Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2014, *8,* 1033-1037

Open Access

Question Retrieval Based on Probabilistic Latent Semantic Analysis in Q & A Community

Chengfang Tan[1, 2,*], Hong Li[1, 2], Yundong Liu[1, 2] and ZhenggaoPan[1, 2]

[1]*School of Information Engineering, Suzhou University, Suzhou Anhui, 234000, China;* [2]*Intelligent Information Processing Lab, Suzhou University, Suzhou, Anhui, 234000, China*

Abstract: With the increasingly popularity of Q & A Community, it has become an important means for people to retrieve question from question library to find the answer. Similarity calculation is the core issue in Q & A community, and the appropriate calculation method is the key factor that affects the quality of question retrieval. This paper proposes a retrieval method based on PLSA model. Firstly, we modelled the question library, and got the probability distribution of "question document –latent semantic -word". Secondly, we calculated the semantic similarity between questions and classify them. Finally, based on user retrieval content, we calculated the similarity between question documents and query, then the query results will be returned to the user in descending order according to the value. Compared with other similarity calculation methods that use VSM, HNS and SD, the experimental results show that this proposed method has a high precision rate.

Keywords: Probabilistic latent semantic analysis, Q & A community, semantic analysis, similarity calculation.

1. INTRODUCTION

In recent years, with the rapid development of Web2.0, Q & A community is becoming more and more popular, which attracts a large number of users and has accumulated massive information. Q & A community provides the question and answer network platform for the user. At present, the common Q & A communities have Yahoo! Answers, Baidu Knows, Sina Love Asked Knowledge, etc. In these Q & A communities, every day there are some users raise lots of questions, while other users retrieve the existing question library to find results or answer some questions. The content of Q & A community is generated spontaneously by the user, which leads to a large number of similar or the same questions exist in it. Therefore, in order to quickly and accurately obtain the needed information for the user, it is necessary to classify the questions and divide the similar questions into a class, which enables user to easily find a class of interested questions.

The study on Q & A community is mainly focused on similarity calculation and answer quality evaluation. Adamic *et al.* comprehensively analyzed the knowledge sharing behaviour in Yahoo! Answers community, examined what is the best answer, and verified the credibility by the asker to choose the best answer [1]. Cheng Zhang used the probability latent semantic analysis method to establish auto answer recommendation mechanism [2]. Jeon *et al.* proposed the machine translation method to retrieve and identify the similar questions in Q & A community, but their work did not consider the answer quality [3]. Agichtein *et al.* used classification framework method to integrate all kinds of text information in Q & A community, studied the prediction of text quality and user satisfaction [4]. Against Baidu Knows, WeiZe Kong at al proposed the feature based on timing, based on the size of question and based on the community users to evaluate the answer quality from multiple perspectives, which can efficiently predict the best answer [5]. Although lots of research has been done on Q & A community, but research on question retrieve is relatively less.

This paper applies the probabilistic latent semantic analysis technology to Q & A community, and use probabilistic model to represent the relationship between "question document-latent semantic-word", which maps question documents and words to the same semantic space, so the similarity between question documents and words can be quantified by calculating the angle of semantic space.

2. THORY OF PROBABILISTIC LATENT SEMANTIC ANALYSIS

Thomas Hofmann proposed probabilistic latent semantic analysis (PLSA) in1999 [6]. The PLSA model has important advantages. On the one hand, synonymy and polysemy can get reasonable representation, on the other hand, by calculating the distance between vectors in the semantic space, this makes many text information processing applications have been quantified to solve.

The PLSA model is from the perspective of probability to analyze the latent semantic of documents. For the given documents set x $D=\{d_1, d_2, \cdots, d_i\}$, words set $W=\{w_1, w_2, \cdots, w_j\}$ and co-occurrence between document and word $A=|a_{ij}|_{n\times m}$, a_{ij} represents the weight of the word w_j

1874-110X/14

in the document d_i. $Z=\{z_1, z_2, \cdots, z_k\}$ represents the latent topics set, where K is a constant that represents the number of latent topics [7]. Figurative representation for PLSA model is as shown in Fig. (**1**).

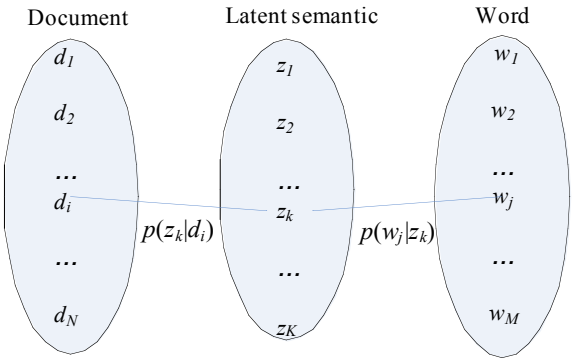

Fig. (1). Representation of probabilistic latent semantic analysis model.

The PLSA model assumes that the distribution between word and document is independent, and the distribution of latent topic in words or documents is also independent. The conditional probability of "word-document" is expressed as follows.

$$p(d_i \mid w_j) = p(d_i)p(w_j \mid d_i) \tag{1}$$

$$p(w_j \mid d_i) = \sum_{k=1}^{k} p(w_j \mid z_k)p(z_k \mid d_i) \tag{2}$$

where $p(d_i)$ represents the probability of document d_i, $p(w_j \mid z_k)$ represents the probability distribution of latent topic z_k on words. Performing the sort of $p(w_j \mid z_k)$, we can get a visual representation of words in latent topic z_k. $p(z_k \mid d_i)$ represents the latent semantic probability distribution of document d_i. In order to explain this group point of view, we need to estimate the parameters $p(w_j \mid z_k)$ and $p(z_k \mid d_i)$. The PLSA model uses the expectation maximization algorithm that is EM algorithm to fit the latent semantic model, by initializing the random number and performing iterative calculation on step E and step M in turn. In step E, it uses the formula (3) to calculate the prior probability that each (d_i, w_j) generates latent semantic z_k.

$$p(z_k \mid d_i, w_j) = \frac{p(w_j \mid z_k)p(z_k \mid d_i)}{\sum_{k=1}^{k} p(w_j \mid z_k)p(z_k \mid d_i)} \tag{3}$$

In step M, it uses the formula (4) and (5) to re-estimate the model.

$$p(w_j \mid z_k) = \frac{\sum_{i=1}^{n} n(d_i, w_j)p(z_k \mid d_i, w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{n} n(d_i, w_j)p(z_k \mid d_i, w_j)} \tag{4}$$

$$p(z_k \mid d_i) = \frac{\sum_{j=1}^{m} n(d_i, w_j)p(z_k \mid d_i, w_j)}{n(d_i)} \tag{5}$$

The state of stopping iteration is when the increasing amount of expected value is less than a threshold, at this time we will get an optimal solution.

$$E(L) = \sum_{i=1}^{n}\sum_{j=1}^{m} n(d_i, w_j)\sum_{k=1}^{k} p(z_k \mid d_i, w_j) \log\left[p(w_j \mid z_k)p(z_k \mid d_i)\right] \tag{6}$$

Because of the very large amount of calculation for EM algorithm, in order to reduce the complexity of the algorithm, this paper uses the iterative calculation method of one thousand cycles to get an approximate value, the value that is considered to be the optimal solution.

3. APPROACH OF QUESTION RETRIEVAL BASED ON PROBABILISTIC LATENT SEMANTIC ANALYSIS

3.1. Text Preprocessing

Chinese segmentation and removing stop words are important factors to be considered when preprocess questions of Q & A community, especially the Chinese segmentation. The main issues to be solved include the common word vocabulary, segmentation specification, segmentation ambiguity and so on. For Chinese word segmentation in this paper, we use ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) which is the most widely used in the field of Chinese word.

3.2. Construct the Matrix of "Question Document-Word"

We statistic the number of appearances that each word of document set appears in each question document, and then obtain the matrix $N(d, w)$ that represents the N * M dimensional "question document-word" matrix. By using TF-IDF formula to normalize this matrix, the influence caused by the different length of documents can be eliminated.

$$a_{ij} = \frac{tf_{ij} * \log(\mathrm{N}/\,\mathrm{n}_i + 0.01)}{\sqrt{\sum_{k=1}^{n}\left[tf_{ik} * \log(\mathrm{N}/\,\mathrm{n}_i + 0.01)\right]^2}} \tag{7}$$

where a_{ij} represents the weight of the word i in the question document j, and tf_{ij} represents the number of frequency that the word i appears in the question document j, N represents the total number of question document, n_i represents the number of frequency of question document that contains the word i, the denominator is the normalization factor.

3.3. Construct PLSA Model

Defining a K-dimensional semantic space Z, probabilistic latent semantic analysis model $U= (P(z_k|d_i))_{k,i}$ is constructed according to the representation of the matrix. U represents the probability distribution of latent semantic in ques-

tion documents. Matrix $V=(P(w_j|z_k))_{j,k}$ represents the probability distribution of latent semantic in keywords. Randomly generating each component of U and V, it makes the accumulation of each row of the matrix U and each column of the matrix V equal to 1.

EM algorithm is used to optimize the entire model. In step E, according to the current value of the matrix U and the matrix V, we use the equation (3) to calculate the prior probability of each (d_i, w_j) for generating the latent semantic z_k. In step M, we use the formula (4) and (5) respectively to re-estimation the probability value on the matrix U and the matrix V, perform steps E and steps M alternately until the value of the formula (6) begin to converge. So the probabilistic latent semantic model for Q & A community is constructed.

3.4. Calculate The Similarity of Questions And Classify Them

According to the following procedure to process each question document q that is to be classified. First of all, the document q needs to be preprocessed with removing stop words, word segmentation and so on. Based on extracted keywords, the keyword vector $n(q,w_j)$ of question document can be constructed and the weight also can be calculated. Then, EM algorithm is used for $p(z_k \mid q)$ to get $p(z \mid q)$ that is the probability distribution of the probabilistic latent semantic in question document q that is to be classified. $p(z \mid c)$ represents the center probability vector of all kinds of document sets, and the similarity between $p(z \mid q)$ and $p(z \mid c)$ can be calculated, namely the distance between vectors in semantic space. According to the size of the distance, the question document will be divided into the corresponding category. The similarity calculation is expressed as follows.

$$sim(z_q \mid z_c) = \frac{\sum_k p(z_k \mid q) p(z_k \mid c)}{\sqrt{\sum_k \left[p(z_k \mid q)\right]^2} \sqrt{\sum_k \left[p(z_k \mid c)\right]^2}} \tag{8}$$

3.5. Calculate The Similarity of "Question Document - Query"

For a query task, the first step is to construct the query vector $n(q,w_j)$. Performing EM algorithm on $P(z_k|q)$, we can get the probability distribution $p(z|q)$ of latent semantic in query text. Then by calculating the cosine of $p(z_k|d)$ and $p(z_k|q)$, we can obtain the similarity between document and query, which is expressed as follows:

$$C\text{os}S\text{im}(x,y)\frac{\sum_{i=1}^{n} w_{xi} * w_{yi}}{\sqrt{\sum_{i=1}^{n} w_{xi}^2} \sqrt{\sum_{i=1}^{n} w_{yi}^2}} \tag{9}$$

where n is the dimension of vector space, w_{xi} is i-dimensional component of query vector x, w_{yi} is i-dimensional component of document vector y.

The algorithm idea of "question document -query" similarity calculation is described as follows.

Input: user query q, question document d of question library, the similarity threshold u.

Output: similar questions sequence.

Step 1 perform text preprocessing for user query q and select keywords.

Step 2 find out the category Q_i in which the q falls, the category is consisted of a number of specific questions, namely $Q_i=\{q_1, q_2, \ldots, q_n\}$.

Step 3 calculate the semantic similarity between q and q_i, that is $CosSim\ (q, q_i)$.

Step 4 based on the semantic similarity $CosSim\ (q, q_i)$, sort them from high to low, select Top N question documents from the category Q_i and then return them to the user.

4. EXPERIMENTS AND RESULT ANALYSIS

4.1. Experimental Data

This experimental data is from the Q & A community "Baidu Knows". Since data scarcity has an influence on PLSA model, so we selected the question category which is relatively popular. Using web crawler to crawl questions of five categories, there are computer, education, business, life and entertainment, a total of 9496 question documents which were used for the construction of question library, the specific distribution of questions is shown in Table **1**. 2/3 of question library were used for training corpus and 1/3 for testing corpus. We modelled the question library based on PLSA model and classified questions, and further divided them into 30 sub-categories. Then by using the similarity calculation method of "question document-query" proposed in this paper, we calculated the cosine of $p(z_k|d)$ and $p(z_k|q)$ to obtain the similarity of document and query, and determined the final returned question documents sequence for the user.

Table 1. The distribution of corpus.

Category	The Number of Questions
Computer	2145
Education	1439
Business	1991
Life	1904
Entertainment	2017

4.2. Evaluation Index

There are many evaluation standards for text information processing results. Among them, the most common is the precision rate and the recall rate. Because of the particularity of the Q & A community, the recall rate can be ignored in the analysis results. Therefore, this paper uses the precision rate as evaluation index, which is expressed as follows.

$$p = \frac{a}{a+b} \tag{10}$$

Table 2. The probability of words in each topic.

The Topic "Black Screen"		The Topic "Apple 6"	
Feature word	**Probability**	**Feature word**	**Probability**
Computer	0.462	Phone	0.378
Boot	0.213	IPhone6	0.157
Monitor	0.087	Price	0.081
Malfunction	0.049	Function	0.056
……		……	

Table 3. Similarity calculations under different methods.

Question in Q & A Community	VSM	HNS	SD	PLSA
What is the suitable quote for Apple 6?	0.58	0.73	0.79	0.84
How to buy Apple 6	0.41	0.62	0.74	0.79
When is Apple 6 on the market?	0.27	0.42	0.53	0.51

where a represents the number of question documents which are assigned to certain category correctly, b represents the number of misclassified question documents in the classification results

4.3. Other Methods of Experiment Comparisons

This paper proposes a similarity calculation method based on PLSA model, which is referred to as PLSA. In order to verify the effectiveness of our proposed method, experiments were carried out by comparing with the following calculation methods.

(1) The similarity calculation based on VSM (referred to as VSM)

Each question in Q & A community can be expressed with an *N*-dimensional vector. The feature value of vector is obtained by calculating the TF-IDF. Set the questions $S_1 = (T_1, T_2, \cdots, T_n)$ and $S_2 = (T_1^{'}, T_2^{'}, \cdots, T_n^{'})$, the similarity between S_1 and S_2 can be calculated as follows:

$$S\,\mathrm{im}(S_1, S_2) = \frac{\sum_{i=1}^{n}(T_i \times T_i^{'})}{\sqrt{\sum_{i=1}^{n} T_i^2}\sqrt{\sum_{i=1}^{n} T_i^{'2}}} \tag{11}$$

(2) The similarity calculation based on HowNet semantic (referred to as HNS)

Assume that A_i and B_j respectively represent the keyword of question S_1 and S_2, the similarity $S(A_i,B_j)$ can be calculated based on HowNet [8]. Set a_i=max($S(A_i,B_1)$, $S(A_i,B_2)$, …, $S(A_i,B_n)$) and b_i=max($S(B_i,A_1)$, $S(B_i,A_2)$, …, $S(B_i,A_n)$). The similarity between S_1 and S_2 can be calculated as follows:

$$S\,\mathrm{im}(S_1, S_2) = \frac{(\frac{\sum_{i=1}^{m} a_i}{m} + \frac{\sum_{i=1}^{n} b_i}{n})}{2} \tag{12}$$

(3) The similarity calculation based on semantic dependency (referred to as SD)

Dependency parsing analysis is used for the question firstly. Through the effective match in dependency syntactic tree, the syntax similarity $Sim_1(S_1,S_2)$ and the semantic similarity $Sim_2(S_1,S_2)$ can be calculated respectively [9]. And finally the similarity between questions can be obtained by weight calculation, which is expressed as follows.

$$Sim(S_1,S_2) = \lambda Sim1(S_1,S_2) + (1-\lambda) Sim2(S_1,S_2) \tag{13}$$

4.4. Analysis of Experimental Results

After running the PLSA modelling program, we can get the probability distribution of feature words of question documents in each latent topic. Due to limited space, this paper take the topic "black screen" and the topic "Apple 6" as examples, the probability of words in each topic is as shown in Table **2**.

Table **2** only shows part of the information on each topic, that is part of keywords and corresponding probabilistic information. As can be seen from Table **2**, the topics obtained from PLSA model have higher accuracy on the corresponding keywords, and independence between topics is relatively stronger. It is easy to find out the related topic according to the keywords.

Taking the user question “how much is the Apple 6?" as an example, we calculated the similarity between related sentences in question library with different calculation methods. The result is shown as Table **3**.

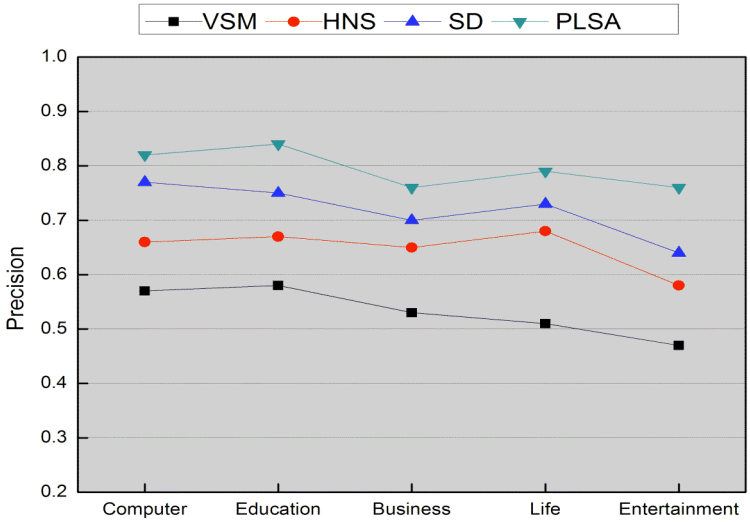

Fig. (2). The precision comparison of four different experimental methods.

From Table **3**, we can see the comparison results. Although the structure of questions has certain difference, but the calculation result will be better when the semantic between questions is similar.

Comparative experiments were carried out based on different similarity calculation methods, which are the method of VSM, HNS, SD and PLSA. We compared the precision rate p under five categories of computer, education, business, life and entertainment. The experimental result is as shown in Fig. (**2**).

As can be seen from Fig. (**2**), since the probabilistic latent semantic analysis introduces the concept of "latent topic", the proposed similarity calculation method based on PLSA model in this paper can get a better precession than the other three experimental methods.

CONCULSION

In this paper, PLSA model is applied to question retrieval of Q & A community. The relationships of "question document-latent semantic-word" can be calculated by using probability model, which realizes the latent semantic analysis for questions. Compared with other similarity calculation methods, the experiments results verify the effectiveness of our proposed method based on PLSA model, which can effectively improve the retrieval precession of Q & A community. The word order of sentence has a certain influence on similarity calculation, so the next step we will research on the word order of sentence, and try to design an effective algorithm for the similarity of word order.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Key University Science Research Project of Anhui Province (No. KJ2014A250) Software engineering projects (No. 2013zytz074) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No. 2014YKF41, No. 2013YKF19, No. 2011YKF10).

REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshye, and S. A. Mark, "Knowledge sharing and yahoo answers: everyone knows something", *Proceedings of the 17th International Conference on World Wide Web*, Beijing, pp. 665-674, 2008.

[2] C. Zhang, M. C. Qu, N. Ni, G. Qiu, and J. Bu, "Automatic answer selection based on probabilistic latent semantic analysis model", *Computer Engineering*, vol. 27, no. 14, pp. 70-74, 2011.

[3] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives ", *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 84-90, 2005.

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media", *Proceedings of the International Conference on Web Search and Web Data Mining, ACM*, pp. 183-194, 2008.

[5] W. Kong, Y. Liu, M. Zhang, and S. Ma, "Answer quality analysis on community question answering", *Journal of Chinese Information Processing*, vol. 25, no. 1, pp. 3-8, 2011.

[6] H. Thomas, "Pobabilistic latent semantic analysis", *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Stockholm: [s.n.]*, pp. 289-296, 1999.

[7] J. Luo, and X. Tu, "Chinese information retrieval based on probabilistic latent semantic analysis", *Computer Engineering*, vol. 34, no. 2, pp. 199-201, 2008.

[8] S. Li, "The research of relevancy between sentences based on semantic computation", *Computer Engineering and Applications*, vol. 38, no. 7, pp. 75-76, 2002.

[9] B. Li, and T. Liu, B. Qin, and S. Li, "Chinese sentence similarity computing based on semantic dependency relationship analysis", *Application Research of Computers*, vol. 20, no. 12, pp. 5-17, 2003.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014