Knowl Inf Syst (2015) 43:389–416
DOI 10.1007/s10115-014-0737-z

REGULAR PAPER

Learning to suggest questions in social media

**Tom Chao Zhou · Michael Rung-Tsong Lyu ·
Irwin King · Jie Lou**

Received: 16 June 2013 / Revised: 12 January 2014 / Accepted: 17 January 2014 /
Published online: 4 March 2014

Abstract Social media systems with Q&A functionalities have accumulated large archives of questions and answers. Two representative types are online forums and community-based Q&A services. To enable users to explore the large number of questions and answers in social media systems effectively, it is essential to suggest interesting items to an active user. In this article, we address the problem of question suggestion, which targets at suggesting questions that are semantically related to a queried question. Existing bag-of-words approaches suffer from the shortcoming that they could not bridge the lexical chasm between semantically related questions. Therefore, we present a new framework, and propose the topic-enhanced translation-based language model (TopicTRLM), which fuses both the lexical and latent semantic knowledge. This fusing enables TopicTRLM to find semantically related questions to a given question even when there is little word overlap. Moreover, to incorporate the answer information into the model to make the model more complete, we also propose the topic-enhanced translation-based language model with answer ensemble. Extensive experiments have been conducted with real-world datasets. Experimental results indicate our approach is very effective and outperforms other popular methods in several metrics.

T. C. Zhou (✉)
Baidu Inc., Shenzhen, China
e-mail: zhouchao03@baidu.com; tom.chaozhou@gmail.com

M. R.-T. Lyu · I. King
Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China

M. R.-T. Lyu · I. King
Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: lyu@cse.cuhk.edu.hk

I. King
e-mail: king@cse.cuhk.edu.hk

J. Lou
Department of Information Systems, City University of Hong Kong, Kowloon Tong, Hong Kong
e-mail: paggy922@gmail.com

Keywords Social media · Online forum · Community-based Q&A · Question suggestion · Language model · Topic modeling

1 Introduction

With the inception of question answering (Q&A) services in social media, very large archives of questions and their answers have been collected over time in social media systems. Typical social media systems with Q&A functionalities include community-based question answering (Q&A) services and online forums. Community-based Q&A services, such as Yahoo! Answers, Baidu Knows, and Quora, are online communities that adopt the Web 2.0 model and organize knowledge exchange in the form of asking and answering questions [34]. Different from traditional FAQs that focus on factoid and specific domain questions [63], community-based Q&A services address various users' needs, including information seeking, recommendation [69,70] and socializing [57]. They complement the mainstream search engines by enabling users to obtain straightforward answers to their natural language questions [1,3]. It is reported that 10 questions and answers are posted per second in Yahoo! Answers.[1] About 218 million questions have been solved in Baidu Knows.[2] An online forum is a Web application, which involves highly interactive and semantically related discussions on domain specific questions, such as travel, sports, and programming. Questions are usually the focus of forum discussions and a natural means of resolving issues [56]. A travel forum TripAdvisor has 45 million reviews until 2011.[3] Previous research efforts show that mining forum knowledge in the form of Q&A pairs could improve forum management [17]. Over times, a large amount of historical Q&A pairs have been built up in community-based Q&A and online forum archives, providing information seekers viable alternatives to general purpose Web search [7].

To utilize the large archives of questions and their answers, *question search* (Jeet05,; [14]), a functionality facilitating users to search previous questions with answers is usually provided along with the social media systems. Typically, question search is conducted by first retrieving questions that are semantically equivalent to a queried question and then returning the related answers to users [15]. For example, given the queried question Q1 in Table 1, question Q2 can be returned for question search. However, a user may not be aware that his/her query may only capture one aspect of the particular topic he/she is interested in. For example, given question Q1 in Table 1, he/she may not be aware that existing questions Q3 and Q4 in Table 1 can satisfy his/her information needs more thoroughly by complementing Q2. Under these circumstances, it is necessary and desirable to suggest semantically related questions. Thus, in this paper, we propose a new functionality for Q&A in social media, named *question suggestion*, a functionality facilitating a user to explore a topic he/she is interested in by suggesting semantically related questions to a queried question.

Performing question suggestion in social media systems has three benefits: (1) helping users explore their information needs thoroughly from different perspectives; (2) increasing page views by enticing users' clicks on suggested questions to increase potential revenues; and (3) providing social media systems a relevant feedback mechanism by mining users' click through logs to improve search quality.

[1] http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/.

[2] http://zhidao.baidu.com/.

[3] http://www.prnewswire.com/news-releases/tripadvisor-grows-and-grows-and-grows-119678844.html.

Table 1 Examples of question search and question suggestion

Query:

Q1: How is Orange Beach in Alabama?

Question search:

Q2: Any ideas about Orange Beach in Alabama?

Question suggestion:

Q3: Is the water pretty clear this time of year on Orange Beach?

Q4: Do they have chair and umbrella rentals on Orange Beach?

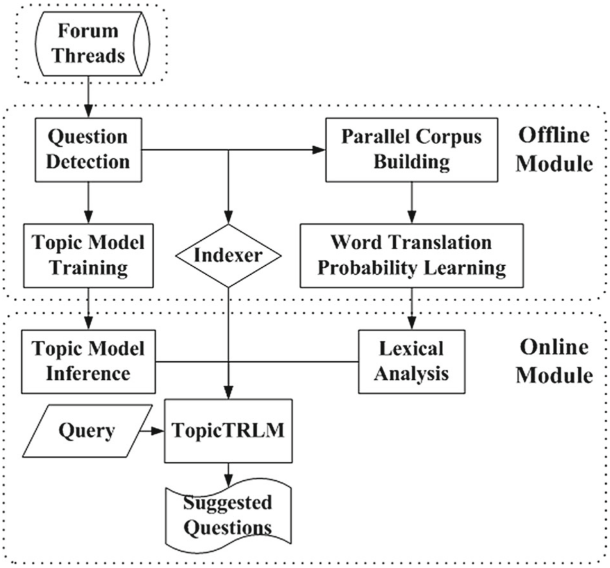


Fig. 1 System framework of question suggestion in online forum

Most existing methods in question search only employ bag-of-words approaches with lexical knowledge, failing to bridge the lexical chasm between semantically related questions [15,57]. Our previous work studies question suggestion in online forums [67], but not in other social media systems with Q&A functionality. To the best of our knowledge, none of the existing work studies the problem of learning to suggest questions in social media systematically by fusing lexical knowledge with latent semantic knowledge.

In this article, we propose to address the problem of learning to suggest questions in social media systems. Specifically, we focus on two representative types of social media systems with Q&A functionality, community-based Q&A services and online forums. We first propose an effective question suggestion framework in online forums as shown in Fig. 1. This framework consists of three major steps: (1) detecting questions in forum threads; (2) learning word translation probabilities from questions in forum threads; and (3) calculating semantic relatedness between a queried question and a candidate question using topic-enhanced translation-based language model (TopicTRLM). In the proposed framework, we utilize interactive nature of forum threads to learn word translation probabilities and fuse both the lexical and latent semantic knowledge to calculate the semantic relatedness between two questions.

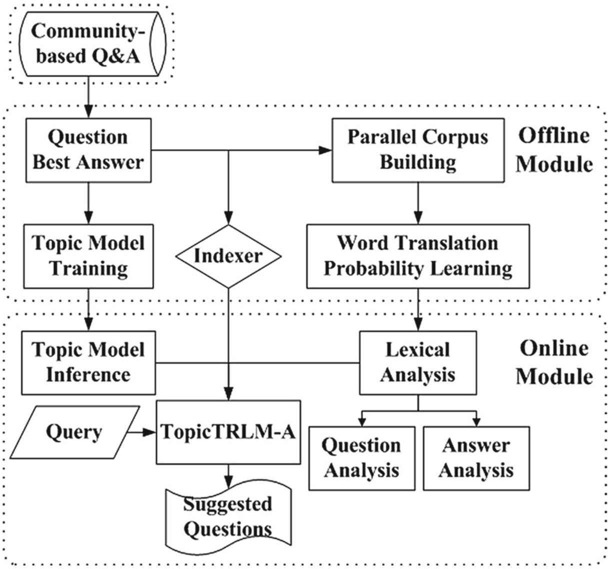


Fig. 2 System framework of question suggestion in community-based Q&A

We also propose a question suggestion framework in community-based Q&A. In community-based Q&A sites, a best answer is readily available for each resolved question. The best answer is chosen by the asker or voted by community users. The framework is shown in Fig. 2. This framework consists of two modules: offline module and online module. In the offline module, we utilize questions and their best answers to learn word translation probabilities and train topic models. In the online module, we compute semantic relatedness between a queried question and a candidate question using topic-enhanced translation-based language model with answer ensemble (TopicTRLM-A).

We empirically verify the effectiveness of proposed models in TripAdvisor and Yahoo! Answers. TripAdvisor is a popular online forum that attracts a large number of discussions about hotels, traveler guides, etc. Yahoo! Answers is a renowned community-based Q&A service around the world. Experimental results show that our approach outperforms previous methods in many metrics.

The rest of the paper is organized as follows. We start with a review of the studies related to automatic Q&A, proliferation of community-based Q&A and online forums, followed by summarizing studies investigating question search and latent topic modeling. Then, we present the proposed approaches to question suggestion in online forums and community-based Q&A services. This is followed by a comprehensive experimental analysis and evaluation. Conclusion and future work are discussed finally.

2 Related work

2.1 Automatic question answering (Q&A) from the web

Automatic Q&A has been a long-standing research problem, which attracts contributions from the information retrieval and natural language-processing communities. Automatic Q&A ranges from automatic subjective Q&A [32,58,71] to automatic factual Q&A [19,21,24]. Most work of retrieving answers directly from the Web focus on factual Q&A.

To set up a baseline for factual Q&A on the Web, how successful search engines are at retrieving accurate answers when unmodified factual natural language questions are asked was studied [48]. An architecture that augments existing search engines so that they support natural language Q&A was developed [47]. To guide future system development, a specialized Q&A test collection was constructed for research purpose [33]. To surmount the barrier of question understanding, the concept of a query language, which provides an intermediate system for capturing the essence of a user's information need and matching that information need to the desired items in a repository of texts, was introduced [55]. In the absence of a standard query language across search engines, words were suggested to be added to the question to guide the search process [2]. Determining question taxonomy is another critical component of process of machine understanding of questions. Five question taxonomies were identified at four levels of linguistic analysis [43]. Semantic enrichment of texts was studied to improve factual Q&A [41].

However, *the existing methods for automatic Q&A do not utilize readily available Q&A pairs in social media as they just extract answers directed from the Web for questions*. In contrast, question suggestion that we address in this article does utilize readily available Q&A pairs in online forums and community-based Q&A services and suggest semantically related questions.

2.2 Proliferation of community-based Q&A services and online forums

With the popularization of social media systems with Q&A functionalities, people have come together to post their questions, answer other users' questions and interact with each other. Community-based Q&A services and online forums are two representative platforms for this purpose. Overtimes, a large amount of historical Q&A pairs have been gathered in their archives, providing information seekers viable alternatives to Q&A from the Web [1,17,27].

Social cues were shown important for continuation of Q&A activity online [46]. Six classes of relevance criterion were identified for selecting best answers in community Q&A [31]. The perceived importance of relevance, quality and satisfaction in contributing to a good answer was explored [54]. An approach based on structuration theory and communities of practice that could guide investigation of dynamics of community Q&A was proposed [53]. A review and analysis of the research literature in social Q&A was conducted [22]. The motivational factors affecting the quantity and quality of voluntary knowledge contribution in community-based Q&A services was investigated [34–36].

However, *the existing methods for studying proliferation of community-based Q&A services and online forums do not consider how to utilize the large number of Q&A pairs in social media for other applications as they just investigate the user behavior*. In contrast, question suggestion that we address in this article is a useful application that does utilize readily available Q&A pairs.

2.3 Question search

Question search aims at finding semantically equivalent questions for a user question. Addressing the lexical chasm problem between user questions and the questions in a Q&A

archive is the focus of most existing work. Berger et al. [4] studied four statistical techniques for bridging the lexical chasm, which include adaptive TFIDF [50], automatic query expansion [39], statistical translation models [5] and latent semantic models [26]. The history of question search originated from FAQ retrieval. The FAQ Finder combined lexical similarity and semantic similarity between questions to rank FAQs, where a vector space model was employed to compute the lexical similarity, and the WordNet [38] was utilized to capture the semantic similarity [12]. Riezler et al. [51] proposed a translation model for question search in FAQ. Their translation model was trained on a large amount of data extracted from FAQ pages on the Web.

Recently, question search has been re-visited with the Q&A data in social media, which mainly includes community-based Q&A services and online forums. Jeon et al. [28,29] employed translation model to tackle the question search problem in community-based Q&A. The translation model, proposed by Berger et al. [5], has been extensively employed in question finding and answer retrieval [4,20,30,51]. Realizing that the translation model may produce inconsistent probability estimates and make the model unstable, Xue et al. [64] proposed translation-based language model which balanced between language model and translation model.

Learning monolingual word-to-word translation probability is the most essential part of solving the lexical gap problem in translation-based models. It can be obtained by training statistical translation models on parallel monolingual corpora. IBM model 1 was commonly employed to learn the translation probabilities [9]. Jeon, Croft and Lee considered question–question pairs as a parallel corpus if their answers are similar [28,29]. Xue et al. [64] treated question–answer pairs as a parallel corpus. Bernhard and Gurevych [6] proposed to use a parallel training dataset of the definitions and glosses provided for the same term by different lexical semantic resources.

Besides translation models, other approaches were also investigated. Bian et al. [7] proposed a learning framework for factual information retrieval within the community-based Q&A data. Particularly, they modeled the retrieval problem as one of learning ranking functions and then introduced an algorithm called GBRank for learning the ranking functions from a set of labeled data. Duan et al. [20] summarized questions in a data structure consisting of question topic and question focus, and performed question search based on tree-cut model. Question topic and question focus was also utilized to perform question clustering, which improves question search [15]. Cao et al. [13,14] exploited category information of questions for improving the performance of question search. Specifically, they applied the approach to vector space model, Okapi BM25 model, language model, translation model and translation-based language model. A systematically evaluation of the performance of different classification methods on question topic classification was studied [45]. Wang et al. [61] proposed a syntactic tree-matching approach instead of a bag-of-word approach to find similar questions. Cao et al. [16] represented each question as question topic and question focus using tree-cutting approach and employed minimum description length (MDL) for question recommendation. In our previous work, we proposed to suggest semantically related questions by learning from interactive and semantically related discussions on various questions from online forums [67].

However, *most existing methods of question search only find equivalent questions instead of semantically related questions. A few methods targeting at finding related questions either only utilize lexical information or just focus on only one type of social media system*. In contrast, we conduct a systematic analysis on how to suggest semantically related questions by utilizing both lexical and latent topic modeling across different social media systems.

2.4 Latent topic modeling

Representing the content of text documents is a critical part of any approach to information retrieval. Typically, documents are represented as a bag-of-words, meaning that the words are assumed to occur independently. To capture important relationships between words, researchers have proposed approaches that group words into topics. Word clustering, for example, was studied in Sparck Jones [59]. The well-known latent semantic indexing (LSI) technique was introduced in Deerwester et al. [18]. More recently, Hoffman [26] described the probabilistic latent semantic indexing (pLSI) technique. This approach used a latent variable model that represents documents as mixture of topics. Using topic models for document representation has also recently been area of considerable interest in machine learning. Latent Dirichlet Allocation (LDA) [8], which was used to build topic models based on a formal generative model of documents [25], was heavily cited in the machine learning literature due to its property of possessing fully generated semantics. Wei et al. [62] proposed a LDA-based document model within language modeling framework for ad hoc retrieval. Ramage et al. [49] employed latent topic models to automatically clustering web pages into semantic groups so as to improve search and browsing on the Web. To the best of our knowledge, this is the first work to systematically investigate question suggestion task by incorporating latent topic modeling across different social media systems. We believe our work is an important step toward understanding how latent topic modeling could help question suggestion.

3 Question suggestion framework

In this section, we present our approach of learning to suggest questions in online forums and community-based Q&A services. Online forums and community-based Q&A services are two different types of social media systems with Q&A functionality. Table 2 summarizes characteristics and differences between two types of systems.

From Table 2, we can see that online forums and community-based Q&A services have different characteristics. Thus, we propose two different frameworks for learning to suggest questions in two systems. Details would be discussed in subsequent sections.

Table 2 Characteristics and differences of online forums and community-based Q&A services

Social media	Number of questions in each thread	Answer directly available for each question	Question readily available in a system
Online forums	Multiple	No. There is no structural indicator of answers for each question. Forum posts are also mixed with spams, chitchat, etc	No. questions are mixed with other posts in each thread. Question detector is needed to identify each question
Community-based Q&A Services	Single	Yes. There is structural indicator of answers for each question	Yes. Each thread is organized around a question

Table 3 An Example of forum discussions

TS: anyone has any suggestion on lounges/clubs/live band in gulfshores/orange beach during memorial weekend?

TF: The Pink Pony in Gulf Shores always has a band. Lulu's has live music, but they close at 10 PM. Lester's on Canal Road, VIP Lounge on Canal Road. There's always the FloriBama, which can have as many as three or four bands on at one time

TS: Also, any suggestions on nice restaurants?

TF: As for restaurants...Mango's is nice, the restaurant in Perdido Beach Resort is nice, The Beach club restaurant is very nice

TS: Could it be Voyagers or Cafe Palm Breeze at the Perdido?

TF: Voyager's, that's it. Nice atmosphere. Okay food. High prices

Thread starter (TS) means people who initiates a thread, and thread follower (TF) means people who replies in a thread

3.1 Question suggestion in online forums

In online forums, a discussion thread usually originates from a root post by the thread starter [17]. Table 3 gives an intuitive description of a discussion thread, in which bold sentences are questions asked by the thread starter.

From Table 3, we can see that all the questions asked by the thread starter are semantically related and center around different aspects of a discussion topic, such as "travel in orange beach" in this example. In addition, the interactive nature of forum discussions ensures that questions in a forum thread are very likely to explore different aspects of a topic.

In this section, we start by introducing the method of question detection in online forums; then, we explain TopicTRLM for measuring semantic relatedness of two questions, followed by learning word translation probabilities in online forums. To make this article self-contained, we briefly introduce LDA.

3.1.1 Question detection in online forums

Questions are usually the focus of forum discussions and a natural means of resolving issues. But in online forums, questions are mixed with normal forum posts. To identify questions in online forums, we adopt the method proposed by Cong et al. [17] for question detection since that method can achieve both high recall and high precision. For details, please refer to Cong et al. (2008)'s paper.

3.1.2 Topic-enhanced translation-based language model

Two types of methods are typically used to represent the content of text documents. One is the bag-of-words representation, which means that words are assumed to occur independently. A bag-of-words model is a fine-grained representation of a text document. The other method to represent text documents is topic model. Topic model assigns a set of latent topic distributions to each word by capturing important relationships between words. Comparing with bag-of-words representation, topic model is a coarse-grained representation for documents.

We investigate applications of a series of statistical translation models, which are typical bag-of-words approaches in the Q&A research. The results are summarized in Table 4.

From Table 4, we could see that statistical translation models were successfully employed to bridge the lexical chasm between two semantically equivalent questions or to capture the

Table 4 Investigation of applications of statistical translation models

Berger et al. [4]	Employing statistical translation models to inspect a collection of questions and answers from FAQ documents, in order to characterize the relation between questions and answers
Riezler et al. [51]	Presenting an approach to query expansion in answer retrieval that uses statistical machine translation techniques to bridge the lexical gap between questions and answers in FAQ pages
Jeon et al. [28,29]	Employing the translation model to tackle the question search problem in community-based Q&A services
Xue et al. [64]	Proposing translation-based language model which balanced between language model and translation model to tackle the question search problem in community-based Q&A services
Bernhard and Gurevych [6]	Investigating several methods to train monolingual translation probabilities so as to solve the lexical gap problem in answer finding
Cao et al. [13,14]	Exploiting category information of questions on vector space model, BM25 model, language model, translation model and translation-based language model to improve the performance of question search

Table 5 Investigation of applications of latent topic models

Wei and Croft [62]	Using LDA to improve ad hoc retrieval by smoothing the probabilities in the document model
Phan et al. [42]	Employing latent topic modeling on a large-scale external data to help build a classifier
Ramage et al. [49]	Employing latent topic models to automatically clustering web pages into semantic groups so as to improve search and browsing on the Web
Rosen-Zvi et al. [52]	Learning author-topic models to describe authors' interests and papers

relations between questions and answers. Thus, the statistical translation model is a nice module to find lexically related questions for a queried question.

We also investigate applications of a series of latent topic models, and we summarize results in Table 5.

From Table 5, we may find that latent topic models are commonly used to cluster semantically related documents. However, these cluster are relatively coarse-grained representations.

As for the question suggestion application, suggested questions should be semantically related to the queried question, and they should explore different aspects of a discussion context with respect to the queried question. Based on our investigations, fine-grained bag-of-words representation of question would contribute to finding lexically similar questions under the similar contexts, and topic model representation would contribute to finding related questions that explore different aspects under the similar contexts. To achieve the goal of adopting both bag-of-words and topic model representations, we propose the TopicTRLM model. It fuses the latent topic information with lexical information to measure the semantic relatedness between two questions systematically. Specifically, we employ the translation-based language model (TRLM) to measure the semantic relatedness of bag-of-words representations of two questions and employ Latent Dirichlet Allocation (LDA) to calculate the latent topics' similarities between two questions.

Equation (1) shows TopicTRLM approach to calculate the semantic relatedness of a queried question and a candidate question:

$$\begin{aligned}&\mathrm{P(q|D)} = \textstyle\prod_{w\in q} P(w|D),\\ &\mathrm{P(w|D)} = \gamma * P_{\text{trlm}}(w|D) + (1-\gamma)P_{\text{lda}}(w|D),\end{aligned} \tag{1}$$

where q is the queried question, D is a candidate question, w is a query term in q. The first line of Eq. (1) means we employ a unigram representation for each question, and this is a common approach in Q&A research. To measure the relatedness between two questions q and D, we only need to multiply the probability of $P(w|D)$ for all w in q. Unigram representation is a simple yet effective approach in many tasks [7,14,64]. $P_{\text{trlm}}(w|D)$ is the TRLM score, and $P_{\text{lda}}(w|D)$ is the LDA score. TRLM score captures the contribution of bag-of-words representation, and LDA score considers the topic model representation. Equation (1) employs Jelinek-Mercer smoothing [66] to fuse the TRLM score with LDA score, and γ is the parameter to balance the weights of bag-of-words representation and topic model representation. The benefit of combining two parts is that the suggested questions for a queried question would be lexically similar to some extent, meaning they are questions under the same context; these questions would also potentially explore different aspects under the same context, meaning they are topically related. A larger γ means that we would like to find more lexically related questions for the queried question; a smaller γ would emphasize more on two questions' latent topic distributions' similarity. When we set $\gamma = 0$, TopicTRLM only employs latent topic analysis, and when we set $\gamma = 1$, TopicTRLM only employs lexical analysis. Thus, TopicTRLM is a generalization of both lexical analysis and latent topic analysis in the question suggestion task. Equation (2) describes TRLM which employs Dirichlet smoothing:

$$\begin{aligned}&\mathrm{P}_{\text{trlm}}(w|D) = \tfrac{|D|}{|D|+\lambda} P_{\text{mx}}(w|D) + \tfrac{\lambda}{|D|+\lambda} P_{\text{mle}}(w|C),\\ &P_{\text{mx}}(w|D) = \delta P_{\text{mle}}(w|D) + (1-\delta)\textstyle\sum_{t\in D} T(w|t)P_{\text{mle}}(t|D),\end{aligned} \tag{2}$$

where $|D|$ is the length of the candidate question, C is the question collection extracted from the forum posts. λ is the Dirichlet smoothing parameter to balance the collection smoothing and empirical data. The reason we employ the collection smoothing is that we may better capture the relatedness between a word w and a question D by considering the large amount of collection statistics. In addition, if the word w does not appear in D, the term $P_{\text{trlm}}(w|D)$ would be zero if we do not employ the collection smoothing. If we increase λ, then we would rely more on smoothing. Dirichlet smoothing has the advantage that for longer candidate questions, its smoothing effect would be smaller. δ is the parameter to balance between language model and translation model. A larger δ would have the effect to retrieve lexically similar questions. A smaller δ would have the effect to retrieve lexically related questions. $T(w|t)$ is the translation probability from source word t to target word w, $P_{\text{mle}}(\bullet)$ is the maximum-likelihood estimation. An essential part of TRLM is to learn the word-to-word translation probabilities $T(w|t)$, which would be discussed later. Equation (3) describes employing LDA to calculate the similarity between a query term w and a candidate D:

$$\mathrm{P}_{\text{lda}}(w|D) = \sum_{z=1}^{K} P(w|z)P(z|D), \tag{3}$$

where K is the number of latent topics, and z is a latent topic. The physical meaning of Eq. (3) is that there is a probability distribution of topics z for the question D, and there is a probability distribution of words w for the topic z. After employing the Bayes equation and summation rule, we could get the similarity between a query term w and a candidate question D.

3.1.3 Learning translation probability in online forums

Learning word-to-word translation probabilities is the most essential part in TRLM. IBM model 1 [9] is employed to learn the translation probabilities, and a monolingual parallel corpus is needed. The construction of the parallel corpus should be tailored to the specific task. To find similar questions, three kinds of approaches are employed previously to build parallel corpus: (1) question and question pairs are considered as a parallel corpus if their answers are similar [28,29], (2) question and answer pairs are considered as a parallel corpus [64], and (3) question and its manually labeled question reformulation pairs are considered as a parallel corpus [6]. However, neither of above three methods is suitable to build the parallel corpus for the question suggestion task in forums. The reason is that the presence of spam within the discussion forum would make all questions subjected to the same spam appear equivalent. To build a parallel corpus for learning word-to-word translation probabilities for question suggestion, we turn to investigating the properties of forum discussions. Because questions are usually the focus of forum discussions and a natural means of resolving issues, questions posted by a *thread starter* during the discussion are very likely to explore different aspects of a topic. It is very likely that these questions are semantically related. Thus, we propose to utilize these semantically related questions posted by the thread starter in each thread to build the parallel corpus. The procedure of generating a parallel corpus of related questions from forums is as follows: (1) extract questions posted by the thread starter in a thread, and create a question pool Q; (2) construct question–question pairs by enumerating all possible combinations of question pairs in the Q; (3) repeat step (1) and (2) for each forum thread; and (4) build the parallel corpus by aggregating all question-question pairs constructed from each forum thread.

3.1.4 Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) [8], as a topic model method that possesses fully generative semantics, has attracted a lot of interests in the machine learning field. The graphical model of LDA is shown in Fig. 3.

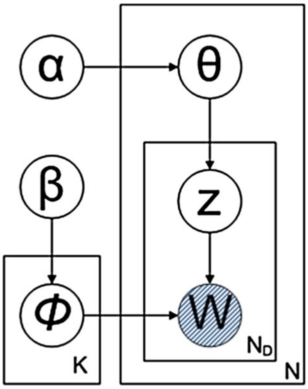


Fig. 3 Graphical model of LDA. N is the number of documents; N_D is the number of words in document D; K is the number of topics

The process of generating a corpus in the smoothed LDA is as follows: (1) pick a multinomial distribution φ_z for each topic z from a Dirichlet distribution with parameter β; (2) pick a multinomial distribution θ_D from a Dirichlet distribution with parameter α for each question D; (3) pick a topic $z \in \{1, ..., K\}$ from the multinomial distribution θ_D for each word token w in question D; (4) pick word w from the multinomial distribution φ_z.

We calculate the semantic relatedness between a query word w and a candidate question D as follows:

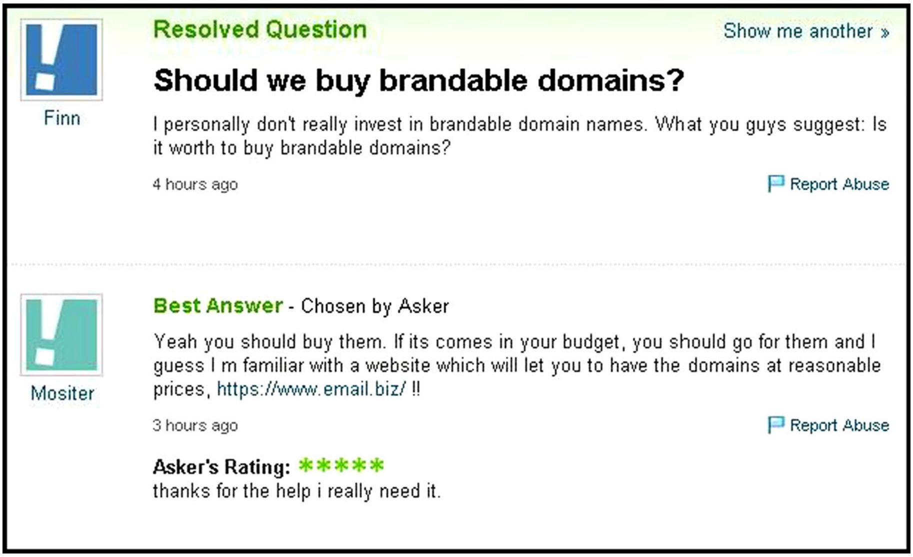
$$\mathrm{P}_{\mathrm{lda}}(w|D, \theta, \varphi) = \sum_{z=1}^{K} P(w|z, \varphi) P(z|\theta, D), \tag{4}$$

where θ and φ are the posterior. The physical meaning of Eq. (4) is that there is a probability distribution of topics z for the question D, and there is a probability distribution of words w for the topic z. After employing the Bayes equation and summation rule, we could get the similarity between a query term w and a candidate question D. We employ Gibbs sampling to directly obtain the approximation of θ and φ because the LDA model is quite complex and cannot be solved by exact inference [23]. In a Gibbs sample, φ is approximated with $(n_{-i,j}^{(w_i)} + \beta_{w_i}) / \sum_{v=1}^{V} (n_{-i,j}^{(v)} + \beta_v)$, and θ is approximated with $(n_{-i,j}^{(D_i)} + \alpha_{z_i}) / \sum_{m=1}^{M} (n_{-i,m}^{(D_i)} + \alpha_m)$ after a certain number of iterations being accomplished. $n_{-i,j}^{(w_i)}$ is the number of instances of word w_i assigned to topic $z = j$, not including the current token. α and β are hyperparameters that determine how heavily this empirical distribution is smoothed. $n_{-i,j}^{(D_i)}$ is the number of words in document D_i assigned to topic $z = j$, not including the current token. The total number of words assigned to topic $z = j$ is $\sum_{v=1}^{V} n_{-i,j}^{(v)}$. The total number of words in document D not including the current one is $\sum_{m=1}^{M} n_{-i,m^{(D_i)}}$. Based on these derivations, we rewrite Eq. (3) as Eq. (5):

$$\mathrm{P}_{\mathrm{lda}}(w|D) = \sum_{z=1}^{K} \frac{n_{-i,j}^{w_i} + \beta_{w_i}}{\sum_{v=1}^{V} (n_{-i,j}^{(v)} + \beta_v)} \times \frac{n_{-i,j}^{D_i} + \alpha_{z_i}}{\sum_{m=1}^{M} (n_{-i,m}^{(D_i)} + \alpha_m)}. \tag{5}$$

3.2 Question suggestion in community-based Q&A services

Community-based Q&A services, such as Yahoo! Answers, are question-centric, in which users are socially interacting by engaging in multiple activities around a specific question [1,71]. Thus, we do not need to perform question detection as we do in online forums. When a user asks a new question, he/she also assigns it to a specific category, within a predefined hierarchy of categories, which should best match the question's general topic. The new question remains "open" for four days with an option for extension or for less if the asker chose a best answer within this period. Registered users may answer a question as long it is "open." If after this time period the question remains unresolved, its status changes from "open" to "in-voting," in which users can only vote for a best answer till a clear winner arises [68]. Thus, a best answer is always available for a resolved question either chosen by the asker or voted by communities. Most community-based Q&A services allow users to write a "question title" to describe their questions in one sentence, and write a "question detail" to elaborate their question in detail. An example of a resolved question in Yahoo! Answers is shown in Fig. 4. In it, we can find that a question detail is provided along with the question title to help elaborate the background, and a best answer is chosen by the asker for this resolved question.



Resolved Question

Should we buy brandable domains?

Finn

I personally don't really invest in brandable domain names. What you guys suggest: Is it worth to buy brandable domains?

4 hours ago

Best Answer - Chosen by Asker

Mositer

Yeah you should buy them. If its comes in your budget, you should go for them and I guess I m familiar with a website which will let you to have the domains at reasonable prices, https://www.email.biz/ !!

3 hours ago

Asker's Rating: *****
thanks for the help i really need it.

Fig. 4 Example of a resolved question in Yahoo! Answers with question title, detail and best answer

After investigating the characteristics in community-based Q&A services, we introduce TopicTRLM-A for measuring questions' relatedness in community-based Q&A services. After that, we discuss the method of learning words' translation probability in community-based Q&A services.

3.2.1 Topic-enhanced translation-based language model with answer ensemble

Since the best answer for each resolved question in community-based Q&A services is always readily available, we propose to incorporate it into our model and propose the TopicTRLM-A. The intuition is that the best answer of a question could also explain the semantic meaning of the question. Thus, when we measure the semantic relatedness of a queried question and a candidate question, we also consider the semantic relatedness between the queried question and the best answer of a candidate question. The mathematical equation of TopicTRLM-A is shown in Eq. (6):

$$\begin{aligned}&P(q|(Q,A)) = \textstyle\prod_{w\in q} P(w|(Q,A)),\\ &P(w|(Q,A)) = \varepsilon P_{\text{trlm}}(w|(Q,A)) + (1-\varepsilon)P_{\text{lda}}(w|Q),\end{aligned} \tag{6}$$

where q is a queried question, (Q, A) is a candidate question with its best answer, w is a word in the queried question. The first line of Eq. (6) means that we employ the unigram representation to calculate the relatedness between q and (Q, A). $P_{\text{trlm}}(w|(Q, A))$ is the lexical score, and $P_{\text{lda}}(w|Q)$ is the latent semantic score. ε is a parameter to balance lexical score and latent semantic score. If we set a large ε, we would reply more on lexical score, if we set a small ε, we would reply more on latent semantic score. The fusion of two parts provides the possibility of finding semantically related questions for the queried question under the similar context. Equation (7) presents the details of lexical score calculation:

$$\begin{aligned}&P_{\text{trlm}}(w|(Q,A)) = \tfrac{|(Q,A)|}{|(Q,A)|+\lambda} P_{\text{mx}}(w|(Q,A)) + \tfrac{\lambda}{|(Q,A)|+\lambda} P_{\text{mle}}(w|C),\\ &P_{\text{mx}}(w|(Q,A)) = \eta P_{\text{mle}}(w|Q) + \theta \textstyle\sum_{t\in Q} T(w|t) P_{\text{mle}}(t|Q) + \mu P_{\text{mle}}(w|A),\end{aligned} \tag{7}$$

where we employ the Dirichlet smoothing between the candidate question and the collection. $|(Q, A)|$ is the length of a candidate question with its best answer. If we set a large λ, we would have a larger smoothing effect. If a term w does not appear in Q and A, the result of the first line would be zero. Adding collection smoothing could avoid this situation. We employ translation-based language model on the question part, and incorporate the best answer using language model. The reason we use language model on the best answer part is that the best answer could capture the meaning of the question to some extent. By adding a contribution from the answer part, we may better capture the semantic relatedness between a q and the (Q, A). η, θ and μ are parameters to represent weights on each part, where $\eta + \theta + \mu = 1$.

3.2.2 Learning translation probability in community-based Q&A services

Learning word translation probability in community-based Q&A services is an important part of TopicTRLM-A model. Different from online forums, there is usually only one question in each thread in community-based Q&A services. After observing the real-world data, we find the question detail is usually a reformulation of the corresponding question title. Thus, we aggregate question title and question detail as a monolingual parallel corpus, and employ IBM model 1 to learn the word translation probabilities.

4 Experiments and results

In this section, we describe the experiments we conducted to test our novel models. We conducted experiments on both TripAdvisor and Yahoo! Answers, to demonstrate the effectiveness of the proposed models in online forums and community-based Q&A services separately. TripAdvisor is a popular forum attracts many discussions on travel related topics, and Yahoo! Answers is one of the most renowned community-based Q&A services. In each part, we first describe our experimental setup, including the dataset we used, metrics we employed and methods we compared. Then, we present the results of our experiments and analyses that shed more light on the performance of our models.

4.1 Experiments in online forums

We consider the question suggestion task as a retrieval task in our experiments. We aim to address three research questions on question suggestion in online forums:

RQ1: How effective is the proposed method to learn the word-to-word translation probabilities in online forums?

RQ2: How is TopicTRLM compared with other approaches on labeled questions in question suggestion task in online forums?

RQ3: How is TopicTRLM compared with other approaches on the joint probability distributions' similarity of topics with ground truth?

4.1.1 Experimental setup

Methods: To evaluate the performance of the proposed methods, we compared the proposed algorithms with alternative approaches. Specifically, we compared our method TopicTRLM with LDA [8], query likelihood language model using Dirichlet smoothing (QL) [66], translation model (TR) [28,29] and the state-of-the-art question search method translation-based language model (TRLM) [64].

Dataset: TripAdvisor is a popular online forum that attracts a large number of discussions about hotels, traveler guides, etc. TripAdvisor forum consists of a large number of threads, which contain posts from thread starters and other participants. For evaluation purpose, we crawled data from the travel forum TripAdvisor. The crawling process was conducted from the thread level. We employed the same settings with Cong et al. [17] to mine LSPs, and the classification-based question detection method was reported to score 97.8 % in Precision, 97.0 % in Recall, and 97.4 % in F_1-score.

After employing the question detection method in crawled data, we randomly sampled 300 questions, we removed questions that are not comprehensible, e.g., "*What to see?*" is not a comprehensible question; while "*How is the Orange Beach in Alabama?*" is a comprehensible question. Finally, we obtained 268 questions. We used the unigram language model to represent questions, and applied IBM model 1 to learn unigram to unigram translation probabilities. We deployed Porter Stemmer [44] to stem question words. We adopted the stop word list used by SMART system [10], but 5W1H words were removed from the stop word list. For each model, the top 20 retrieval results were kept. We used *pooling* [37] to put results from different models for one query together for annotation, and all models were used in the pooling process. If a returned result was considered as semantically related to the queried question, it was labeled with "relevant"; otherwise, it was labeled with "irrelevant." Two assessors were involved in the initial labeling process. If two assessors had different opinions on a decision, a third assessor was asked to make a final decision. The kappa statistics between two assessors was 0.74. This test set was referred to as "TST_LABEL."

We tried to create a reasonable ground truth data without involving laborious manual labeling. Thus, we assumed that questions posted by the same user in a thread were related. We built the unlabeled testing dataset by randomly selecting threads until there were 10,000 threads that contained at least two questions posted by thread starters. The first question in each thread was treated as the queried question. This test set was referred to as "TST_UNLABEL."

The remaining questions, referred to as "TRAIN_SET," were used in three purposes: (1) building parallel corpus to learn the word-to-word translation probabilities, (2) LDA training data and (3) question repository to retrieve questions to offer question suggestion service. TRAIN_SET contained 1,976,522 questions extracted from 971,859 threads. We conducted a detailed analysis on the TRAIN_SET to acquire a deeper understanding of the forum activities.

This paper leveraged thread starters' activities in forums, so we first conducted a post level analysis on thread starters' activities. The statistics is shown in Table 6.

From Table 6, we can see thread starters replied on average 1.9 posts to the thread he or she initiated, and this indicates our expectation that forum discussions are quite interactive. We also plotted the distribution of replied posts from thread starter in Fig. 5, and this distribution follows a power law distribution.

We also conducted a question level analysis on thread starters' activities. Table 7 presents statistics of question level activities of thread starter. We found over 68.8 % thread starters asked on average 2 questions in each thread. These findings supported our motivation that

Table 6 Statistics of post level activities of thread starter (TS)

#Threads	#Threads that have replied posts from TS	Avg. # replied posts from TS
1,412,141	566,256	1.9

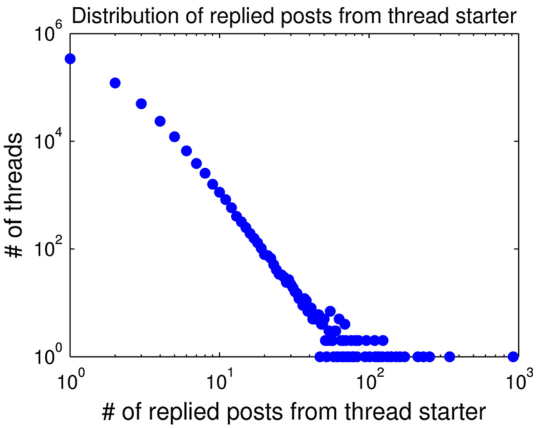


Fig. 5 Post level distribution of thread starters' activity

Table 7 Statistics of question level activities of thread starter (TS)

#Threads	#Threads TSs' posts contain questions	Avg. # questions in TSs' posts
1,412,141	971,859	2.0

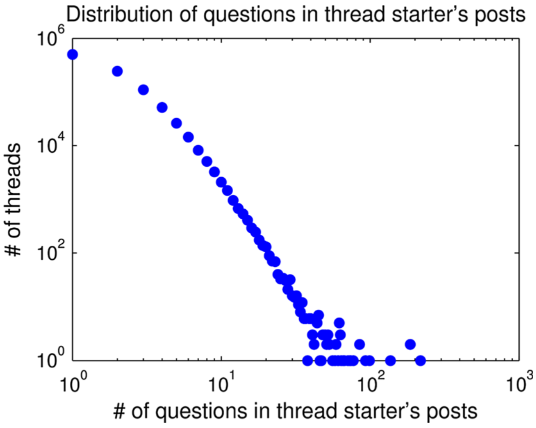


Fig. 6 Question level distribution of thread starters' activity

question is a focus of forum discussions, and forum data is an ideal source to train the proposed model for question suggestion.

Figure 6 depicts a view of distribution of questions in thread starter's posts. We can see this distribution also follows a power law distribution.

We used 5 as the threshold for sequential pattern mining in question detection. In this paper, we mined LSPs by considering both minimum support threshold and minimum confidence threshold. We empirically set minimum support at 0.5 % and minimum confidence at 85 % using a development corpus. Each discovered LSP formed a binary feature as the input for a classification model. We used a rule-based classification algorithm Ripper to perform the question detection task. These parameter values are set empirically according to the paper [17]. We used GIZA++ [40] to train the IBM model 1. We used GibbsLDA++ [42] to conduct LDA training and inference.

Table 8 The first row shows the source words

Words rank	Shore		Park		Condo		Beach	
	IBM 1	LDA	IBM 1	LDA	IBM 1	LDA	IBM 1	LDA
1	Shore	Shore	Park	Park	Condo	Condo	Beach	Beach
2	Beach	Groceri	Drive	Hotel	Beach	South	Resort	Slope
3	Snorkel	Thrift	Car	Stai	Area	North	What	Jet
4	Island	Supermarket	How	Time	Unit	Shore	Hotel	Snowboard
5	Kauai	Store	Area	Area	Island	Pacif	Water	Beaver
6	Condo	Nappi	Where	Recommend	Maui	Windward	Walk	Huski
7	Area	Tesco	Walk	Beach	Rent	Seaport	Area	Steamboat
8	Water	Soriana	Time	Nation	Owner	Alabama	Room	Jetski
9	Boat	Drugstor	Ride	Tour	Shore	Opposit	Snorkel	Powder
10	Ocean	Mega	Hotel	Central	Rental	Manor	Restaur	Hotel

Top 10 words that are most semantically related to the source word are presented according to IBM translation model 1 and LDA. All the words are lowercased and stemmed

Metrics: For the evaluation of the task, we adopted several well-known metrics that evaluate different aspects of the performance of the proposed method, including Precision at Rank R ($P@R$), mean average precision (MAP), mean reciprocal rank (MRR) and Kullback-Leibler divergence (KL-divergence). Reciprocal rank is an accepted measure in Q&A evaluation. It favors hits that are ranked higher, however, gives appropriate weights to lower ranked hits [60].

Parameter tuning: There are several parameters need to be determined in our experiments. We used 20 queries from the TST_LABEL and employed MAP to tune the parameters. We chose $\alpha = 50/\text{K}$, $\beta = 0.1$ in LDA estimation like in Phan et al. [42], and K is the number of topics in LDA. We set $K = 200$ according to the previous study [67]. We ran 200 iterations to estimate the LDA model and ran 30 iterations to do LDA inference. The parameter λ controls how much we incorporate collection smoothing into the empirical observation, and we set $\lambda = 2{,}000$ according to the existing work [66]. The parameter δ controls the weights between the language model and the translation model, and we set $\delta = 0.2$ empirically according to the previous paper [64]. The parameter γ balances the weights between the lexical score and the latent topic score. We ran a grid search on [0, 1] with a step size to be 0.1; then, we selected the value that maximizes the MAP for 20 queries. So we set $\gamma = 0.7$ after our grid search. In summary, optimal parameters are as follows: $\alpha = 0.25$, $\beta = 0.1$, $K = 200$, $\lambda = 2{,}000$, $\delta = 0.2$ and $\gamma = 0.7$.

4.1.2 Experiment on word translation

To answer RQ1, we used the proposed method to build the parallel corpus, and the constructed parallel corpus contains 2,629,533 question-question pairs. Table 8 shows the top 10 words that are most related to the given words under certain conditions after employing IBM model 1 and LDA.

Various relatedness between words were discovered using IBM model 1. For example, when a user is asking a question about *shore*, *snorkel* is related because *snorkeling* is a popular activity in *shore*, and *condo* is also related because the user also needs to rent a *condo* for living. *Walton* is a beach name in Florida's Emerald Coast near *Pensacola* and *Destin*. Its

Table 9 Comparison on labeled questions (a larger metric value means a better performance)

Metrics	LDA	QL	TR	TRLM	TopicTRLM
P@R	0.2411	0.3370	0.4135	0.4555	0.5140
MAP	0.3684	0.4089	0.4629	0.5029	0.5885
MRR	0.5103	0.5277	0.5311	0.5317	0.5710

full name is *Fort Walton Beach*. *Atlanta* is also related to *Walton* because the nearest Airport of *Walton* provides frequent flights to *Atlanta*. Recall that the proposed method considers that questions in a thread could translate to each other, leading to capturing the relatedness of words from related questions. In other words, it characterizes relations in related events that happen in related questions. We could find that LDA captures different relations, and the reason is that LDA describes co-occurrence relations because it considers words in a question. For example, people ask questions such as "*Is there any grocery store at Orange Beach?*," and LDA is capable of capturing this kind of word relations between *grocery* and *beach* in a sentence. Thus, we believe both approaches capture different relatedness between words.

4.1.3 Experiment on labeled question

We conducted an experiment on TST_LABEL to answer RQ2. We employed the word-to-word translation probabilities learnt from the parallel question–question corpus in TR, TRLM and TopicTRLM. The experimental results on metrics P@R, MAP and MRR are shown in Table 9. All the results are statistically significant according to the sign test compared with the previous method.

Table 9 shows that LDA performs the worst. Because LDA is a coarse-grained representation to measure the relatedness between questions, it is not able to capture accurate meaning of each question. TR has better question suggestion performance compared with QL. This finding is consistent with the previous work [28,29]. The reason is that the translation model has the potential to bridge the lexical chasm between related questions. It also confirms the effectiveness of the proposed method to build parallel corpus of related questions from forum thread. TRLM has better performance than TR because TR set the probability of self-translation to 1. This introduces inconsistent probability estimates and makes the model unstable. The proposed TopicTRLM outperforms other approaches in all metrics. This confirms the effectiveness of TopicTRLM in the question suggestion task. The advantage of TopicTRLM compared with other approaches is that it fuses the latent semantic meanings of questions with lexical similarities, and this fusion promises to benefit from both the bag-of-words representation and topic model representation.

4.1.4 Experiment on topics' joint probability distribution

In order to answer RQ3, we conducted another experiment on TST_UNLABEL to evaluate topic level performances of the proposed method. For each queried question q, we consider its first subsequent question q' posted by the thread starter in the actual thread as its relevant result. For all the 10,000 queried questions and their relevant results, we used the trained LDA model to infer the most probable topic. We aggregated the counts of topic transitions

Table 10 Comparison on differences between ground truth and methods' topics' joint probability distribution (a smaller KL-divergence value means a better performance)

Methods	Kullback-Leibler Divergence
LDA	0.1127
QL	0.1067
TR	0.0955
TRLM	0.0911
TopicTRLM	**0.0906**

Bold values are the best comparing with other methods

in the actual threads as ground truth and applied maximum-likelihood estimation approach to calculate topics' joint probability using Eq. (6):

$$P(\text{topic}(q), \text{topic}(q')) = P(\text{topic}(q')|\text{topic}(q)) \times P(\text{topic}(q)) \tag{8}$$

We used a 200 * 200 ($K = 200$) matrix to represent ground truth topics' joint probability distributions. In addition, for each queried question, we employed different approaches to retrieve results and considered the first result as its suggested question. We measured the difference between two probability distributions using the Kullback-Leibler divergence. The experimental results are shown in Table 10. Results in Table 10 confirm the effectiveness of the proposed TopicTRLM.

4.2 Experiments in community-based Q&A services

In community-based Q&A services, we also consider the question suggestion task as a retrieval task in our experiments. We aim to address three research questions on question suggestion in community-based Q&A services:

RQ1: How is TopicTRLM-A compared with other approaches on categories "computers & internet" and "travel" on different metrics?
RQ2: What are the suggested questions look like for each method?
RQ3: How is the parameter sensitivity in TopicTRLM-A?

4.2.1 Experimental setup

Methods: To evaluate the performance of the proposed methods, we compared the proposed algorithms with alternative approaches. Specifically, we compared our methods TopicTRLM and TopicTRLM-A with LDA [8], query likelihood language model using Dirichlet smoothing (QL) [66] and translation-based language model (TRLM) [64].

Dataset: We used Yahoo! Answers dataset from Yahoo! Webscope program (Yahoo!). The dataset includes 4,483,032 questions and their answers. More specifically, we utilized the resolved questions under two of the top-level categories at Yahoo! Answers, namely "travel" and "computers & internet." We randomly sampled 100 questions from each category as test questions. The remaining questions and their corresponding best answers in each category were used as question repository, as well as training set for learning word translation probabilities and building LDA model. We used the unigram language model to represent questions and applied IBM model 1 to learn unigram to unigram translation probabilities. We used Porter Stemmer [44] to stem question words. We adopted the stop word list used by SMART system [10], but 5W1H words were removed from the stop word list. For each model, the top 20 retrieval results were kept. We used *pooling* [37] to put results from different models

Table 11 Performance of different models on category "computers & internet" (a larger metric value means a better performance)

Methods	MAP	Bpref	MRR	P@R
LDA	0.2397	0.136	0.2767	0.1594
QL	0.346	0.2261	0.416	0.2594
TRLM	0.3532	0.2368	0.4271	0.2777
TopicTRLM	0.4235	0.2755	0.5559	0.3197
TopicTRLM-A	**0.6228**	**0.4673**	**0.7745**	**0.5467**

Bold values are the best comparing with other methods

for one query together for annotation, and all models were used in the pooling process. If a returned result was considered as semantically related to the queried question, it was labeled with "relevant"; otherwise, it was labeled with "irrelevant." Two assessors were involved in the initial labeling process. If two assessors had different opinions on a decision, a third assessor was asked to make a final decision. The kappa statistics between two assessors was 0.78. We used GIZA++ [40] to train the IBM model 1. We used GibbsLDA++ [42] to conduct LDA training and inference.

Metrics: For the evaluation of the task, we adopted several well-known metrics that evaluate different aspects of the performance of the proposed method, including Precision at Rank R ($P@R$), mean average precision (MAP), mean reciprocal rank (MRR) and Bpref. Bpref is proposed by Buckley et al. [11] and is the score function of the number of non-relevant candidates.

Parameter Tuning: There are several parameters need to be determined in our experiments. We used 20 queries from each category, and employed MAP to tune the parameters. Specifically, parameter values for $\alpha, \beta, K, \lambda, \delta, \gamma$ were set in the same manner as discussed in section 4.1.1. ε is the parameter to balance the contributions between the lexical score and the latent TopicTRLM-A. To make a fair comparison between TopicTRLM and TopicTRLM-A, we set $\varepsilon = \gamma = 0.7$, since γ controls the weights between the lexical score and the latent topic score in TopicTRLM. Due to the same reason, we set $\eta = \delta = 0.2$, since both parameters control the weight of language model in TopicTRLM-A and TopicTRLM, respectively. For parameters θ and μ, we ran an alternative grid search with the constraint that $\theta + \mu = 0.8$. The step size was set as 0.1. The maximal MAP for 20 queries was achieved when $\theta = 0.6$ and $\mu = 0.2$. Thus, optimal parameter values are as follows: $\alpha = 0.25$, $\beta = 0.1$, $K = 200$, $\lambda = 2{,}000$, $\delta = 0.2$, $\gamma = 0.7$, $\varepsilon = 0.7$, $\eta = 0.2$, $\theta = 0.6$ and $\mu = 0.2$.

4.2.2 Experiment on Yahoo! answers dataset

Table 11 demonstrates the results of different models on category "computers and internet" and Table 12 shows the results on category "travel." All the results are statistically significant according to the sign test compared with the previous method.

From Tables 11 and 12, we can find that TopicTRLM-A achieves the best performance on different metrics on two categories. The reason is TopicTRLM-A combines contributions from both questions and their answers through utilizing lexical and latent semantic relatedness, thus getting the best performance. TopicTRLM performs better than methods which utilizes lexical only or latent topic information only by fusing both the lexical and latent semantic knowledge. TRLM performs better than QL, which is consistent with previous research [64]. LDA performs the worst since it is too coarse grained.

We also look into concrete results for test questions. Tables 13 and 14 show results for two test questions from the category "computers and internet." Table 15 shows results

Table 12 Performance of different models on category "travel" (a larger metric value means a better performance)

Methods	MAP	Bpref	MRR	P@R
LDA	0.1345	0.0612	0.1616	0.0675
QL	0.316	0.1902	0.388	0.2048
TRLM	0.3222	0.2034	0.3923	0.2234
TopicTRLM	0.3615	0.244	0.4406	0.2644
TopicTRLM-A	**0.467**	**0.3167**	**0.5963**	**0.387**

Bold values are the best comparing with other methods

Table 13 The results for "Hi, I lost my Yahoo password?

Methods	Results
LDA	1. How can I send my MSN password to my other account if my MSN password is lost? 2. I lost my administrator password and I only have a guest as a user how can I get my password or another one? 3. My other Yahoo email password is stolen by someone, how can I report it and get it back as soon as possible?
QL	1. I keep having a problem with my password. It keeps changing my password or not letting me sign on? 2. I need a program that can help me figure out a password without actually changing the password, or altering it? 3. My minor daughter's Yahoo name and password were changed by someone along with them changing email. What do I?
TRLM	1. I need a program that can help me figure out a password without actually changing the password, or altering it? 2. My minor daughter's Yahoo name and password were changed by someone along with them changing email. What do I? 3. I'm on Myspace and I changed my password but I forgot my password and my email password. What do I do?
TopicTRLM	1. I lost my administrator password and I only have a guest as a user how can I get my password or another one? 2. I forgot my security question, how will I get my lost Yahoo password back for my other id? 3. I have lost my Yahoo password. I don't remember any of the information I fed into the sign u form?
TopicTRLM-A	1. I have lost my Yahoo password. I don't remember any of the information I fed into the sign u form? 2. I forgot my security question, how will I get my lost Yahoo password back for my other id? 3. If I forget my Yahoo password, is there any way to get it back or does Yahoo have to send me a new one?

How can I get my old password back without any changing with my email?" of category "computers and internet"

for one test question from the category "travel." Let's take the question in Table 15 as an example, the queried question is "Why can people only use the air phones when flying on commercial airlines, i.e., no cell phones etc.?" Thus, the underlying information need of the user is to know why cell phone could not be used in commercial airlines. The first result of TopicTRLM-A model is "Why are you supposed to keep cell phone off during flight in commercial airlines?" We can find the first result is semantically equivalent to the test question, thus, the best answer of the first result should answer the user's information need accurately. The second result of TopicTRLM-A model is "Why don't cell phones from the ground at or near airports cause interference in the communications of aircraft?" This question

Table 14 The results for “I want to just know how to use Outlook Express to send mail, and if possible to sync with Yahoo or MSN?” of category “computers and internet”

Methods	Results
LDA	1. How can I do to synchronize my Yahoo or MSN email account with Microsoft Outlook to send and receive messages? 2. How can you set up to read your sent messages on MSN Hotmail, like Yahoo? 3. What’s my Yahoo incoming mail (POP3/IMAP)?!! At Yahoo mail & outgoing mail (SMTP) required 4 email notifier?
QL	1. Can I sync Outlook Express inbox with Yahoo inbox? 2. Is there a way to sync Yahoo calendar to Blackberry 7,100t without having Outlook? 3. How do I sync contacts from Ipaq with Outlook? Most of them are not getting synched?
TRLM	1. Can I sync Outlook Express inbox with Yahoo inbox? 2. I have several contacts in my Outlook that are not showing up on my Treo when I sync. How do I fix??? 3. Is there a way to sync Yahoo calendar to Blackberry 7100t without having Outlook?
TopicTRLM	1. What’s my Yahoo incoming mail (POP3/IMAP)?!! At Yahoo mail & outgoing mail (SMTP) required 4 email notifier? 2. How do I synchronize my Yahoo mail with Outlook Express i.e. I wish to use Outlook Express to check my Y! mail 3. How can I do to synchronize my Yahoo or MSN email account with Microsoft Outlook to send and receive messages?
TopicTRLM-A	1. I want to use Yahoo mail in my Outlook Express. Please tell me POP3 and SMTP address of Yahoo? 2. Is it possible to configure my Yahoo id in Outlook Express 6? 3. Sir, please tell me I am all Yahoo mail converlet in Outlook please tell me how I do?

Table 15 The results for “Why can people only use the air phones when flying on commercial airlines, i.e. no cell phones etc.?” of category “travel”

Methods	Results
LDA	1. Why are you supposed to keep cell phone off during flight in commercial airlines? 2. I will be flying from CA to FL. Any tips on how I can get over my fear of flying? 3. I need the contact number of emirates airlines here in Philippines?
QL	1. Cell phones and pagers really dangerous to avionics? 2. Do cell phones work on cruise ships? T-mobile? 3. Cell phones in Singapore, Bali or Kuala Lumpur?
TRLM	1. Why don’t cell phones from the ground at or near airports cause interference in the communications of aircraft? 2. Should I bring my Vertu phones in my carry-on luggage or send through? I have 3 of them? 3. Cell phones and pagers really dangerous to avionics?
TopicTRLM	1. Cell phones and pagers really dangerous to avionics? 2. Why don’t cell phones from the ground at or near airports cause interference in the communications of aircraft? 3. Do cell phones work on cruise ships? T-mobile?
TopicTRLM-A	1. Why are you supposed to keep cell phone off during flight in commercial airlines? 2. Why don’t cell phones from the ground at or near airports cause interference in the communications of aircraft? 3. Cell phones and pagers really dangerous to avionics?

Fig. 7 The effect of parameter ε on the MAP of question suggestion

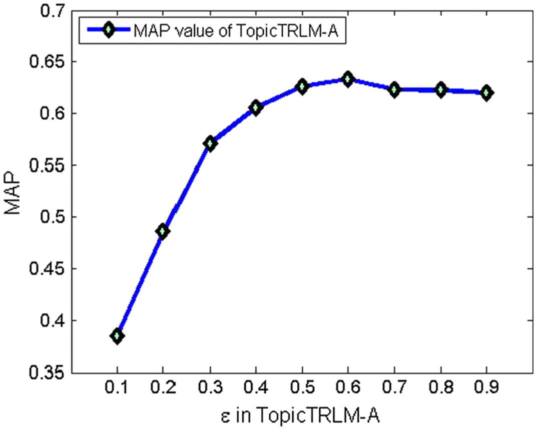
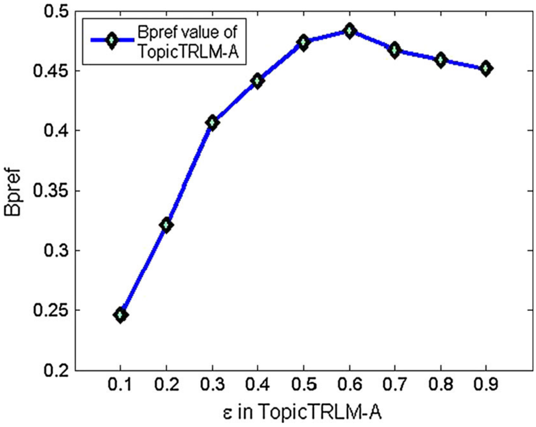


Fig. 8 The effect of parameter ε on the Bpref of question suggestion



is quite related to the test question since it also discusses the interference of cell phones to the communications of aircraft, and it also belongs to the topic of “interference of aircraft.” The third result is “Cell phones and pagers really dangerous to avionics?” This question would open the asker’s mind that not only cell phones, but also pagers maybe dangerous to aircraft systems, more specifically, to avionics. We can find that TopicTRLM-A could not only find questions that are semantically equivalent to the queried question, but also find questions that are semantically related to the queries question. Thus, TopicTRLM-A could satisfy users’ information needs more thoroughly. Tables 13 and 14 show similar findings.

We also test the sensitivity of parameter ε in TopicTRLM-A. A larger ε means we reply more on lexical score, and a smaller ε means we reply more on latent semantic knowledge. Figure 7 shows the MAP of employing different ε on category “computers and internet,” Fig. 8 shows the Bpref, Fig. 9 shows the MRR and Fig. 10 shows the P@R. TopicTRLM-A performs the best when ε is between 0.5 and 0.7 on different metrics. The optimal parameter range indicates that only by fusing both lexical and latent semantic knowledge together, the model could achieve the best performance. Figures 7, 8, 9 and 10 also demonstrate that TopicTRLM-A is sensitive to the parameter ε, but the parameter ε is still feasible to tune since the optimal range is relatively wide. The optimal parameter range of ε is similar on the other category “travel.”

Fig. 9 The effect of parameter ε on the MRR of question suggestion

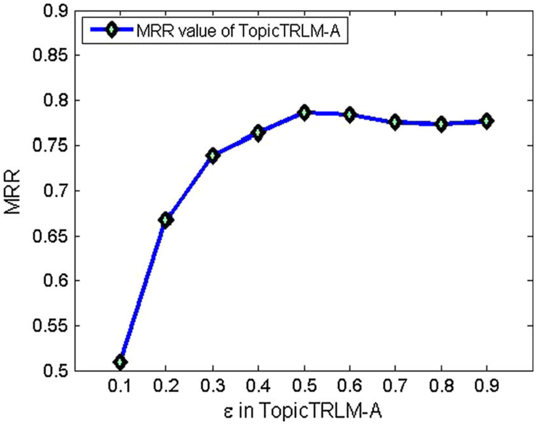
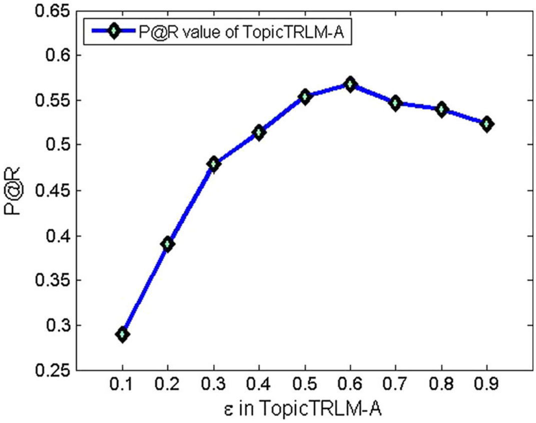


Fig. 10 The effect of parameter ε on the P@R of question suggestion



5 Conclusion and future work

In this paper, we address the issue of question suggestion in social media. Given a queried question, we are to suggest questions that are semantically related to the queried question and that can explore different aspects of a topic tailored to users' information needs. We tackle the problem on two types of the most representative social media systems with Q&A functionality: online forums and community-based Q&A services. In online forums, we propose an effective method to build the parallel corpus of related questions from forum threads, and we propose TopicTRLM, which fuses lexical knowledge with latent semantic knowledge to measure the relatedness between questions. In community-based Q&A services, we also propose an effective method to build the parallel corpus of related questions, and we propose TopicTRLM-A, which incorporates answer information into question to measure the semantic relatedness more thoroughly. Extensive experiments indicate that our method to build parallel corpus is effective and the TopicTRLM and TopicTRLM-A methods outperform other approaches.

Because we want to assist users in exploring different aspects of the topic that he/she is interested in by offering question suggestion service, it is worthwhile to investigate how to measure and how to diversify the suggested questions. Moreover, as users' voting and rating behaviors in community-based Q&A and online forums are important to infer both question and answer quality, we would investigate how to incorporate these aspects to suggest high-quality semantically related questions in the future.

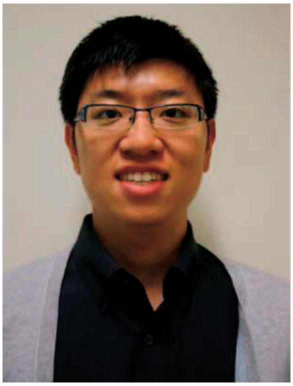
Acknowledgments The work described in this paper was fully supported by the Basic Research Program of Shenzhen (Project No. JCYJ20120619152419087 and JC201104220300A), and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413212 and CUHK 415212). The authors would like to thank the anonymous reviewers for their insightful comments and helpful suggestions.

References

1. Adamic LA, Zhang J et al (2008) Knowledge sharing and yahoo answers: everyone knows something. In: Proceedings of the 17th international conference on World Wide Web. ACM
2. Agichtein E, Lawrence S et al (2001) Learning search engine specific query transformations for question answering. In: Proceedings of the 10th international conference on World Wide Web. ACM
3. Agichtein E, Liu Y et al (2009) Modeling information-seeker satisfaction in community question answering. ACM Trans Knowl Discov Data (TKDD) 3(2):10
4. Berger A, Caruana R et al (2000) Bridging the lexical chasm: statistical approaches to answer-finding. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM
5. Berger A, Lafferty J (1999) Information retrieval as statistical translation. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. ACM
6. Bernhard D, Gurevych I (2009) Combining lexical semantic resources with question and answer archives for translation-based answer finding. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: volume 2-volume 2. Association for Computational Linguistics
7. Bian J, Liu Y et al (2008) Finding the right facts in the crowd: factoid question answering over social media. In: Proceedings of the 17th international conference on World Wide Web. ACM
8. Blei DM, Ng AY et al (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
9. Brown PF, Cocke J et al (1990) A statistical approach to machine translation. Comput Linguist 16(2): 79–85
10. Buckley C, Singhal A et al (1995) New retrieval approaches using SMART: TREC 4. In: Proceedings of the 4th text REtrieval conference (TREC-4)
11. Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. ACM
12. Burke RD, Hammond KJ et al (1997) Question answering from frequently asked question files: experiences with the faq finder system. AI Mag 18(2):57
13. Cao X, Cong G et al (2010) A generalized framework of exploring category information for question retrieval in community question answer archives. In: Proceedings of the 19th international conference on World Wide Web. ACM
14. Cao X, Cong G (2012) Approaches to exploring category information for question retrieval in community question-answer archives. ACM Trans Inf Syst (TOIS) 30(2):7
15. Cao Y, Duan H et al (2011) Re-ranking question search results by clustering questions. J Am Soci Inf Sci Technol 62(6):1177–1187
16. Cao Y, Duan H et al (2008) Recommending questions using the mdl-based tree cut model. In: Proceedings of the 17th international conference on World Wide Web. ACM
17. Cong G, Wang L et al (2008) Finding question-answer pairs from online forums. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. ACM
18. Deerwester SC, Dumais ST et al (1990) Indexing by latent semantic analysis. JASIS 41(6):391–407
19. Demner-Fushman D, Lin J (2007) Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist 33(1):63–103
20. Duan H, Cao Y et al (2008) Searching questions by identifying question topic and question focus. In: Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies
21. Ferrucci D, Brown E et al (2010) Building Watson: an overview of the deepQA project. AI Mag 31(3): 59–79

22. Gazan R (2011) Social Q&A. J Am Soc Inf Sci Technol 62(12):2301–2312
23. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Nat Acad Sci USA 101(Suppl 1):5228–5235
24. Harabagiu S, Moldovan D et al (2001) Answering complex, list and context questions with LCC's question-answering server. In: Proceedings of the text retrieval conference for question answering (TREC 10)
25. Heinrich G (2005) Parameter estimation for text analysis. Fraunhofer IGD
26. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. ACM
27. Huston S, Croft WB (2010) Evaluating verbose query processing techniques. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM
28. Jeon J, Croft WB et al (2005) Finding semantically similar questions based on their answers. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. ACM
29. Jeon J, Croft WB et al (2005) Finding similar questions in large question and answer archives. In: Proceedings of the 14th ACM international conference on information and knowledge management. ACM
30. Jijkoun V, de Rijke M (2005) Retrieving answers from frequently asked questions pages on the web. In: Proceedings of the 14th ACM international conference on information and knowledge management. ACM
31. Kim S, Oh S (2009) Users' relevance criteria for evaluating answers in a social Q&A site. J Am Soc Inf Sci Technol 60(4):716–727
32. Li B, Liu Y et al (2008) CoCQA: co-training over questions and answers with an application to predicting question subjectivity orientation. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics
33. Lin J, Katz B (2006) Building a reusable test collection for question answering. J Am Soc Inf Sci Technol 57(7):851–861
34. Lou J, Fang YL et al (2013) Contributing high quantity and quality knowledge to online Q&A communities. J Am Soc Inf Sci Technol 64(2):356–371
35. Lou J, Lim KH et al (2011) Drivers of knowledge contribution quality and quantity in online question and answering communities. In: Proceedings of the 15th pacific conference on information systems
36. Lou J, Lim KH et al (2012) Knowledge contribution in online question and answering communities: effects of groups membership. In: Proceedings of the 2012 international conference on information systems
37. Manning CD, Raghavan P et al (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
38. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41
39. Mitra M, Singhal A et al (1998) Improving automatic query expansion. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM
40. Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51
41. Ofoghi B, Yearwood J et al (2009) The impact of frame semantic annotation levels, frame-alignment techniques, and fusion methods on factoid answer processing. J Am Soc Inf Sci Technol 60(2):247–263
42. Phan XH, Nguyen LM et al (2008) Learning to classify short and sparse text and web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on World Wide Web. ACM
43. Pomerantz J (2005) A linguistic analysis of question taxonomies. J Am Soc Inf Sci Technol 56(7):715–728
44. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
45. Qu B, Cong G et al (2012) An evaluation of classification models for question topic categorization. J Am Soc Inf Sci Technol 63(5):889–903
46. Raban DR (2009) Self-presentation and the value of information in Q&A websites. J Am Soc Inf Sci Technol 60(12):2465–2473
47. Radev D, Fan W et al (2005) Probabilistic question answering on the web. J Am Soc Inf Sci Technol 56(6):571–583
48. Radev DR, Libner K et al (2002) Getting answers to natural language questions on the web. J Am Soc Inf Sci Technol 53(5):359–364
49. Ramage D, Heymann P et al (2009) Clustering the tagged web. In: Proceedings of the second ACM international conference on web search and data mining. ACM
50. Ramos J (2003) Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning
51. Riezler S, Vasserman A et al (2007) Statistical machine translation for query expansion in answer retrieval. In: Proceedings of the 45th annual meeting of the Association for Computational Linguistics

52. Rosen-Zvi M, Chemudugunta C et al (2010) Learning author-topic models from text corpora. ACM Trans Inf Syst (TOIS) 28(1):4
53. Rosenbaum H, Shachaf P (2010) A structuration approach to online communities of practice: the case of Q&A communities. J Am Soc Inf Sci Technol 61(9):1933–1944
54. Shah C, Kitzie V (2012) Social Q&A and virtual reference–comparing apples and oranges with the help of experts and users. J Am Soc Inf Sci Technol 63(10):2020–2036
55. Liu GZ (1998) Automated information retrieval: theory and methods. J Am Soc Inf Sci 49(10):953–955
56. Shrestha L, McKeown K (2004) Detection of question-answer pairs in email conversations. In: Proceedings of the 20th international conference on computational linguistics. Association for Computational Linguistics
57. Shtok A, Dror G et al (2012) Learning from the past: answering new questions with past answers. In: Proceedings of the 21st international conference on World Wide Web. ACM
58. Soricut R, Brill E (2004) Automatic question answering: Beyond the factoid. In: Proceedings of the HLT-NAACL
59. Sparck Jones K (1971) Automatic keyword classification for information retrieval. Butterworths, London
60. Voorhees E, Tice DM (1999) The TREC-8 question answering track evaluation. In: Proceedings of the eighth text retrieval conference (TREC-8). http://trec.nist.gov/pubs/trec8/t8_proceedings.html
61. Wang K, Ming Z et al (2009) A syntactic tree matching approach to finding similar questions in community-based qa services. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. ACM
62. Wei X, Croft WB (2006) LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval
63. Wu CH, Yeh JF et al (2005) Domain-specific FAQ retrieval using independent aspects. ACM Trans Asian Lang Inf Process (TALIP) 4(1):1–17
64. Xue X, Jeon J et al (2008) Retrieval models for question and answer archives. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. ACM
65. Yahoo! Yahoo! Webscope dataset, ydata-yanswers-all-questions-v1_0. http://research.yahoo.com/Academic_Relations
66. Zhai C, Lafferty J (2004) A study of smoothing methods for language models applied to information retrieval. ACM Trans Inf Syst (TOIS) 22(2):179–214
67. Zhou TC, Lin CY et al (2011) Learning to suggest questions in online forums. In: Proceedings of the 25th AAAI conference on artificial intelligence
68. Zhou TC, Lyu MR et al (2012) A classification-based approach to question routing in community question answering. In: Proceedings of the 21st international conference companion on World Wide Web. ACM
69. Zhou TC, Ma H et al (2009) Tagrec: leveraging tagging wisdom for recommendation. Computational Science and Engineering, 2009. CSE'09. International Conference on IEEE
70. Zhou TC, Ma H et al (2010) UserRec: a user recommendation framework in social tagging systems. AAAI
71. Zhou TC, Si X et al (2012) A data-driven approach to question subjectivity identification in community question answering. In: Proceedings of the twenty-sixth AAAI conference on artificial intelligence



Tom Chao Zhou received the B.S. (2008) in Computer Science and Technology from Zhejiang University, China, and the Ph.D. (2013) in Computer Science and Engineering from the Chinese University of Hong Kong. He is a Senior Research Engineer in Baidu. His research interests include information retrieval, natural language processing and data mining, where he has published 11 technical publications in journals (IEEE TSC, ACM TOIS) and conferences (AAAI, SIGIR, ISSRE, etc.) with hundreds of citations. He has been granted two patents. He has visited Google Research, Microsoft Research Redmond, Microsoft Research Asia, and National University of Singapore during his PhD study. He served as program committee member or reviewer for ACL, SDM, IJCAI, WSDM, WWW, KDD, CIKM, EMNLP, MM, TOIS, IJCNN, etc. He received the award of Best New Employee of NLP Department of Baidu, AAAI Student Scholarship, PCCW Foundation Scholarship, and Global Scholarship Program for Research Excellence of CUHK.

Michael Rung-Tsong Lyu received the B.S. (1981) in electrical engineering from National Taiwan University, the M.S. (1985) in computer engineering from University of California, Santa Barbara, and the Ph.D. (1988) in computer science from University of California, Los Angeles. He is a Professor in the Computer Science and Engineering Department of the Chinese University of Hong Kong. His research interests include software reliability engineering, software fault tolerance, distributed systems, data mining, social networks, machine learning, multimedia information retrieval, and mobile networks, where he has published over 400 papers. He has been an Associate Editor of IEEE Transactions on Reliability, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Services Computing, and Wiley Software Testing, Verification and Reliability Journal. Dr. Lyu is an IEEE Fellow, an AAAS Fellow, a Croucher Senior Fellow, and received IEEE Reliability Society 2010 Engineer of the Year Award.

Irwin King is Associate Dean (Education), Faculty of Engineering and Professor at the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He received his B.Sc. degree in Engineering and Applied Science from California Institute of Technology (Caltech), Pasadena and his M.Sc. and Ph.D. degree in Computer Science from the University of Southern California (USC), Los Angeles. His research interests include machine learning, social computing, web intelligence, data mining and multimedia information processing. In these research areas, he has over 200 technical publications in journals and conferences. In addition, he has contributed over 30 book chapters and edited volumes. Moreover, Prof. King has over 30 research and applied grants. He is an Associate Editor of the ACM Transactions on Knowledge Discovery from Data (ACM TKDD) and Journal of Neural Networks. He is a member of the Board of Governors and Vice-President for both INNS and APNNA. He is the General Chair of WSDM2011, General Co-Chair of RecSys2013, and in various capacities in a number of top conferences such as WWW, NIPS, ICML, IJCAI, AAAI, etc.

Jie Lou is now a Ph.D. student at City University of Hong Kong. She is interested in the research areas of knowledge management, social media, mobile technology usage, etc. Her research articles have appeared in several international journals and conferences, such as Journal of American Society for Information Science and Technology (JASIST), International Conference on Information Systems (ICIS) and Pacific Asia Conference on Information Systems (PACIS).