

commendation, established the corresponding relation between music and emotion, and verified the feasibility of the system through studying the recommendation accuracy, response time and comfort level.

Acknowledgments

This work was supported by project of Technology Department of JiangXi Province [No 20143BBM26048] and project of The Education Department of JiangXi Province [No GJJ14765], the project [20151BBE50079] also give lots of supports.

References

1. Wang Xin (2012) Meta-analysis on intervention effect of music on anxiety symptom, *China Music*, No.4, p.p.201-208.
2. Wang Huan (2014) Music and sleep, *Drama forum*, No.6, p.p:87-89.
3. Qin Ruisheng (2011) Music therapy: enlighten the soul of mentally disordered children, *Journal of Social Welfare*, No.1 ,p.p:48-50
4. Li Xue (2015) Study on music and emotion regulation, *Journal of Jilin College of The Arts*, No.1, p.p:7-10.
5. Eric Jensen. (2005) *Art education and brain development*[M]. Beijing: China light industry press, 2005. .
6. Knutzen H, Kvifte T, Wanderley M M. (2014) Vibrotactile Feedback for an Open Air Music Controller. *Sound, Music, and Motion. Springer International Publishing*, p.p.41-57.
7. McCauley, Jack J., Brian Bright, and John Devecka (2014) "Music video game and guitar-like game controller." U.S. Patent No. 8,827, 806. 9 Sep. 2014.
8. Mays T, Faber F. (2014) A Notation System for the Karlax Controller. *Proceedings for New Interfaces for Musical Expression*, p.p:100-104.
9. Mutthuraju K S, Vijaya P A.(2015) Design and Implementation of Rover Controller and Music Player Control using Sixth Sense Technology. *International Journal of Engineering Research and Technology*, 4(5), p.p:54-60.
10. Barrington L, Oda R, Lanckriet G R G. (2009) Smarter than Genius? Human Evaluation of Music Recommender Systems. *ISMIR*, No.9, p.p.357-362.



Duplicate Questions Removal Based on LDA in Q&A Community

Shoufei Gan, Shizhen Lu, Chengfang Tan

School of Information Engineering, Suzhou University, Suzhou 234000, Anhui, China

Abstract

There are a large number of similar or duplicate questions exist in Q&A community, resulting in poor efficiency retrieval and other issues. This paper puts forward a kind of removing duplicate questions method based on LDA, fully considering deep semantic knowledge of questions. First of all, we use LDA to model for question set, and execute parameter inference by Gibbs sampling of MCMC to calculate model parameters indirectly. Then through mining the hidden relationship between different

topics and words in questions, the probability distribution of topic and word can be obtained, which is used to calculate the similarity between questions. Lastly, setting a similarity threshold, questions which have high degree of duplication will be screened and removed. Compared with other traditional similarity calculation methods, experimental results show that the proposed similarity calculation method has a better precision rate and obtains good effect on removing duplicate questions.

Key words: Q&A COMMUNITY, LATENT DIRICHLET ALLOCATION, SIMILARITY CALCULATION, DUPLICATE QUESTIONS REMOVAL

1. Introduction

With the growing prosperity and the explosive growth of data access in Q&A community, it has gradually become an important medium for Internet users on information transmission and knowledge sharing, such as Yahoo! Answers, Baidu knows, and other community sites. Q&A community releases tens of thousands of questions every day. Common questions and corresponding answers are usually stored in the database. User can search the similar question to get corresponding answers directly, which can save a lot of time. However, with the growing number of questions in Q&A community, there are many duplicated questions, which seriously affect the user to obtain the required information quickly and accurately. Therefore, a key study is how to remove duplicate questions from Q&A community.

Many theories and algorithms have emerged in Q&A community. R.D.Burke et al. [1] and V.Jijkoun et al. [2] used vector space model to calculate angle cosine between the query question vector and the candidate question vector. M.Collins et al. [3] proposed a tree method to compare the similarity between syntactic trees by calculating the number of the same tree fragments between two syntactic trees. J.Jeon et al. [4] estimated the similarity between two questions through calculating the similarity of two answers, but without taking into account syntactic and semantic-features. Meanwhile, according to the time spent, the

quality of answer and other factors, some scholars predicted user satisfaction for answers [5, 6].

This paper proposes a kind of method to remove duplicate questions from Q&A community. It uses LDA model to express the relationship between question, topic and word. So questions and words will be mapped to the same semantic space. The similarity between questions can be calculated by using the topic information of questions. Based on this, we can screen and remove questions with the high degree of duplication, thereby reducing duplicate questions in the retrieval results to improve information retrieval efficiency.

2. Theory of LDA model

LDA model is proposed by Blei et al. in 2003 [7], which is based on LSA (Latent Semantic Analysis) and PLSA (Probability Latent Semantic Analysis). It belongs to a kind of implicit variable model. LDA model is trained by unsupervised method, and is independent of the number of training samples, so it is more suitable for dealing with large-scale text corpus. The model consists of three layers of word, topic and document, which is a three layer Bayesian probability model. The main idea of this model is that it represents a document into the probability distribution of each topic, and each topic is represented as a probability distribution of different words in all documents. LDA is a probability model, which is shown in Figure 1.

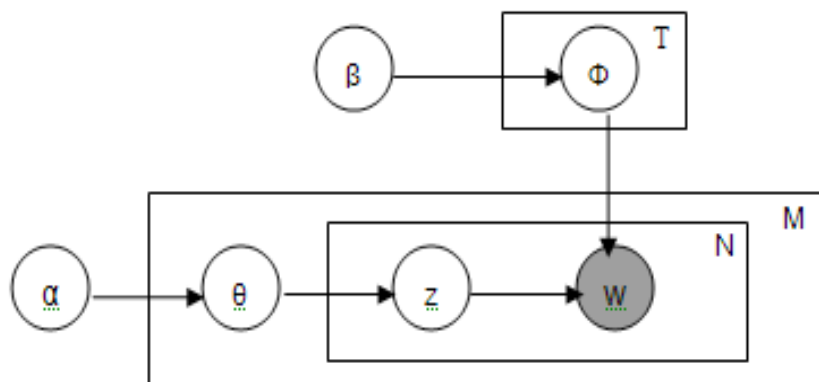


Figure 1. LDA model

The meaning of each symbol in LDA model diagram is shown in Table 1.

Table 1. The meaning of each symbol

Symbol	Meaning	Symbol	Meaning
α	Super parameter of θ	w	Word
β	Super parameter of Φ	N	The number of words
θ	Document-topic probability distribution	T	The number of topics
Φ	Topic-word Probability distribution	M	The number of documents
z	Word probability distribution		

LDA model describes the process of generating word in the document based on the latent topic. The model is determined by two parameters α and β , where α reflects the relative strength of implicit topics in the document and β represents its probability distributional of implicit topics.

3. Duplicate questions removal based on LDA

3.1 LDA modeling

LDA model uses probability way to generate model. In order to model the question set, a question is seen as a document, and each question can be expressed as a series of topics mixture distribution, denoted as $p(z)$. At the same time each topic is the probability distribution of all the words, denoted as $p(w|z)$. Therefore, the probability distribution of each word in a question can be calculated as follows:

$$p(w_i) = \sum_{j=1}^T p(w_i | z_i = j)p(z_i = j) \tag{1}$$

Where z_i is a latent variable, $p(w_i|z_i = j)$ represents the probability of the word w_i that belongs to the j th topic, $p(z_i = j)$ represents the probability of the document d that belongs to the j th topic.

The process of training to generate text by LDA model in question set is as follows:

A word polynomial distribution $\phi^{(t)}$ is obtained by extracting the relationship between topic and word from *Dirichlet*(β).

A topic polynomial distribution θ_d is obtained by extracting the relationship between question and topic from *Dirichlet*(α).

For each word w_i in each document, a topic t is extracted from the topic polynomial distribution θ_d , and then a word w_i is extracted from the word polynomial distribution $\phi^{(t)}$ on this topic.

3.2 Parameter estimation

In the process of building LDA model, it needs to estimate model parameters. Common estimation methods are mainly variation Bayesian inference, desirable propagation algorithm and Gibbs sampling, etc. Among them, Gibbs sampling method is easy to understand and implement, which can effectively extract topics from large text set. Therefore, Gibbs sampling algorithm has become the most popular extraction algorithm of LDA model.

In the LDA model, the most important two parameters are the probability distribution of each topic and the probability distribution of each document. This paper estimates model parameters by using Gibbs sampling algorithm based on MCMC, and indirectly calculates the probability distribution of topic and the probability distribution of word by Gibbs sampling. The calculation formula is given by [8]:

$$\phi_{z,w} = \frac{n_z^w + \beta_w}{\sum_{w=1}^V n_z^w + \beta_w} \tag{2}$$

$$\theta_{q,z} = \frac{n_q^z + \alpha_z}{\sum_{z=1}^T n_q^z + \alpha_z} \tag{3}$$

Where $\phi_{z,w}$ represents the probability of the word w in the topic z , $\theta_{d,w}$ represents the probability of the question q in the topic z , n_z^w represents the number of the word w appeared in the topic z , β_w is the Dirichlet prior of the word w , n_q^z represents the number of the topic z appeared in the question q , α_z is the Dirichlet prior of the topic z , V stands for the number of questions and T represents the number of topics.

After Gibbs sampling algorithm, we can map the

word vector space of question to the topic vector space of question, and then the result can be used as the input of question similarity calculation.

3.3 Similarity calculation

Through the construction of LDA model, we can obtain probability distribution of topic and probability distribution of word, where the probability distribution of topic is a simple mapping of question vector space. Therefore, the similarity of the two questions can be achieved by calculating the corresponding probability distribution of topics.

In this paper, we use JS (Jensen-Shannon) distance to calculate the distance of topic probability vector between $p = (p_1, p_2, \dots, p_m)$ and $q = (q_1, q_2, \dots, q_n)$, the specific calculation formula is expressed as follows:

$$D_{js}(p, q) = \frac{1}{2} [D_{KL}(p, \frac{p+q}{2}) + D_{KL}(q, \frac{p+q}{2})] \quad (4)$$

Where p and q represent the probability distribution of topic respectively,

$$D_{KL}(p, q) = \sum_{j=1}^r p_j \ln \frac{p_j}{q_j}$$

3.4 Duplicate questions removal

According to the similarity between questions in question set, it is necessary to set a similarity threshold to remove duplicate questions. The algorithm of duplicate questions removal is as follows:

Input: question set Q to be removed, $Q = \{q_1, q_2, \dots, q_n\}$, the value of similarity threshold u .

Output: question set Q' after removing duplicate questions, duplicate set R .

Step1 preprocess the question set.

Step2 construct LDA model for the question set.

Step3 calculate the similarity between the question q_i and the question q_j , namely $Sim(q_i, q_j)$.

Step4 if $Sim(q_i, q_j) > u$ then add the question q_i into duplicate set R and remove it from the question set Q .

4. Experiment and analysis

4.1 Experimental data

In this paper, experimental data was selected from

the Q &A community of Baidu knows. Since the sparsity of data will affect the LDA model, so we selected some relatively popular question category. Using network spiders to crawl questions of three categories, there is computer, education and life, a total of 5217 questions which were used for the construction of question set, the specific distribution of questions is shown in Table 2. 2/3 of question set were used for training data and 1/3 for testing data.

Table 2. The distribution of experimental data

Category	Computer	Education	Life
The number of questions	2184	1607	1426

For questions of each category, three members of the experimental group labeled them by manual to identify duplicate questions in each category.

This paper used Chinese lexical analysis system-ICTCLAS as segmentation tool. After pretreatment, the question set was modeled by LDA. The experience values of priori parameters α and β in LDA model were $\alpha = 50 / K$, $\beta = 0.01$, $K=30$. The number of Gibbs sampling iterations was set to 1000 times.

4.2 Experimental analysis

This paper adopted the precision rate P to evaluate the performance of algorithm, namely $P = \frac{A}{B}$, where A represents the number of questions that are correctly removed, and B represents the number of questions removed.

1 The determination of the similarity threshold u .

The similarity threshold u has an important impact on final removal result. In order to carry out the high quality similarity calculation without introducing too much noise data, we executed comparison experiments on different thresholds, by using artificial way to judge the result and statistic the precision rate P of three categories questions, which is as shown in Table 3.

Table 3. The precision rate on different similarity threshold

Category	Precision rate (%)				
	$u=0.5$	$u=0.6$	$u=0.7$	$u=0.8$	$u=0.9$
Computer	61.23	67.69	75.23	74.11	69.37
Education	65.87	72.35	78.61	73.08	69.29
Life	58.36	65.17	73.49	71.34	68.41

From the Table 3, we can see that the change of the threshold also brings the change of the precision rate P . When the threshold is 0.7, the precision rate P is the highest in each category, so we determined the threshold is 0.7 for followed experiments.

Based on the determined threshold, we calculated the repetition rate of question for the above three categories, repetition rate = the number of duplicate question removal / total questions. The calculation result is shown in Table 4.

Table 4. The repetition rate of three categories

Category	Computer	Education	Life
Repetition rate (%)	3.57	2.14	2.51

As can be seen from Table 4, the repetition rates of three categories are different. The reason is that the distribution of duplicate questions in Q &A community has randomness, and the selected experimental data is independent of the size.

2. Comparison experiments on the effect of dupli-

cate question removal.

Experiments were divided into three groups, respectively using the statistical model based on VSM, question classification algorithm based on HowNet, and the proposed method.

The first group experiment used the traditional VSM method to calculate question similarity, namely calculated the TF-IDF value of every word in questions to get the corresponding vector representation of questions, and then obtained the similarity by calculating the angle cosine between two vectors.

The second group experiment calculated the similarity based on HowNet semantic [9], by calculating the similarity of words in questions to get the similarity of two questions.

The third group experiment was based on the proposed LDA method. Through modeling the questions set to calculate the topic probability distribution of questions, so the similarity between two questions also can be calculated.

Experimental comparison result of three groups is as shown in Figure 2.

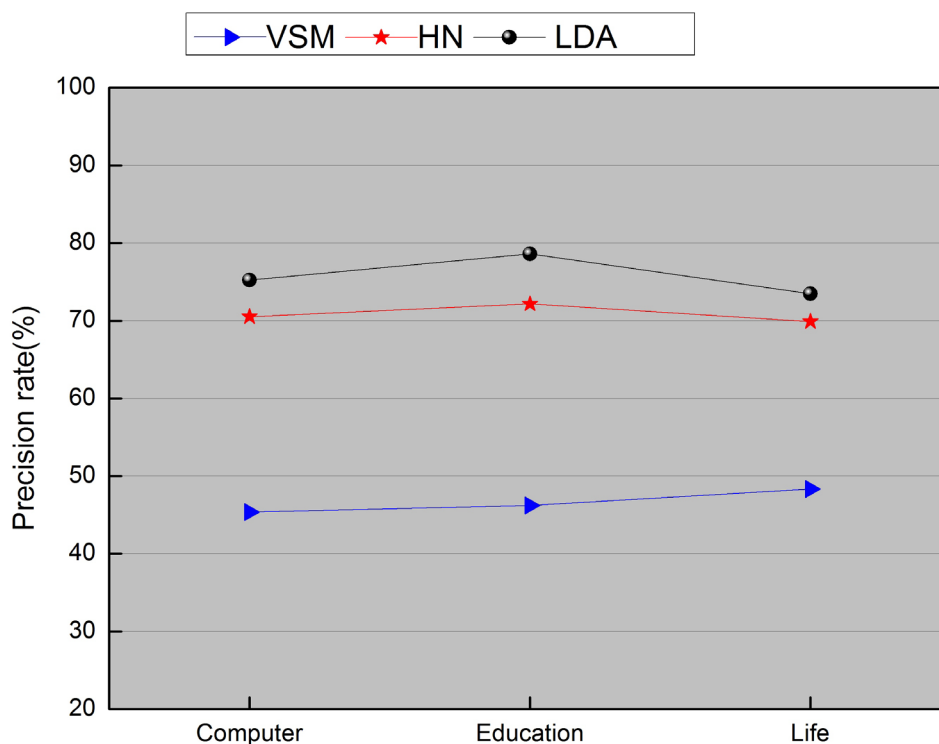


Figure 2. The comparison result of three experimental methods

As can be seen from the Figure 2, comparing with the other two experimental methods, the similarity calculation method proposed in this paper obtains a higher precision rate. This suggests that the similarity calculation algorithm based on LDA topic model ef-

ficiently enhances the semantic expression ability of question and improves the computational result.

5. Conclusions

In order to handle a large number of duplicate questions in Q&A community, according to the

characteristics of LDA topic model, this paper introduces LDA to model the question set and calculate the similarity between questions. By setting an appropriate similarity threshold value, we remove the duplicate questions in Q&A community. Experimental results show that the proposed method in this paper efficiently improves the precision rate by LDA model, reduces the data sparsity and high-dimension feature space. In addition, the whole precision rate is not high. On the one hand, the reason is that the question is a short text and irregularity, on the other hand, the result is influenced by its quality of divided experimental data. So how to improve the precision and efficiency of question feature is the focus of the next research.

Acknowledgements

This work was supported by Key University Science Research Project of Anhui Province (No. KJ2014A247) and Major University Science Research Project of Anhui Province (No. KJ2014ZD31) and University Students Innovative Undertaking Training Program of China (No. 201410379003) and University Students Research Project Suzhou University (No. KYLXLKYB15-28)

References

1. Burke R D, Hammond K J, Kulyukin V A, et al. (1997) Question answering from frequently asked question files: experiments with the faq finder system. *AI Magazine*, 18(2), p.p.57-66.
2. Jijkoun V, Rijke M. (2005) Retrieving answers from frequently asked questions pages on the web. *Proc. Conf. on Information and Knowledge Management*, ACM, New York, p.p.76-83.
3. Collins M, Duffy N. (2001) *Convolution kernels for natural language*. Massachusetts: MIT Press, p.p.625-632.
4. Jeon J, Croft W, Lee J. (2005) Finding similar questions in large question and answer archives. *Proc. Conf. On Information and Knowledge Management*, ACM, New York, p.p.84-90.
5. Agichtein E, Castillo C, Donato, et al. (2008) Finding high-quality content in social media. *Proc. Conf. On Web search and web data mining*, p.p.183-194.
6. Agarwal A, Raghavan H, Subbian K, et al. (2012) Learning to Rank for Robust Question Answering. *Proc. Conf. On Information and Knowledge Management*, ACM, New York, p.p.833-842.
7. Blei D M, Lafferty A J D. (2007) A correlated topic model of science. *Annals of Applied Statistics*, 1(1), p.p.17-35.
8. Quan X J, Liu G, et al. (2010) Short text similarity based on probabilistic topics. *Knowledge Information System*, 25(3), p.p.473-491.
9. Sujian Li. (2002) The research of relevancy between sentences based on semantic computation. *Computer Engineering and Applications*, 38(7), p.p.75-76.



METAL
JOURNAL

www.metaljournal.com.ua