is represented by a VSM vector. The cosine similarities between cluster labels and the questions are calculated. For each question, the closest cluster is chosen. Finally, the question clusters are sorted for display based on their group score, which is the product of the cluster label score and the number of questions in this cluster.

From the procedure described above, we can conclude that the question clustering model has several major advantages: unlike the traditional clustering approaches which calculate similarity among questions firstly and then label the discovered groups, it attempts to find good cluster labels firstly and then mark questions with appropriate labels to form groups. By doing this, the question clustering model can focus on finding meaningful cluster labels. It can extract a proper phrase to describe each cluster. As shown in next section, these labels do represent different topic focuses of the topic. Moreover, questions grouped into each cluster substantially relates to its topic focus.

In our system, we implement the question clustering model based on Carrot2 (Carrot2, 2002), which is an open source project for search results clustering engine.

4. System implementation

Based on the above algorithms, we implemented a hot topic discovery and trend analysis system. In this section, we empirically test each component of our system, i.e. hot term extraction, question clustering and trend analysis.

4.1. Question corpus

We downloaded solved questions from Yahoo! Answers from three categories, i.e Elections, Economics and Cameras. Since there are some limitations for Yahoo! Answers applications, we can only collect about 10,000 questions for each category. In total, 24,263 questions are collected. The statistics of each category is listed in Table 1.

After questions are downloaded, fields such as question title, content, asker, ask date stamp etc., are extracted. An inverted index is created for the extracted questions using lucene.

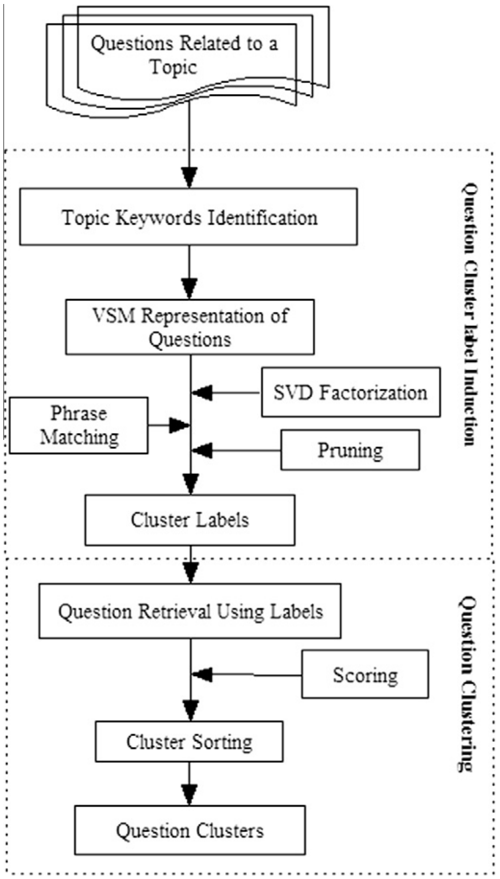

Questions Related to a Topic
Topic Keywords Identification
VSM Representation of Questions
SVD Factorization
Phrase Matching
Pruning
Cluster Labels
Question Cluster label Induction
Question Retrieval Using Labels
Scoring
Cluster Sorting
Question Clusters
Question Clustering

Fig. 3. The procedure of question clustering model.

Table 1
Dataset characteristics for each category.

Category	Date from	Date to	# of questions
Elections	January 25th, 2009	March 3rd, 2009	8,316
Economics	December 1st, 2008	March 3rd, 2009	8,006
Cameras	December 14th, 2008	March 6th, 2009	7,941

Table 2
Hot terms extracted by TF * PDF.

Dataset	Elections	Economics	Cameras
1	Obama	Economi	Camera
2	Peopl	Monei	Pictur
3	Think	Econom	Digit
4	Republican	Peopl	Canon
5	Presid	Price	Len
6	Vote	Good	Card
7	Know	Govern	http
8	Want	Countri	Take
9	Bush	Rate	Good
10	Make	Make	Nikon
11	Democrat	Market	Want
12	Year	Increas	Batteri
13	Time	Help	Help
14	Elect	Year	Look
15	http	Demand	Memori
16	Monei	Think	Know
17	Countri	Product	Work
18	Govern	Bank	Imag
19	Parti	Work	Photo
20	Go	Dollar	Make

Table 3
Hot terms extracted by our system.

Dataset	Elections	Economics	Cameras
1	Obama	Price	Camera
2	Stimulu	Econom	Remov
3	Rush	Peopl	Copi
4	Packag	Monei	Select
5	Bill	Demand	Click
6	Peopl	Curv	Inform
7	Republican	Bank	Mega
8	Million	Tax	Charg
9	Think	suppli	Pictur
10	Black	Depress	Batteri
11	Senat	Unemploy	Card
12	Billion	Loan	Digit
13	Presid	Reserv	Memori
14	Plan	Point	Viewfind
15	Work	Decreas	Pixel
16	Year	Govern	Wide
17	Democrat	Rate	Amazon
18	Monei	Wage	Len
19	Fail	Unit	Reader
20	Make	Industri	File

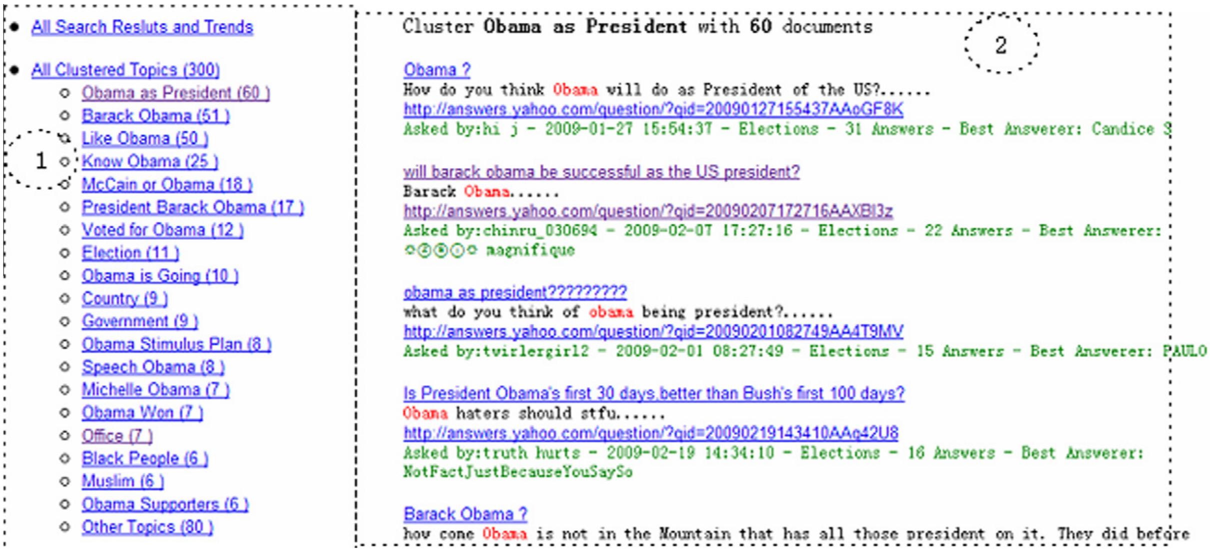
4.2. Hot term extraction

Table 2 shows the extracted hot terms using TF * PDF only for each dataset. Table 3 shows the extracted hot terms using our algorithm described in Section 3.2. We can see that more name entities are discovered by our algorithm, and hot terms in Table 3 seem to be more suitable to represent topics. The noisy terms, such as "make", "year", "good" etc. are down weighted or removed from the hot term list by considering the temporal information.

4.3. Question clustering

Fig. 4 shows the system interface for the hot term "Obama" in the Elections dataset. It is split into two parts with dashed rectangles. Part 1 is the cluster result for the top 300 questions about "Obama". Each cluster is represented by a cluster label, followed by the number of questions in this cluster. These clusters represent different aspects of the topic "Obama". In this example, questions talking about "Obama" are clustered into "Obama as President" (which indicates that Obama is USA president), "Barack Obama" (The full name of Obama), "Like Obama" (discussing whether people like their new president Obama), etc. Part 2 lists all questions grouped into the cluster "Obama as President". We can see that almost all the questions are asking for opinions about Obama as new USA president.

Table 4 lists the cluster labels for hot term "recess" in the Economics category. Here, "recess" is a stemming form for the term "recession", which refers to the economic recession spreading across the world. From these cluster labels, it can be concluded that questions about economic recession cover many aspects of this topic: "Economy is in a Recession", "Economic Recession", etc. are questions talking about the current economic situation; "Recession Affected", "Impact on the Economy", etc. are questions discussing the recession impacts on people's daily lives and economy; "Economy got so Bad", etc. are questions about the anxieties



- All Search Resluts and Trends
- All Clustered Topics (300)
  - Obama as President (60)
  - Barack Obama (51)
  - Like Obama (50)
  - Know Obama (25)
  - McCain or Obama (18)
  - President Barack Obama (17)
  - Voted for Obama (12)
  - Election (11)
  - Obama is Going (10)
  - Country (9)
  - Government (9)
  - Obama Stimulus Plan (8)
  - Speech Obama (8)
  - Michelle Obama (7)
  - Obama Won (7)
  - Office (7)
  - Black People (6)
  - Muslim (6)
  - Obama Supporters (6)
  - Other Topics (80)

Cluster **Obama as President** with **60** documents

Obama ?
How do you think Obama will do as President of the US?......
http://answers.yahoo.com/question/?qid=20090127155437AAoGF8K
Asked by:hi j - 2009-01-27 15:54:37 - Elections - 31 Answers - Best Answerer: Candice S

will barack obama be successful as the US president?
Barack Obama......
http://answers.yahoo.com/question/?qid=20090207172716AAXBI3z
Asked by:chinru_030694 - 2009-02-07 17:27:16 - Elections - 22 Answers - Best Answerer: ○◎◎○○ magnifique

obama as president?????????
what do you think of obama being president?......
http://answers.yahoo.com/question/?qid=20090201082749AA4T9MV
Asked by:twirlergirl2 - 2009-02-01 08:27:49 - Elections - 15 Answers - Best Answerer: PAULO

Is President Obama's first 30 days better than Bush's first 100 days?
Obama haters should stfu......
http://answers.yahoo.com/question/?qid=20090219143410AAq42U8
Asked by:truth hurts - 2009-02-19 14:34:10 - Elections - 16 Answers - Best Answerer: NotFactJustBecauseYouSaySo

Barack Obama ?
how come Obama is not in the Mountain that has all those president on it. They did before

Fig. 4. Question clustering for hot term "Obama".

Table 4
Cluster labels for questions about "recess".

Cluster Labels	# of Related Questions
Economy is in a recession	79
Economic recession	55
Recession depression	51
Recession going	33
Recession just	25
World recession	22
Current recession	20
Long is this recession	16
Global recession	15
Explain	14
Government	14
Caused the recession	12
Recession affected	12
Recession crisis	12
Economy got so bad	11
Difference between a recession and a depression	8
Recession right	8

and fears that this economic recession brought to the world; "Difference between a Recession and a Depression" lists questions trying to distinguish the two economic terms recession and depression, and so on.

From the above two examples, we can conclude that our system can successfully mine the topic focus distribution for a specified hot topic. It can help users focus on the interested subjects conveniently. For instance, in the "recession" case, users can browse through the clusters "Economic Recession" etc. to gain information about the current economic situation. On the other hand, the economic professionals can focus on "Economic got so Bad", "Recession Affected" etc. to analyze people's reactions toward the economic recession, allowing them to take proper actions to cope with the recession.

4.4. Trend analysis

Trend analysis iterates the same topic for all dates within a specified data range, and bins the counts into time buckets and plots the result. By trend graphing, we aim to offer users an overall view of the topic distribution over time. We also implemented trend search which can compare trends among several different topics. Points in the graph indicate the number of questions asked about a topic within a time bin (one day by default).

Fig. 5 displays the trend graphs for the two major political parties– "republican" and "democratic" during Febraury 2009. Fig. 6 shows the Google Trends (2006) search for the same subjects. We can conclude that our findings are consistent with that from Google search.

Figs. 7 and 8 show trends search by our system and Google Trends for four famous cameras brands—Canon, Nikon, Sony, Kodak—during Febraury 2009. Their rankings in terms of CQA question hit count is different from their rankings in terms of Google Trends. "Sony" ranks lower than Nikon in our system, although it has higher market share. But it ranks highest in Google Trends. This

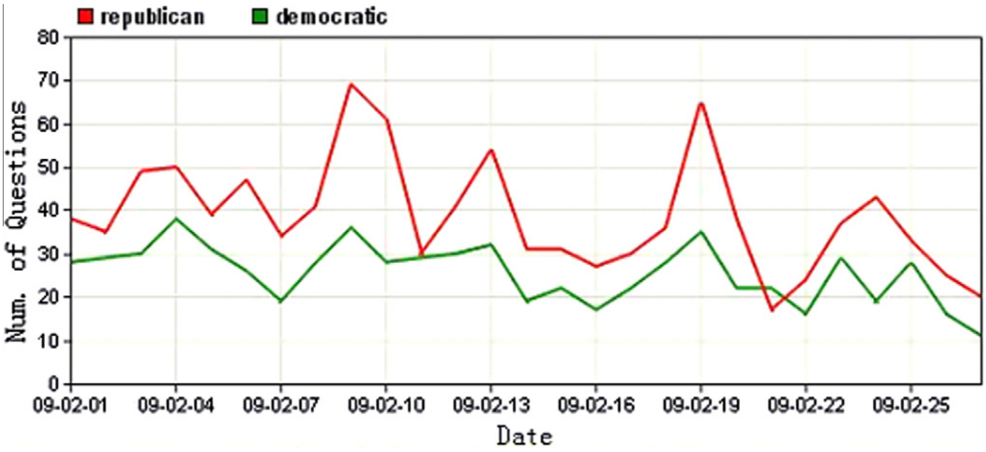


Fig. 5. Trend search for "republican" and "democratic".

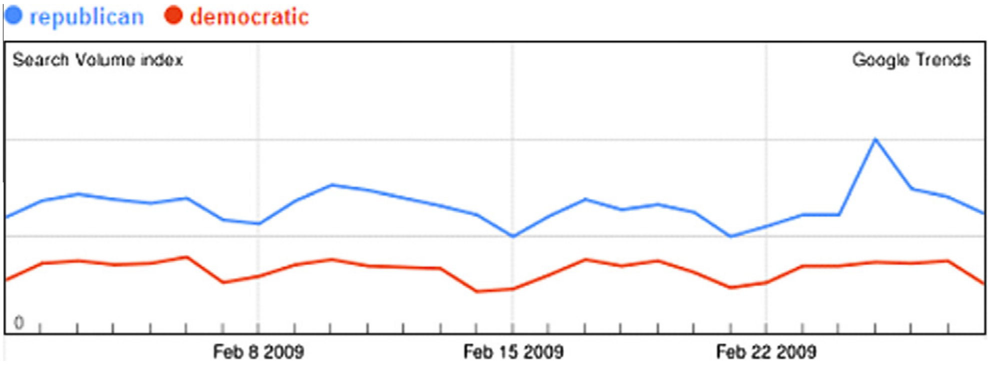


Fig. 6. Google Trends for "republican" and "democratic".

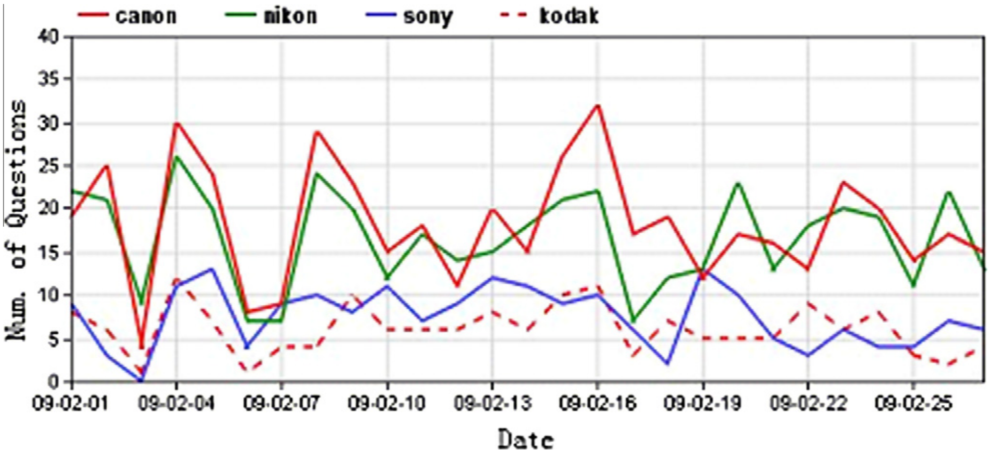


Fig. 7. Trend search for four cameras brands.

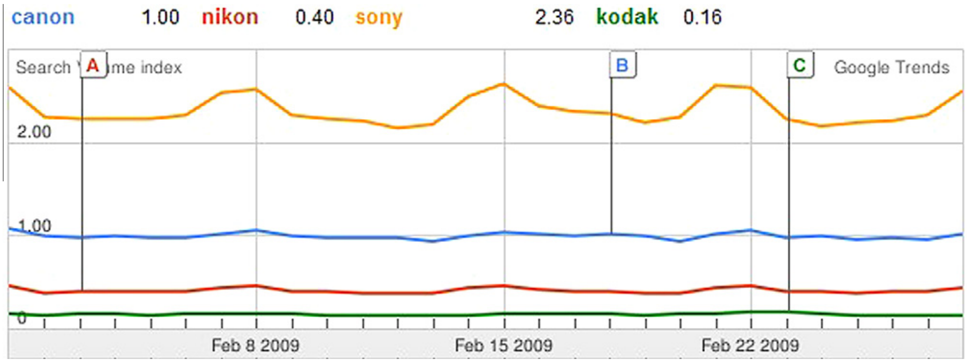


Fig. 8. Google Trends for four cameras brands.

may be because that besides cameras, Sony is also a successful brand in many other electronic products. Google Trends made no distinction between these different products, whereas our system only focuses on cameras. Apart from this, Fig. 7 corresponds to market shares of these four camera brands, i.e. canon gains the biggest market share, while Kodak the lowest.

With trend graphing, the evolution of topics is quite straight forward. In a sense, the trend variation for a brand reveals its market share, as shown in Fig. 7. By comparing question trends between their own products and products from competitors, it can help manufactures master the current market situation, such that the manufactures can adjust their market strategy in time to achieve better sale performances. Moreover, by clustering questions about a product, manufactures can conveniently grasp the market demands and analyze consumers' opinions about their products. This information provides basis for them to improve their products and services.

5. Conclusions and future work

In this paper, we propose a hot topic discovery mechanism for CQA services. A hot topic discovery and trend analysis system is implemented to illustrate the effectiveness and feasibility of our mechanism. It can be used as a platform to test new algorithms and new ideas. As shown in Section 4, our system can discover interesting topics and has potential applications.

In this paper, we only take the questions into consideration for hot topic discovery. As part of our ongoing and further research, the question qualities and users are being exploited for hot topic discovery. Sometimes, a hot term may fail to describe a topic, due to the ambiguity of a single term. We are exploring to represent a topic with a set of keywords using probabilistic topic models, which may better describe a topic. The evolution relationships among topics during different time periods are also under consideration.

Acknowledgments

This research is supported by the projects 863 (No. 2006AA010106), 973 (No. 2007CB311007), NSFC (No. 60703085).

References

Allan, J., Carbonell, J., & Doddington, G. (1998). Topic detection and tracking pilot study: Final report. In *Proceeding of the DARPA broadcast news transcription and understanding workshop* (pp. 194–218). San Francisco, CA: Morgan Kaufmann Publishers, Inc..

Bian, J., Liu, Y., Agichtein, E., & Zha, H. (2008). Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of WWW* (pp. 467–476).

Bian, J., Liu, Y., Zhou, D., Agichtein, E., & Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of WWW* (pp. 51-60).

Bun, K. K., & Ishizuka, M. (2002). Topic extraction from news archive using TF * PDF algorithm. In *3rd International conference on Web information systems engineering, (WISE)* (pp. 73–82).

Cao, Y., Duan, H., Lin, C., Yu, Y., & Hon, H. (2008). Recommending questions using the MDL-based tree cut model. In *Proceedings of the 17th International World Wide Web Conference, (WWW)* (pp. 81-90).

Carrot2 Project: (2002) <http://project.carrot2.org/>.

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

Dey, L., Mahajan, A., & Haque, S.M. (2009). Document clustering for event identification and trend analysis in market news. In *Proceedings of ICAPR* (pp. 103–106).

Duan, H., Cao, Y., Lin C., & Yu, Y. (2008). Searching questions by identifying question topic and question focus. In *Proceedings of ACL* (pp. 156–164).

Glance, N., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for Weblogs. In *The workshop on the Weblogging ecosystem at the 13th international world wide Web conference, WWW*.

Google Trends, 2006. http://www.google.com/trends.

Jeon, J., Croft, W. B., & Lee, J. H. (2005). Finding similar questions in large question and answer archives. In *Proceedings of CIKM* (pp. 84–90).

Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities using link analysis. In *Proceedings of CIKM* (pp. 919–922).

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of SIGKDD* (pp. 91–101).

Liu, Y., Bian, J., & Agichtein, E. (2008). Predicting information seeker satisfaction in community question answering. In *Proceedings of SIGIR* (pp. 483–490).

Microplaza: (2009) <http://microplaza.com/public>.

Nagano, S., Inaba, M., Mizoguchi, Y., & Iida, T. (2008). Takahiro Kawamura: Ontology-based topic extraction service from Weblogs. In *IEEE international conference on semantic computing, (ICSC)* (pp. 468–475).

Osiński, S., & Weiss, D. (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems, 20*, 48–54. May/June, 3.

Rajaraman, K., Tan, A. (2001). Topic detection, tracking, and trend analysis using self-organizing neural networks. In *Proceedings of PAKDD* (pp. 102–107).

Sekiguchi, Y., Kawashima, H., Okuda, H., & Oku, M. (2006). Topic detection from blog documents using users' interests. In *The 7th international conference on mobile data management, (MDM)* (p. 108).

The 2004 Topic detection and tracking. (TDT2004). Task definition and evaluation plan, <http://www.nist.gov/speech/tests/tdt/>.

Twopular: (2008) <http://twopular.com/>.

Zhang, C., & Zhang, Q. (2008). Topic navigation generation using topic extraction and clustering. In *International symposium on knowledge acquisition and modeling* (pp. 333–339).