An intelligent Q&A system based on the LDA topic model for the teaching of Database Principles

Lin Cui & Caiyin Wang

Suzhou University
Suzhou, Anhui, People's Republic of China

ABSTRACT: With the development of computer networks, network education has received more and more attention. In the network environment, due to the limitation of time, teachers cannot answer in a timely manner all the questions that students ask. Therefore, an intelligent Q&A (questions and answers) system based on the LDA (latent Dirichlet allocation) topic model was developed, and is discussed in this article. In order to solve the difficult problems under network circumstances, considered in this research were the knowledge points and characteristics of FAQs (frequently asked questions) for the course, Database Principles. The intelligent Q&A system was based on the LDA model and on topics-documentation-knowledge points. This intelligent Q&A system allows users to describe problems in natural language and, then, the problems are submitted to the system. The system returns accurate answers related to the topic. Students opined that the Q&A system performed very well and could answer all the questions they posed about Database Principles.

INTRODUCTION

Network teaching frees people from the limitations of time and space since they can receive education whenever and wherever they like. But there are problems in the network teaching environment. For example, it is impossible to answer all the questions students ask of teachers due to time limitations and this saps students' enthusiasm. Furthermore, teachers have to answer repeated questions [1].

In order to solve these problems, the commonly used methods are BBS (bulletin board system), on-line FAQ (frequently asked questions), e-mail and message boards [2]. The BBS, e-mail and message boards do not produce an immediate response [3]. Traditional search engines, such as Google and Baidu have many defects in that the results they return are not the direct answer to the question [4].

By analysing the frequently asked questions in the course Database Principles, an intelligent Q&A system based on the LDA (latent Dirichlet allocation) topic model was developed. Through analysing the FAQs, it was determined that each question belonged to a specific type, and a method was proposed to determine a candidate question set, which greatly improves efficiency and accuracy.

RELATED WORK

As early as the 1960s, when research on artificial intelligence had just started, scholars proposed that computers should answer questions using natural language, which could be regarded as the rudimentary question answering system [5]. The Q&A systems had become fashionable for a time in the field of natural language in the 1980s [6]. However, with the rise of large-scale text processing technology, the research on question answering systems died away.

In recent years, with the rapid development of network and information technology, the desire by people to obtain information faster has again promoted the development of question answering systems. Many companies and research institutes are involved in this development, including Microsoft, IBM and MIT (Massachusetts Institute of Technology) [7]. In 1999, TREC (Text Retrieval Conference) introduced automatic question answering for tracking projects. Since then, the Q&A track has gradually become one of the most popular TREC projects [8].

Many countries have developed a number of relatively mature Q&A systems. In the open software Q&A systems, there is the Start system (http://start.csail.mit.edu/) developed by the InfoLab Group of the MIT Computer Science and Artificial Intelligence Laboratory and AnswerBus (http://www.answerbus.com) [9]. However, the Q&A systems used to solve problems for a specific course are very rare. Accordingly, an intelligent Q&A system that returns answers to users' questions on the course Database Principles was developed.

RESEARCH FOUNDATION

Knowledge Points

The Q&A system for a specific course aims to answer questions proposed by users. A user's question usually contains detailed information about the course, which is actually specific information or several related pieces of information or one aspect of a *knowledge point*. Knowledge points are the basic organisational unit and the basic transmission unit, and include words, sentences, concepts, definitions, theorems, formulas, rules and laws, etc [10]. In teaching, knowledge is generally organised in the form of knowledge points. In this article, a knowledge tree is organised according to the structure of the chapter, section and knowledge points, as shown in Figure 1.

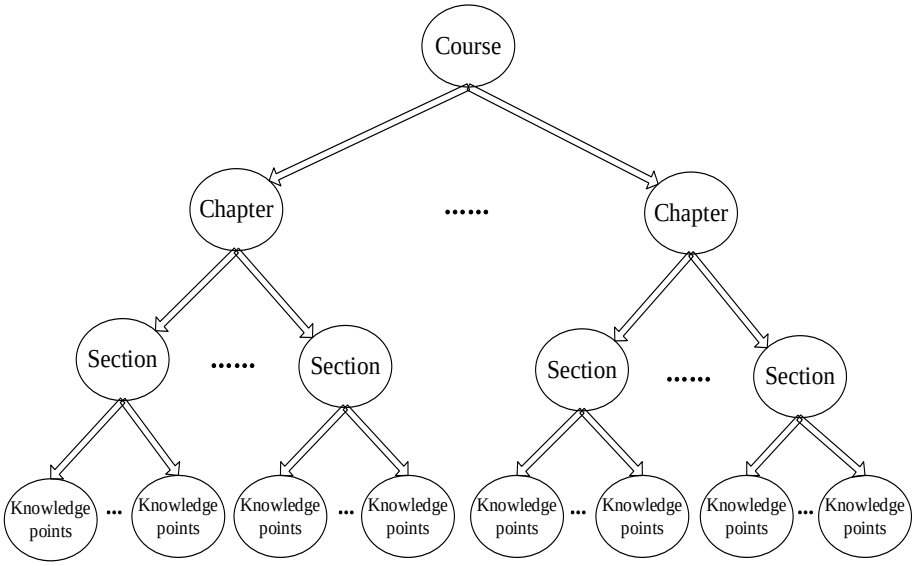

Figure 1. The tree structure of knowledge points.

LDA Model

The LDA model was proposed by Blei in 2003 [11]. The LDA model has been applied successfully to text classification, information retrieval and many other related fields [12]. The LDA model is shown in Figure 2.

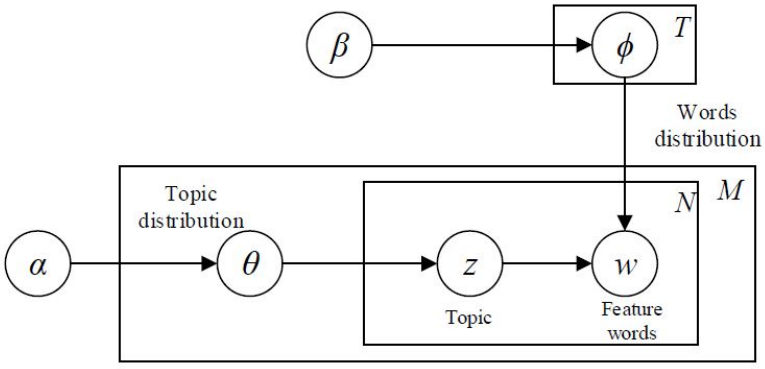

Figure 2: Representation of the LDA probability model.

The LDA model has a three-layer structure, i.e. words, topics and documents. Given a document collection, LDA would represent each document as a collection of topics, each topic is a multinomial distribution used for capturing the relevant information between words. In LDA, these topics are shared by all the documents and each document has a particular theme. LDA is determined by the parameter (α,β) of a document; α reflects the relative strength of hidden topics in document sets, β represents the probability distribution of all the latent topics. θ represents the proportion of each underlying theme in the document, z shows the underlying topic proportion in each word by document, w is the word vector table of documents. N is the total number of documents in the document set and N_d represents the total number of words in the documents [13].

OVERALL ARCHITECTURE OF THE QUESTION ANSWERING SYSTEM

The proposed intelligent Q&A system for the course Database Principles is illustrated in Figure 3.

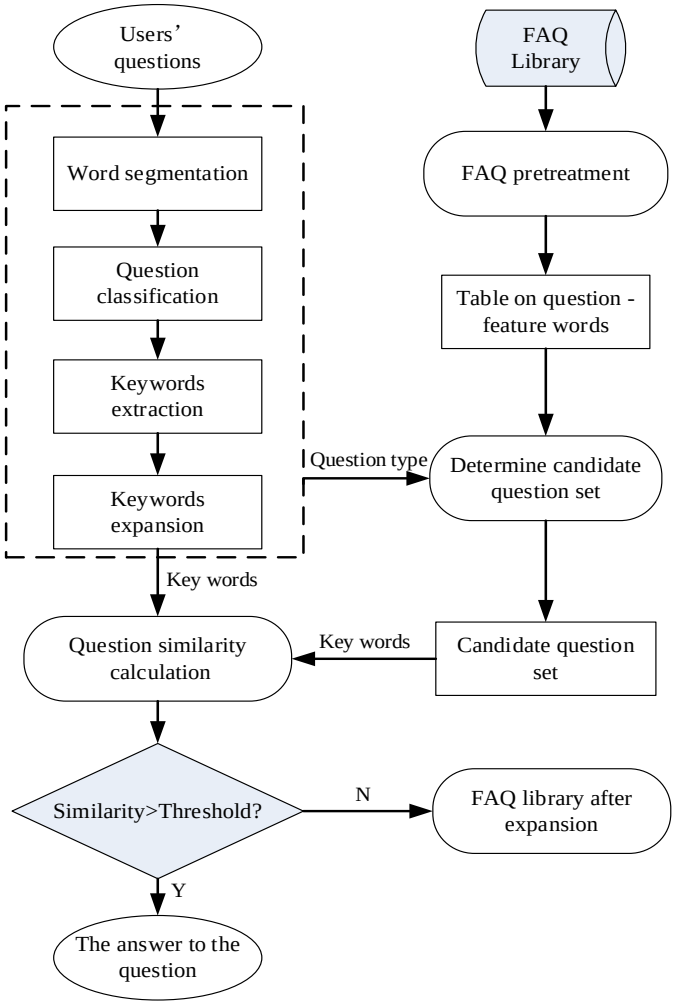

Figure 3: Overall framework for the system.

As shown in Figure 3, this system includes five major functional modules, which are: understanding the question, FAQ pretreatment, determining the candidate set of questions, question similarity calculation and FAQ library expansion. Every function module is introduced below.

The module of understanding the question includes word segmentation, question classification, keywords extraction and keywords expansion and other processes. Its main function is to analyse the input question.

The module for determining the candidate question set can narrow the search space according to problem type, in that the candidate question set is screened from the FAQ library. Thus, the operation of searching problems is facilitated.

The module of FAQ library preprocessing implements word processing for the frequently asked questions, so as to obtain the feature words contained in each question and avoid reprocessing every time in similarity calculation, which improves the efficiency of the system.

The module of question similarity calculation refers to the similarity calculation between the problems input by users and the problems in the candidate problems, which returns the most similar answers to users.

The module of FAQ expansion adds questions that the system cannot answer within the FAQ library. Similarity values are calculated by the module. The system determines whether the maximum similarity value for a user's question is already present in the FAQ library and to judge whether the similarity is larger than a preset threshold. If the maximum similarity value is less than the threshold, then, the problem input by the user does not exist in the FAQ library. This problem would be added to the FAQ library, which would further improve the library [14].

RESEARCH ABOUT CORE ISSUES

The Domain Knowledge Library

The domain knowledge library can be implemented by constructing a knowledge points table, a field characteristics table, a feature words similarity table and a question type table. In this article, the concepts of database principles are

described by using domain feature vocabulary. The relationships between concepts are represented through feature words similarity tables. The domain feature table stored the special vocabulary of database principles, which is used to describe the course knowledge. The domain feature table is the data base of word segmentation processing and is also the data foundation of problems' similarity calculation. Also, defined was the structure of question types, which are divided into six types, viz. the definition type, reason type, list type, relation type, method type and other type. The table of question types is shown in Table 1:

Table 1: Question type table.

Typeid	Type
1	Definition type
2	Reason type
3	List type
4	Relation type
5	Method type
6	Other type

Question Similarity Calculation

Whether the question similarity calculation is accurate or not directly determines the performance of the question answering system. The LDA model regards question answering systems as containing multiple topics and these topics have different weightings. The topic retrieval model includes implicit topic information. The specific calculation is as follows [15]:

$$P(C|B) = \lambda \prod_{w \in C} p(w|B) + (1-\lambda) \prod_{w \in C} \sum_{t \in t_B} \prod_{w \in t} p(w|t) * p(t|B) \quad (1)$$

Where, t_B is the topic set of question B, t is one of the topics in t_B, λ is a parameter. $p(w|t)$ is the probability that word w emerges in topic t, $p(t|B)$ is the probability that the theme t emerges in question B. All answers in the Q&A system are needed to calculate $P(C|B)$ and construct a topic retrieval model. If the probability $P(C|B)$ is less than a threshold, the answer is implicit, otherwise the answer is explicit.

SYSTEM IMPLEMENTATION

The intelligent Q&A system, described in this article, was based on the LDA model and developed using Java, JSP and Eclipse, MySQL for the database and Tomcat on the server. The interface for this system is easy to use, but the background program is quite complex. However, users do not need to know about the background program in detail. So, the system is easy to use. The operation of this system is illustrated in the following Figures 4-7.

To the definition type question *What is transaction?* (Figure 4), the system responds as shown in Figure 5.

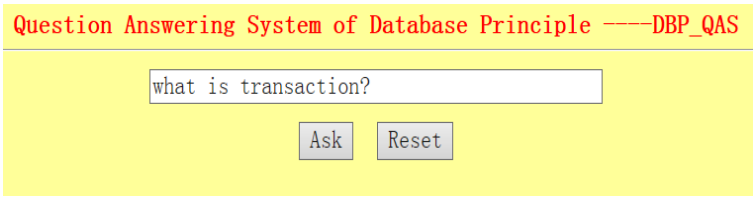

Figure 4: Definition question.

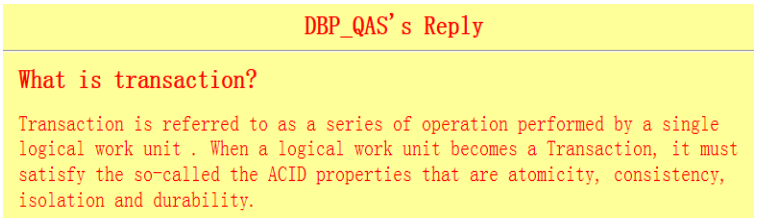

Figure 5: Reply to the definition question.

To the relation type *What is the difference between stored procedure and trigger?* (Figure 6) is the answer as shown in Figure 7:

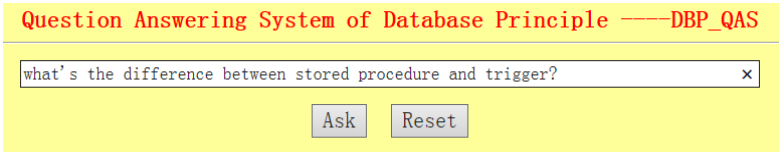

Figure 6: Relation question.

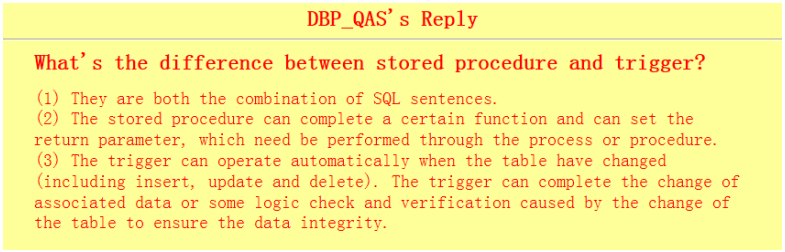

Figure 7: Reply to the relation question.

CONCLUSIONS

With Database Principles as the study object and by using the LDA model, an intelligent Q&A system was designed and implemented. This intelligent Q&A system can reduce the workload of teachers and better meet the personalised needs of students. In the network environment, students can ask questions easily and answers can be rapidly produced, improving the students' learning and the quality of the teaching.

ACKNOWLEDGEMENTS

This work was supported by the ordinary project of Anhui Province Colleges and Universities Natural Science Foundation of China (No.KJ2013B283, No.KJ2012Z401); the open project of Intelligent Information Processing Laboratory at Suzhou University of China (No.2013YKF14); and the project of Anhui Province Higher Education Revitalisation Plan of China (No.2013zytz074).

REFERENCES

1. Guzdial, M., Education: teaching computing to everyone. *J. of Communications of the ACM*, 52, **5**, 31-33 (2009).
2. Lugmayr, A., Applying *design thinking* as a method for teaching in media education. *Proc. 15th Inter. Academic MindTrek Conf.: Envisioning Future Media Environments*, 332-334 (2011).
3. Hardy, N., Pinto, M. and Wei, H., The impact of collaborative technology in IT and computer science education: harnessing the power of Web 2.0. *Proc. 9th ACM SIGITE Conf. on Infor. Technol. Educ.*, Cincinnati, 63-64 (2008).
4. Eastman, C.M. and Jansen, B.J., Coverage, relevance, and ranking: the impact of query operators on Web search engine results. *J. of ACM Trans. on Infor. Systems*, 21, **4**, 383-411 (2003).
5. Hendrix, G.G., Sacerdoti, E.D., Sagalowicz, D. and Slocum, J., Developing a natural language interface to complex data. *ACM Trans. Database Systems*, 3, **2**, 105-147 (1978).
6. Lin, J.J. and Katz, B., Question answering from the Web using knowledge annotation and knowledge mining techniques. *Proc. 12th Inter. Conf. on Infor. and Knowledge Manage.*, 116-123 (2003).
7. Jeongwoo, K., Eric, N. and Luo, S., A probabilistic graphical model for joint answer ranking in question answering. *Proc. 30th Annual Inter. ACM SIGIR Conf.*, 343-350 (2007).
8. Jijkoun, V. and Rijke, M.D., Retrieving answers from frequently asked questions pages on the Web. *Proc. ACM Conf. on Infor. and Knowledge Manage.*, 76-83 (2005).
9. Agichtein, E., Lawrence, S. and Gravano, L., Learning to find answers to questions on the Web. *ACM Trans. on Internet Technol.*, 4, **2**, 129-162 (2004).
10. Hijikata, Y., Takenaka, T. and Kusumura, Y., Interactive knowledge externalization and combination for SECI model. *Proc. 4th Inter. Conf. on Knowledge Capture*, 151-158 (2007).
11. Blei, D.M., Ng, A.Y. and Jordan, M.I., Latent Dirichlet allocation. *J. of Machine Learning Research*, **3**, 993-1022 (2003).
12. Wang, D., Thint, M. and Al-Rubaie, A., Semi-supervised latent Dirichlet allocation and its application for document classification. *Proc. IEEE/WIC/ACM Inter. Joint Conferences on Web Intelligence and Intelligent Agent Technol.*, 306-310 (2012).
13. Bhardwaj, A., Reddy, M. and Setlur, S., Latent Dirichlet allocation based writer identification in offline handwriting. *Proc. 9th IAPR Inter. Workshop on Document Analysis Systems*, 357-362 (2010).
14. Jijkoun, V. and Rijke, M., Retrieving answers from frequently asked questions pages on the Web. *Proc. 14th ACM Inter. Conf. on Infor. and Knowledge Manage.*, 76-83 (2005).
15. Krestel, R., Fankhauser, P. and Nejdl, W., Latent Dirichlet allocation for tag recommendation. *Proc. 3rd ACM Conf. on Recommender Systems*, 61-68 (2009).