The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2398-7294.htm

Received 2 March 2021
Revised 19 April 2021
Accepted 22 April 2021

A study of similar question retrieval method in online health communities

Bufei Xing
Baidu Online Network Technology (Beijing) Co., Ltd, Beijing, China

Haonan Yin
Paul Merage School of Business, University of California, Irvine, California, USA, and

Zhijun Yan and Jiachen Wang
School of Management and Economics, Beijing Institute of Technology, Beijing, China

Abstract

Purpose – The purpose of this paper is to propose a new approach to retrieve similar questions in online health communities to improve the efficiency of health information retrieval and sharing.

Design/methodology/approach – This paper proposes a hybrid approach to combining domain knowledge similarity and topic similarity to retrieve similar questions in online health communities. The domain knowledge similarity can evaluate the domain distance between different questions. And the topic similarity measures questions' relationship base on the extracted latent topics.

Findings – The experiment results show that the proposed method outperforms the baseline methods.

Originality/value – This method conquers the problem of word mismatch and considers the named entities included in questions, which most of existing studies did not.

Keywords Online health communities, Question retrieval, Named entity extraction, Topic model

Paper type Research paper

1. Introduction

The fast popularity of Web 2.0 results in a great change in health area and leads to the emergence of Medicine 2.0 (Eysenbach, 2008; Van De Belt *et al.*, 2010), which provides an interactive and effective communication platform for doctors and patients. More and more patients like to find health information and share their experience in online health communities (OHC), which is widely adopted in Medicine 2.0 (Figueroa, 2017; Roberts

© Bufei Xing, Haonan Yin, Zhijun Yan and Jiachen Wang. Published in *International Journal of Crowd Science*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Funding: This work was supported by National Key Research and Development Plan of China (Grant No: 2017YFB1400101), National Natural Science Foundation of China (Grant No: 71572013, 71872013, 72072011) and Beijing Municipal Social Science Foundation (Grant No: 18JDGLB040).

International Journal of Crowd Science
Vol. 5 No. 2, 2021
pp. 154-165
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-03-2021-0006

and Demner-Fushman, 2016). Many online health communities, such as HealthTap (www.healthtap.com) and GoodDoctor (www.haodf.com), provide the communication channel between doctors and patients. Patients can ask health questions ranging from precaution measures to disease treatment in these communities. Doctors normally answer these questions freely or with a charge. Obviously, OHC become an important and valuable platform for patients to gather information, find support and improve health status (Yan *et al.*, 2016).

With deep involvement of community members, the number of posted questions and answers increases rapidly, which brings the information overload problem. Before patients want to ask health questions, they will normally search for the similar questions and related answers at first. Quite a few questions have already been asked and answered in the past (Figueroa, 2017). It is time consuming for patients to browse each question and answer to locate information they interest. They can hardly find similar ones from the massive number of questions and answers. Hence, finding the relevant similar question-answer pairs in large archives is very important, which can help users to find suitable answers more efficiently (Liu *et al.*, 2014). If the existing answers are satisfactory, patients will not need to post questions in communities. The existing studies mainly used three different kinds of methods to find similar questions, i.e. lexicon-based, syntax-based and topic-based methods. The lexicon-based method retrieves similar questions based on word matching (Samuel *et al.*, 2017; Wu *et al.*, 2008). The syntax-based method uses syntactic tree structure of questions to measure the questions distance and identify the similar questions (Ferrández, 2011). The topic-based method extracts the topic or the focus structure from questions and suggests that questions are similar if they have same topics or focus structure (Chen *et al.*, 2016a; Wu, 2015).

Although prior studies extensively explore the similar question retrieval in online communities, their existing approaches have two limitations in health area. First, different from other communities, there exists significant informal term expression phenomena in OHC. Patients do not have professional health knowledge and frequently use different kinds of informal or even wrong expressions for health terms. The informal expression of terms is ignored in previous studies. Second, the online health question-answer pairs usually include many professional health terms, which is helpful for information seeking and extraction. However, the existing methods did not take this factor which may improve the question retrieval performance into consideration.

To address above problems, we propose a novel domain-knowledge based approach to retrieve similar health questions. We reformulate the similar question retrieval problem as the question similarity calculation problem. The question similarity consists of two parts: domain knowledge similarity and topic similarity. The domain knowledge similarity is defined to evaluate the domain relationship of two questions. And the topic similarity is proposed to measure the latent connection between two questions. This study first applies the Conditional Random Field (CRF) model to recognize health named entities from the question and uses the word move distance measure the domain knowledge similarity. By adopting the Latent Dirichlet Allocation (LDA) method, the latent topics are extracted from questions and topic similarity between questions is computed. By integrating the domain knowledge similarity and topic similarity, the similar questions and answers can be recommended to patients when patients post a new question.

This study contributes to literature in several aspects. First, we applied the similar question retrieval method to a professional service industry, i.e. health service. OHC is becoming one of the most popular and active communities. Many patients get health knowledge and handle their health problems by using question-answer service in OHC.

Prior studies mostly focus on similar question retrieval in general communities. Second, patients may use informal and ambiguous expressions or terms in OHC, which makes it more difficult to identify the similar questions. We propose a novel approach based on distributional semantic vectors to compute domain knowledge similarity of health questions. And the word mismatching problem resulted from informal and ambiguous expression can be well solved. Third, the proposed method can extract the health named entities in OHC. The health entities include the most useful information in the question-answer pair. The recognition of health entities can help to identify the similar health questions in OHC.

The remainder of the paper is organized as follows. Section 2 introduces the related work in finding similar questions. The proposed similar health question retrieval method is presented in Section 3. Section 4 gives the experiment results. Finally, we discuss major research findings and practical implications of this study in Section 5.

2. Related work

Similar question retrieval focuses on finding similar questions in communities (Wang *et al.*, 2009). There are mainly three kinds of approaches in retrieving similar questions from community-based question-answer archives, i.e. lexicon-based, syntax-based and topic-based approaches.

2.1 Lexicon-based methods

Lexicon-based method mostly adopts the word match technique to retrieve similar questions. The more overlap words two documents have, the more similar they are. The lexicon-based method mostly uses the bag-of-words (BoW) model to compute the words' similarity. The similarity calculation method used in the BoW model includes inverse document frequency overlap method, phrase overlap method (Banerjee and Pedersen, 2003), term frequency and inverse document frequency term (TF-IDF) weights (Wu *et al.*, 2008) and Jaccard similarity coefficient method (Niwattanakul *et al.*, 2013). For example, based on the number of shared words between the two documents, Banerjee and Pedersen presented a model to measure the semantic similarity between documents (Banerjee and Pedersen, 2003). Zhang *et al.* explored a key concept identification approach for query refinement and a pivot language translation based approach to explore key concept paraphrasing. Based on these two approaches, they proposed a new similar question retrieval method with high performance (Zhang *et al.*, 2016). Although the lexicon-based methods can be applied in similar questions identification, it just considers the repetition of words in two questions. The word order, word sense and syntactic information are ignored, which results in the unsatisfactory performance (Zhang *et al.*, 2014).

2.2 Syntax-based methods

The syntax-based method mainly uses syntactic tree structure of questions to measure questions' distance. This type of method not only calculates the words' similarity of different questions but also takes the similarity of syntactic structure into account. Wang employed a syntactic tree structure to measure words' distance, and then used WordNet and Leacock's index to evaluate the semantic similarity of questions (Wang *et al.*, 2009). Based on the lexical and syntactic knowledge generated by a Part-of-Speech (PoS) tagger and a syntactic chunker, Ferrández integrated lexical dependency information between terms into traditional information retrieval similarity measurement (Ferrández, 2011). Lian *et al.* divided the question retrieval problem into question classification and question retrieval. Question classification prunes the search space and removes some noise, and then

"dependency syntactic tree" is used to find similar questions within the predetermined categories (Lian, Yuan, Hu, and Zhang, 2013). Despite taking the syntactic tree into consideration, this type of method does not identify the semantic and topic level similarity between two questions (Zhang *et al.*, 2014).

2.3 Topic-based methods

Unlike the above two methods, the topic-based method maps questions to a topic and focus structure (Chen *et al.*, 2016b; Gao *et al.*, 2014). Duan *et al.* represented questions by a topic and focus structure and used the MDL-based (Minimum Description Length) tree cut model to identify question topic and focus automatically. Then the topic and focus similarity can be calculated to find similar questions (Duan *et al.*, 2008). Wu proposed a relevance-dependent topic model by integrating the past queries to enhance accuracy. They first constructed a collection of documents to reveal the users' intentions based on their past behavior, and then applied the topic model to extract latent topics. The retrieval results are ranked by highlighting topics that are highly similar to the query according to their relevance in the updated document model (Wu, 2015). Jiang *et al.* proposed a Topic Enhancing Inverted Index (TEII) method that incorporated the topic information into inverted index (i.e. TFIDF) for top-k document retrieval. By combining the topic similarity with the traditional TFIDF similarity, the method got good performance (Jiang *et al.*, 2015). However, this type of method didn't consider the named entities information which is an obvious feature in OHC.

3. Methodology

To alleviate the problems of existing approaches, we propose a novel method to retrieve similar questions in OHC. The proposed method mainly includes two steps. The first is data preprocess, which includes word segmentation and PoS tagging. The second step is to calculate the question similarity, which includes topic-based and domain-knowledge based similarity computing. We first perform word segmentation and PoS tagging on each question. Then the Conditional Random Field (CRF) model is adopted to extract health named entities in questions. In the similarity computing phrase, we first train word2vector model to get the word-embedding of every term. Each question-answer pair is classified into two term sets: health terms (i.e. health named entities) and non-health general terms. With the word-embedding, we build a linear programming model to get the minimum distance of two terms sets which represents the similarity between two sets and is called domain knowledge similarity. Then we employ LDA model to extract latent topic of questions. Each question will be mapped into the topic space. Finally, we combine domain knowledge similarity and topic similarity, and retrieve the top-N similar questions. The framework of our approach is shown in Figure 1.

3.1 Preprocessing

The first step is data preprocessing that removes useless characters, such as special symbols. Then, word segmentation and PoS tagging are performed on individual question-answer pairs. Chinese sentences have different structure from English sentences and word segmentation is the basic operation to handle text information. It can separate each word from its adjacent others in a Chinese sentence. We use a Chinese lexical analysis tool called NLPIR (Natural Language Processing and Information Retrieval Sharing Platform, http://ictclas.nlpir.org/) for sentence segmentation and PoS tagging.

The following example shows a parsed Chinese sentence with PoS tags:

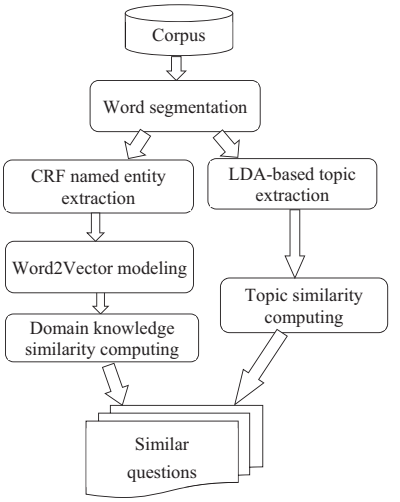

Figure 1.
Overview of the proposed method

Original sentence: "糖尿病十二年,眼底出血,左眼干涩,右眼没有不适症状,是否需要做激光?" (i.e. "With 12 years of diabetes, my fundus is bleeding, and my left eye is dry while right eye does not have any discomfort symptoms. Do I need to do laser?").

Result of PoS tagging by NLPIR: "糖尿病(diabetes)/disease 十二(twelve)/m 年(years)/qt,/wd 眼底(fundus)/n 出血(bleed)/v,/wd 左眼(left eye)/n 干涩(dry)/a,/wd 右眼(right eye)/n 没有(no)/d 不适(discomfort)/a 症状(symptoms)/n,/wd 是否(if)/v 需要(need)/v 做(do)/v 激光(laser)/treatment?/wd".

In which the tag 'n' represents a noun; 'an' represents an adnoun; 'wd' represents a punctuation; 'd' represents an adverb; 'a' represents an adjective; 'v' represents an verb; 'm' represents an number; and 'disease' or 'treatment' represents that the type of the word is disease or treatment.

3.2 Domain knowledge similarity

Domain knowledge similarity measures the domain relationship of different health questions. This study adopts the Word2Vector model to compute the distance of different words in questions based on the extracted health entities. Then, the similarity of different questions can be derived by word mover's distance method.

3.2.1 Named entity extraction. After word segmentation, we then employ CRF model to extract health named entities in posted questions. CRF model is widely used in named entity extraction and has been proved to have a good performance (Lei *et al.*, 2014; Uzuner *et al.*, 2011). In this stage, we use PoS (part-of-speech) feature, body indicator feature, suffix feature as feed features to train CRF model. These features are widely used in previous studies (Yang *et al.*, 2014).

Based on the word segmentation and PoS tagging results, we first label each word with the above features in all questions. For example, for the Chinese word '左眼'(left eye), the label of PoS feature is 'n', and the label of body indicator feature is '1', because '眼'(eye) is a kind of organ in human body. The label of suffix feature is '眼'. After annotating the feature label of each word in each question, we then annotate each word with a named entity label, i.e. B-X, I-X, O, in which, B-X represents the beginning of an entity, I-X represents the

continuation of an entity and the O represents the word is without any relationship with an entity. The character 'X' can be 'D' (disease), 'T' (treatment), 'C' (check), 'S' (symptom). Then the widely used tool, i.e. CRF++ (https://taku910.github.io/crfpp/), is adopted to train our named entity extraction model. With a well-trained CRF model, we extract the health named entities from the whole corpus.

3.2.2 Semantic distance calculation based on Word2Vector model. Word2Vector is a distributional vector space model and can convert words to distributional semantic vectors (i.e. word-embedding). In Word2Vector, each word is represented by a vector and word similarity can be calculated based on vector similarity. The closer they are in the distributional vector space, the more similar two words are.

After training the Word2Vector model with the whole corpus, the distributional semantic vector (word-embedding) of each term can be derived. Supposing that the distributional semantic vector of term e_i and term e_j are v_i and v_j respectively, the semantic distance between e_i and e_j is defined as:

$$d(e_i, e_j) = 1 - \cos(v_i, v_j) \tag{1}$$

Where cos (v_i, v_j) represents the cosine similarity between v_i and v_j.

3.2.3 Domain knowledge similarity calculation. Based on the extracted health named entities, terms in each question can be classified into two sets: health terms and non-health terms. We use E to denote the set of health terms and F to denote the set of non-health terms. For two questions p and q, the domain knowledge similarity is calculated based on the similarity between E_p and E_q and similarity between F_p and F_q. We use the Word Mover's Distance method to derive the domain knowledge similarity.

Word Mover's Distance (WMD) is a novel distance function between text documents based on word-embedding (Kusner *et al.*, 2015). The WMD defines the dissimilarity between two documents as the minimum distance, which the embedded words of one document need to "travel" to reach the embedded words of another document. Assuming that the set S = $\{s_i\}$ (i = 1, 2,. . .,|S|) and the set T= $\{t_j\}$ (j = 1, 2,. . ., |T|), the elements of each set are terms and the weight of each term in each set is 1/|S| and 1/|T|, respectively. The weight here represents the importance of a term in a set, and we assume that the total importance of a set is 1. The semantic distance between the word s_i in set S and the word t_j in set T is d (s_i, t_j), which is calculated as the formula (1). Then, the problem is, how to get the semantic distance between the set S and T with the word distance d (s_i, t_j).

Supposing that the weight of d (s_i, t_j) that contributes to the distance between S and T is w (i,j), we can compute the semantic distance between S and T by summing w (i,j) *d (s_i, t_j), i.e. $\sum_{i}^{|S|}\sum_{j}^{|T|} w(i,j)*d(s_i, t_j)$. Meanwhile, because the weight of each term in S and T is 1/|S| or 1/|T| respectively, there are two restrictions about w(i,j). For each j (j = 1,2,. . .,|T|), $\sum_{i=1}^{|S|} w(i,j)$ should be equal to 1/|T|. For each i (i = 1, 2,. . .,|S|), $\sum_{j=1}^{|T|} w(i,j)$ should be equal to 1/|S|. Moreover, w(i,j) is no less than 0. Then the minimum value of $\sum_{i=1}^{|S|}\sum_{j=1}^{|T|} w(i,j)*d(s_i, t_j)$ is the semantic distance (i.e. WMD) of the set S and T. Thus we have the following linear programming model to get the semantic distance of the set S and T.

$$\min \sum_{i=1}^{|S|}\sum_{j=1}^{|T|} w(i,j)*d(s_i,t_j)$$

subject to :

$$\begin{aligned} &\sum_{i=1}^{|S|} w(i,j) = 1/|T|,\ j = 1,2,\ldots,|T| \\ &\sum_{j=1}^{|T|} w(i,j) = 1/|S|, i = 1,2,\ldots,|S| \\ &w(i,j) \geq 0, \forall i,j \end{aligned} \tag{2}$$

The above optimization problem is a well-studied transportation problem, which can be solved by simplex method or the potential method (with redistributive cycle). By solving the above problem, we can get the minimum distance of set S and set T, which is represented by $d_{min}(S,T)$. Then, we can get the semantic similarity of set S and T as follows.

$$similarity_{set}(S,T) = 1 - d_{min}(S,T) \tag{3}$$

Then, the domain knowledge similarity between question p and q is given as follows:

$$similarity_{domain}(p,q) = \varepsilon * similarity_{set}(E_p,E_q) + (1-\varepsilon) * similarity_{set}(F_p,F_q) \tag{4}$$

Where $similarity_{set}(E_p, E_q)$ is the semantic distance between the health terms set of question p and q; $similarity_{set}(F_p, F_q)$ is the semantic distance between the non-health terms set of question p and q; ε is an adjustable parameter between $similarity_{set}(E_p, E_q)$ and $similarity_{set}(F_p, F_q)$, which ranges from 0.0 to 1.0.

3.3 Topic similarity

3.3.1 Lda model. Latent Dirichlet Allocation (LDA) was first introduced by Blei *et al.* (2003). As an unsupervised generative probability model, LDA model assumes that a document has some different topics, and different models have different topic distributions. Each topic is a distribution over words. LDA model is used to mine the latent topics included in documents.

LDA assumes the following generative process for each document d in a corpus D:

(1) Choose N∼Possion (ξ), where N represents the number of words in a document;

(2) Choose $\theta \sim$Dir (α), where θ is the topic distribution

(3) For each of the N words w_n:

- Choose a topic $z_n \sim$ Multinomial (θ).
- Choose a word w_n from p (w_n $|z_n, \beta$), and p(w_n $|z_n, \beta$) is a multinomial probability conditioned on the topic z_n.

So, given the parameter α and β, the joint distribution of a topic mixture θ, a set of N topics, and a set of N words W is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta). \quad (5)$$

And the probability of a corpus is:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)d\theta_d \quad (6)$$

In this paper, we use Gibbs sampling to generate the topic distribution of each document and estimate the LDA model.

3.3.2 Latent topic similarity. In this step, LDA model is used to derive the topic distribution of each question. Then cosine similarity is adopted to compute the topic similarity between questions. For example, if we got the topic distribution vector of question p as f_p and the topic distribution of question q as f_q, then the topic similarity between p and q is given as follows:

$$Similarity_{topic}(\mathrm{q}, \mathrm{p}) = \cos(f_p, f_q) \quad (7)$$

Where $\cos(f_p, f_q)$ is the cosine similarity of f_p and f_q.

3.4 Question similarity

Based on the domain knowledge similarity and latent topic similarity between questions, we employ a linear equation to derive the final similarity between two questions:

$$similarity(p, q) = \lambda * similarity_{topic}(p, q) + (1 - \lambda) * similarity_{domain}(p, q) \quad (8)$$

where p and q are two questions in the corpus; $\mathrm{similarity}_{\mathrm{topic}}$ is the topic similarity between p and q; $similarity_{domain}$ is the domain knowledge similarity between p and q; λ is an adjustable parameter between topic similarity and domain knowledge similarity, which ranges from 0.0 to 1.0.

4. Experiments

4.1 Data Acquisition

In order to evaluate our method, we first crawled totally 131585 online health question-answer pairs from XunYiWenYao (www.xywy.com) generated between May 2015 and May 2017. This is the dataset1. The dataset1 is used to find the appropriate model parameters and to do the performance experiments. At the same time, we crawled more additional health question-answer pairs from GoodDoctor (www.haodf.com) generated between January 2017 and May 2017 to train the Word2Vector model. This is the dataset2. Both the dataset2 and dataset1 are used to train the Word2Vector model to obtain the word-embedding of terms in the corpus.

Meanwhile, we randomly selected 600 questions from dataset1 as target query questions. To obtain the ground truth, we pooled up 50 similar questions for each target query question by using various models, such as vector space model, and topic-based model (Zhang *et al.*, 2014). We then asked two annotators who were not involved in the design of our method, to independently annotate if the candidate similar questions are really similar to the target query question. And the dissimilar questions will be removed from the candidate similar questions. When the conflict happens, another annotator will make the final decision to decide if the candidate question should be removed from the candidate list.

4.2 Experiment results

4.2.1 Evaluation metrics. We use P@H (including P@1 and P@5) MAP (mean average precision) as evaluation metrics, which are used widely in previous studies (Jeon *et al.*, 2005; Wang *et al.*, 2009; Zhang *et al.*, 2014). P@H represents the query precision and only considers the top H relevant results. The calculation of P@H is given as follows:

$$\text{P@H} = \frac{numbers\ of\ relevant\ results\ in\ top\ H}{H}. \tag{9}$$

Normally, H is predefined and represents the first H retrieved similar questions. MAP is adopted to evaluate the precision of a set of queries and is the mean of the average precision scores for each query. The average precision score of the ith query is defined as follows:

$$\text{AveP (i)} = \frac{\sum_{m=1}^{R}\left(P(m) * rel(m)\right)}{number\ of\ relevant\ results} \tag{10}$$

In which, m is the rank in the sequence of retrieved questions, R is the number of retrieved questions, P (m) is the query precision at cut-off m in the list and is equal to P@m, and rel (m) is an indicator function equaling 1 if the question at rank m is a relevant question, zero otherwise. Then MAP is defined as:

$$\text{MAP} = \frac{\sum_{i=1}^{M}\text{AveP(i)}}{\text{M}} \tag{11}$$

In which, M is the number of queries, and AveP (i) is the average precision of query i, which is given in the formula (10).

4.2.2 Data analysis and results. To evaluate the effectiveness of our method, we compare the experiment results of our method with two baseline methods: LDACF and TEII. LDACF is a LDA guided clustering approach. It first uses LDA topic model to transfer the questions into topic vectors. Based on the lexical feature and topic vectors, it uses an unsupervised approach to cluster questions. Finally, the approach ranks the similar questions by word overlapping factors and Levenshtein distance of PoS and words (W.-N. Zhang *et al.*, 2014). TEII is a Topic Enhancing Inverted Index method which incorporates the topic information into inverted index (i.e. TFIDF) for top-k document retrieval (Jiang *et al.*, 2015). By combining the topic similarity score with the traditional TF-IDF similarity score, TEII enhances the similar document retrieval. Both methods incorporate the topic information and are similar to our method.

Table 1. Performance comparison with other methods

Evaluation metric	our method	LDACF	TEII	our method-LDACF	our method-TEII
MAP	0.443	0.362	0.332	0.081*	0.111**
P@1	0.570	0.478	0.403	0.092**	0.167**
P@5	0.360	0.283	0.280	0.077**	0.080**

Notes: $^{**}p < 0.01$; $^{*}p < 0.05$

Before the comparison, we first do the preprocessing including word segmentation and PoS tagging on the dataset1, and then extract the health named entities. Then we use the two benchmark methods and our proposed method to retrieve similarity questions with the selected 600 query questions on dataset1, respectively. Third, we compare the experiment result of our method with that of the two benchmarks. Finally, we conduct paired T-tests to examine whether the improvements of our method over LDACF and TEII are statistically significant. Table 1 shows the comparison on MAP and P@K (K = 1, 5) between our method and two baseline methods on dataset1.

The results show that our method is better than the baseline methods significantly. Compared with LDACF method, the proposed method reaches a higher precision and the MAP and P@1 increases nearly 0.1. TEII is the method with the worst performance among three methods. The proposed method has a great improvement across three metric indicators compared with TEII. Moreover, we use the parameter λ to balance between topic similarity and domain knowledge similarity. We also test the performance of different value of λ. We find that the experiment has the best performance when λ is equal to 0.2. As the increase of λ, the experiment performance become worse.

5. Conclusions

In this paper, we propose a novel hybrid method to mine similar questions in online health communities by combining domain knowledge similarity with latent topic similarity. The domain knowledge similarity evaluates the similarity of two questions from the view of domain relationship. We apply the CRF model to recognize health named entities, i.e. domain information, from the question. Then the word move distance is adopted to compute the domain knowledge similarity. The LDA method is used to extract topic from questions and topic similarity between questions is derived. Experiment results show that our method outperforms baseline methods.

There are also several limitations of this study. First, because our work are based on the named entity extraction, the result of named entity extraction may influence experiment results. The method this study adopts is the widely used named entity extraction method. In future work, we plan to improve the named entity extraction performance to enhance our method. Second, we only choose two existing methods as benchmarks. There are many other methods developed for similar question retrieval. In the future, we may compare the proposed method with other methods as well. Third, in this paper, we simply set the different terms with the same weights. However, different health terms play different roles in domain similarity evaluation. We should consider the different weight distribution of health terms in future works.

References

Banerjee, S. and Pedersen, T. (2003), "Extended gloss overlaps as a measure of semantic relatedness", Paper presented at the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Chen, L., Jose, J.M., Yu, H., Yuan, F. and Zhang, D. (2016a), "A semantic graph based topic model for question retrieval in community question answering", Paper presented at the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA.

Chen, L., Jose, J.M., Yu, H. and Yuan, F. (2016b), "A hybrid approach for question retrieval in community question answering", *The Computer Journal*, Vol. 60 No. 7, pp. 1019-1031.

Duan, H., Cao, Y., Lin, C.Y. and Yu, Y. (2008), "Searching questions by identifying question topic and question focus", Paper presented at the Meeting of the Association for Computational Linguistics, Columbus.

Eysenbach, G. (2008), "Medicine 2.0: social networking, collaboration, participation, apomediation, and openness", *Journal of Medical Internet Research*, Vol. 10 No. 3, p. e22.

Ferrández, A. (2011), "Lexical and syntactic knowledge for information retrieval", *Information Processing and Management*, Vol. 47 No. 5, pp. 692-705.

Figueroa, A. (2017), "Automatically generating effective search queries directly from community question-answering questions for finding related questions", *Expert Systems with Applications*, Vol. 77, pp. 11-19.

Gao, Y., Xu, Y. and Li, Y. (2014), "A topic based document relevance ranking model", Paper presented at the 23rd International Conference on World Wide Web, Seoul.

Jeon, J., Croft, W.B. and Lee, J.H. (2005), "Finding similar questions in large question and answer archives", Paper presented at the 14th ACM International Conference on Information and Knowledge Management, Bremen.

Jiang, D., Leung, K.W.-T., Yang, L. and Ng, W. (2015), "TEII: topic enhanced inverted index for top-k document retrieval", *Knowledge-Based Systems*, Vol. 89 No. 89, pp. 346-358.

Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K. (2015), "From word embeddings to document distances", Paper presented at the 32nd International Conference on International Conference on Machine Learning, Lille.

Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M. and Xu, H. (2014), "A comprehensive study of named entity recognition in Chinese clinical text", *Journal of the American Medical Informatics Association*, Vol. 21 No. 5, pp. 808-814.

Lian, X., Yuan, X., Hu, X. and Zhang, H. (2013), "Finding similar questions with categorization information and dependency syntactic tree", Paper presented at the International Conference on Web-Age Information Management.

Liu, D.-R., Chen, Y.-H. and Huang, C.-K. (2014), "QA document recommendations for communities of question–answering websites", *Knowledge-Based Systems*, Vol. 57 No. 2, pp. 146-160.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S. (2013), "Using of jaccard coefficient for keywords similarity", Paper presented at the International MultiConference of Engineers and Computer Scientists.

Roberts, K. and Demner-Fushman, D. (2016), "Interactive use of online health resources: a comparison of consumer and professional questions", *Journal of the American Medical Informatics Association*, Vol. 23 No. 4, pp. 802-811.

Samuel, H., Kim, M.-Y., Prabhakar, S., Jabbar, M.S.M. and Zalane, O. (2017), "Community question retrieval in health forums", Paper presented at the 2017 IEEE International Conference on Biomedical and Health Informatics (BHI).

Uzuner, Ö., South, B.R., Shen, S. and Duvall, S.L. (2011), "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text", *Journal of the American Medical Informatics Association*, Vol. 18 No. 5, pp. 552-556.

Van De Belt, T.H., Engelen, L.J., Berben, S.A. and Schoonhoven, L. (2010), "Definition of health 2.0 and medicine 2.0: a systematic review", *Journal of Medical Internet Research*, Vol. 12 No. 2, pp. 1-14.

Wang, K., Ming, Z. and Chua, T.-S. (2009), "A syntactic tree matching approach to finding similar questions in community-based qa services", Paper presented at the 32nd International Conference on Research and Development in Information Retrieval, Boston, MA.

Wu, M.S. (2015), "Modeling query-document dependencies with topic language models for information retrieval", *Information Sciences*, Vol. 312, pp. 1-12.

Wu, H.C., Luk, R.W.P., Wong, K.F. and Kwok, K.L. (2008), "Interpreting TF-IDF term weights as making relevance decisions", *ACM Transactions on Information Systems*, Vol. 26 No. 3, pp. 1-37.

Yan, Z., Wang, T., Chen, Y. and Zhang, H. (2016), "Knowledge sharing in online health communities: a social exchange theory perspective", *Information and Management*, Vol. 53 No. 5, pp. 643-653.

Yang, J., Yu, Q., Guan, Y. and Jiang, Z. (2014), "An overview of research on electronic medical record oriented named entity recognition and entity relation extraction", *Acta Automatica Sinica*, Vol. 40 No. 8, pp. 1537-1562.

Zhang, W.N., Ming, Z.Y., Zhang, Y., Liu, T. and Chua, T.S. (2016), "Capturing the semantics of key phrases using multiple languages for question retrieval", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28 No. 4, pp. 888-900.

Zhang, W.-N., Liu, T., Yang, Y., Cao, L., Zhang, Y. and Ji, R. (2014), "A topic clustering approach to finding similar questions from large question and answer archives", *PloS One*, Vol. 9 No. 3, p. e71511.

Further reading

Rhebergen, M.D. and Hulshof, C.T. (2010), "An online network tool for quality information to answer questions about occupational safety and health: usability and applicability", *BMC Medical Informatics and Decision Making*, Vol. 10 No. 1, p. 63.

Corresponding author

Zhijun Yan can be contacted at: yanzhijun@bit.edu.cn

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**