2012 International Conference on Information and Computer Applications (ICICA 2012)
IPCSIT vol. 24 (2012) © (2012) IACSIT Press, Singapore

Learning Diagnosis for Students' Free-text Answers with Feature-based Approaches

Wen-Juan Hou, Chieh-Shiang Lu, Chuang-Ping Chang and Hsiao-Yuan Chen

Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan

Abstract. For improving the interaction between students and teachers so as to enhance the student's willingness to learn, it is important for teachers to understand students' learning levels and misconceptions, and then make the teaching materials better. In this paper, we want to diagnose students' free-text answers and indicate the inadequate concepts automatically. We first build the assessment corpus from the course in the university, and then extract some features. Finally, the students' answers are analyzed based on extracted features so that the missing core concepts of the students will be produced. We get the best of 77.41% recall rate based on Term Frequency-Inverse Document Frequency (TF-IDF) and Term Frequency (TF) features. The experiments with feature-based approaches show the exhilarating results and some future directions are explored.

Keywords: Learning diagnosis, Core concept, Machine learning, Feature extraction, Text mining

1. Introduction

The developments of computer-assisted tutoring and testing systems were extensively studied due to the popularization of computers and information technologies [2, 3, 5, 8, 9, 12]. In the conventional testing systems, students are given scores or grades to represent their learning status. The information is inadequate for both students and teachers because it cannot explicitly identify the missing concepts of students. It implies that providing students the learning insufficient concepts after testing is an important research issue.

Recently, there are some works related to the learning diagnosis. Some researchers focus on the web-based learning environment [1, 10]. The method restricts students in the internet environment. Some other researchers construct concept map of learning [4, 11]. The concept map is used to represent relationships among learning concepts in order to diagnose the learning problems of students. The approach usually needs the support of experts and it is a complicated and time-consuming task. Different techniques such as the graph related approach [2], the concept-effect model [5, 12], the knowledge network [6, 7], and neuro-fuzzy knowledge processing [14] are proposed. They try to understand student's knowledge space by different techniques. The works show that the learning diagnosis of learners is worthy of investigation.

In this paper, we propose a novel method to diagnose the student's answers and indicate the inadequate concept automatically. Generally, learning diagnosis systems directly compare the students' response and standard answers. Unfortunately, this method is restricted to have to prepare standard answers in advance and it is difficult to ensure which parts the student lacks during the comparison. To overcome this problem, we employ the features that are used in the machine learning method to help the learning diagnosis task.

2. Overall Architecture

Figure 1 shows the overall architecture of our methods for the diagnosis to the students' answers. At first, we preprocess students' answers, including stopword removal, stemming and feature extraction. Then we separate data into five parts. Four parts are served as the training data and one part is taken as the testing data.

Each of the extracted features is considered as a representation of the answer. Finally, the suggestion for student's deficiency in learning is produced after applying the feature comparison procedure.

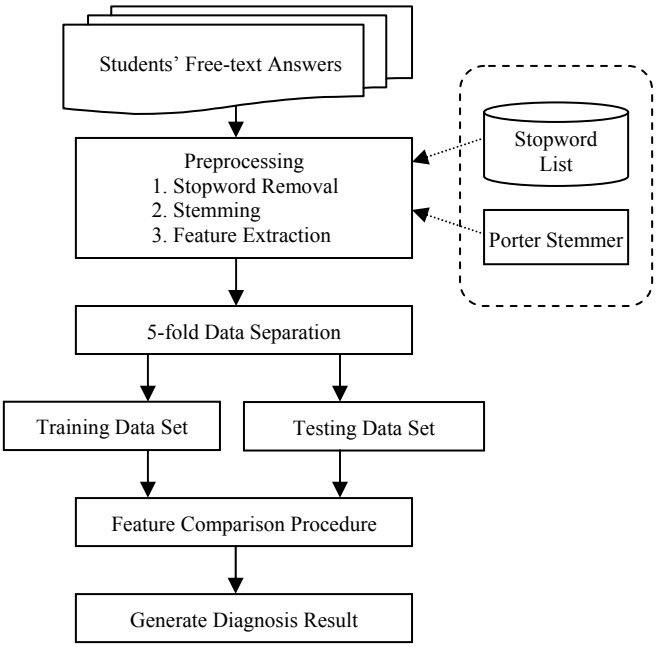

Fig.1. Architecture for the diagnosis of the student's learning status

3. Methods

3.1. Exclusion of Punctuation and Stopword

For gathering the useful information from the students' answers, the first processing step is to remove punctuation, numbers and stopwords because they serve as the noise roles.We also remove the words that both appear in student's answers and the question to avoid only copying the question by students.

3.2. Stemming

Stemming is a procedure of transforming an inflected (or sometime derived) word to its stem, base or root form. Usually, the stem need not be identical to the morphological root of the word. Generally, stemming can group the same word semantics and reflect more information around the word variations. The Porter's stemming algorithm [13] is used in the study.

3.3. Feature Extraction

The features extracted from answers are term frequency (TF), term frequency-inverse document frequency (TF-IDF) and entropy-variation (EV).

The term count in the given answer is simply the number of times a given term appearing in that answer. To give a measure of the importance of the term t_i within the particular answer a_j, we define the term frequency $tf_{i,j}$ as follows.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the term t_i in answer a_j, and the denominator is the sum of number of occurrences of all terms in answer a_j.

The TF-IDF weight is a statistical measure that is frequently used in the information retrieval and the data mining areas. The formula of computing the inverse document frequency idf_i for the term t_i is as follows.

$$idf_i = \log \frac{|D|}{|\{d : t_i \in D\}|} \quad (2)$$

where $|D|$ is the total number of documents in the corpus, and $|\{d: t_i \in D\}|$ is the number of documents where the term t_i appears (that is $n_{i,j} \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. We use $(1+|\{d: t_i \in D\}|)$ instead.

For the term t_i within the particular answer a_j, the formula of the TF-IDF weight is in the following:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

where $tf_{i,j}$ and idf_i are defined in Equations (1) and (2), respectively.

The entropy is the average uncertainty of a single random variable. The definition of the entropy $H(x)$ is expressed in terms of a discrete set of probabilities $p(x_i)$:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (4)$$

In the paper, we use the concept from entropy, and want to quantify the information amount of each word among students' answers. We call it as *Entropy-Variation* (EV). In our study, x_i is mapped to the word x in the ith student's answer, and $p(x_i)$ is the probability of word x in the ith student's answer. Because there are n students' answers, i is from 1 to n. Thus, $H(x)$ represents the average uncertainty of the word x appearance in the answers. To avoid the EV value tending to have large values as the number of students increases, we normalize $H(x)$ so that the value is between 0 and 1. The formula of the normalized $H(x)$, called $normH(x)$, is stated as below.

$$normH(x) = -\frac{1}{n} \sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (5)$$

3.4. 5-Fold Data Separation

In the study, we experiment with 5-fold cross validation. In 5-fold cross validation, the original sample is randomly partitioned into five subsamples. Of the five subsamples, a single subsample is retained as the validation data for testing, and the remaining four subsamples are used as training data. The cross-validation process is then repeated five times (the folds), with each of the five subsamples used exactly once as the validation data. The five results from the folds are then averaged to produce a single estimation.

3.5. Feature Comparison Procedure

In this phase, we want to find concepts that are important in the answers but missed by students. We first compare features TF, TF-IDF and EV between the training data and testing data. If the feature value is greater than zero in the training data but it is zero in the testing data, then we will extract this feature as the missing core concept candidate. After that we sort the missing core concept candidates according to the values in the training data. Finally, because our corpus contains the definitional questions where the answers are not long, we select at most ten missing core concepts with the higher feature values.

4. Experimental results

The evaluation metrics we adopt are the precision rate and the recall rate. The both rates are widely used to evaluate the systems in the NLP domain. The precision rate is the number of relevant retrieved records divided by the total number retrieved in a database search. Ther recall rate is the proportion of relevant retrived records to all relevant records in the database.

We borrow the concepts and make some modification for two reasons. First, the number of core concepts in a definition is not large, so we retreive first five or first ten missing concepts. Second, if the student gets higher scores in the test, it means missing concepts are less, so we have to include the number of the correct core concepts that the student owns during evaluation. Our evaluation metrics are called *core-precision* and *core-recall*. "*core-precision@5*" and "*core-precision@10*" represent five and ten concepts are proposed, respectively. The formulas are given in the following.

$$core - precision@5 = \frac{match + correct}{5 + correct} \quad (6)$$

$$core-precision@10=\frac{match+correct}{10+correct} \tag{7}$$

$$core-precision=\frac{core-precision@5+core-precision@10}{2} \tag{8}$$

$$core-recall=\frac{match+correct}{\#keyword} \tag{9}$$

where *match* is the number of concepts proposed by our system which are identical to core concepts that the teacher supplies in advance, *correct* is the number of core concepts that the student has in the answer, and *#keyword* is the total number of core concepts that the teacher gives.

We experiment with six definitional questions for each student and the final results are listed in Tables 1 and 2.

Tab.1 core-precision experimental results

Question / Feature	Q1	Q2	Q3	Q4	Q5	Q6	Average
TF	57.30%	39.28%	10.83%	16.52%	56.72%	60.50%	40.19%
TFIDF	55.71%	47.00%	10.83%	16.52%	56.21%	61.15%	41.24%
EV	56.24%	28.60%	4.67%	16.52%	54.47%	53.44%	35.66%
TF+TFIDF	55.71%	45.21%	11.21%	16.52%	56.21%	61.53%	41.06%
TF+EV	56.24%	29.62%	5.06%	16.52%	54.47%	53.07%	35.83%
TFIDF+EV	56.24%	28.60%	4.67%	16.52%	54.47%	53.07%	35.60%
ALL	56.24%	29.62%	4.67%	16.52%	54.47%	53.07%	35.77%

Tab.2 core-recall experimental results

Question / Feature	Q1	Q2	Q3	Q4	Q5	Q6	Average
TF	95.51%	65.90%	76.95%	60.28%	82.08%	75.03%	75.96%
TFIDF	93.59%	75.38%	74.38%	60.28%	81.65%	75.45%	76.79%
EV	95.19%	50.26%	30.79%	60.28%	79.94%	68.62%	64.18%
TF+TFIDF	93.59%	73.59%	79.49%	60.28%	81.65%	75.88%	77.41%
TF+EV	95.19%	51.28%	35.90%	60.28%	79.94%	68.40%	65.17%
TFIDF+EV	95.19%	50.26%	33.33%	60.28%	79.94%	68.40%	64.57%
ALL	95.19%	51.28%	33.33%	60.28%	79.94%	68.40%	64.74%

In the experiment, we first try individual features and then make combinations. For the individual feature, TFIDF performs the best both in *core-precision* and *core-recall*. TF+TFIDF reaches the best performance in *core-recall* among all combinations. It suggests TF and TFIDF are good indicators in the learning diagnosis system.

5. Conclusion

In this paper, we propose an approach to diagnosing the student's learning status by analyzing the student's free-text answers. We utilize features extracted from the texts, with different combinations of features and get the best core-recall rate of 77.41%. It means we suggest most of missing core concepts of the student. The core-precision rate needs enhancement furthermore. We analyze that perhaps core concepts in the answer are less but we propose too many, so that the precision rate is reduced.

In the future, we want to find more useful features to increase the performance. Moreover, linking the relevant learning material that core concepts need will help the student more. They will be further explored in the subsequent research.

6. Acknowledgements

The authors wish to thank the Aim for the Top University (ATU) project of National Taiwan Normal University (NTNU) funded by the Ministry of Education and the National Science Council (NSC) of Taiwan under contract NSC 100-2631-S-003-006.

7. References

[1] CHEN Chih-ming; HSIEH Ying-ling; HSU Shih-hsun. Mining Learner Profile Utilizing Association Rule for Web-based Learning Diagnosis, Expert Systems with Applications [J], 2007, 33: PP6-22.

[2] CHEN Ling-hsiu. Enhancement of Student Learning Performance Using Personalized Diagnosis and Remedial Learning System, Computers & Education [J], 2011, 56 (1): PP289-299.

[3] CHIOU Chuang-kai; HWANG Gwo-jen; TSENG Judy C. R. An Auto-scoring Mechanism for Evaluating Problem-Solving Ability in a Web-based Learning Environment, Computers and Education [J], 2009, 53 (2): PP261-272.

[4] CHOI Sook-young. A Concept Map_Based Adaptive Tutoring System Supporting Learning Diagnosis for Students with Learning Disability, In: K. Miesenberger, et al (eds.) Computers Helping People with Special Needs, 9th International Conference, ICCHP 2004, Paris, France, July 7-9, 2004, Proceedings, LNCS 3118, Springer-Verlag Berlin Heidelberg, PP194-201.

[5] CHU Hui-chun; HWANG Gwo-jen; HUANG Yueh-min. An Enhanced Learning Diagnosis Model based on Concept-Effect Relationships with Multiple Knowledge Levels, Innovations in Education & Teaching International [J], 2010, 47 (1): PP53-67.

[6] CHUANG Tung-lin; FANG Kwo-ting. Based on Knowledge Management Perspectives to Construct Knowledge Network Learning Diagnosis Analysis System, Proceedings of 2009 International Conference on Machine Learning and Computing, 107-111, Darwin, Australia, July 10-12, 2009.

[7] HEH Jia-sheng; LI Shao-chun; CHANG Alex; CHANG Maiga; LIU Tzu-chien. Diagnosis Mechanism and Feedback System to Accomplish the Full-Loop Learning Architecture, Journal of Educational Technology & Society [J], 2008, 11 (1): PP29-44.

[8] HOU Wen-juan; TSAO Jia-hao. Automatic Assessment of Students' Free-Text Answers with Different Levels, International Journal on Artificial Intelligence Tools [J], 2011, 20 (2): PP327-347.

[9] HOU Wen-juan; TSAO Jia-hao; CHEN Li; LI Sheng-yang. 2010. Automatic Assessment of Students' Free-text Answers with Support Vector Machines, In: N. García-Pedrajas, et al (eds.) Trends in Applied Intelligent Systems: 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, Córdoba, Spain, June 1-4, 2010, Proceedings, Part I, LNAI 6096, Springer-Verlag Berlin Heidelberg, PP235-243.

[10] HUANG Chenn-jung; LIU Ming-chou; CHU San-shine; CHENG Chih-lun. An Intelligent Learning Diagnosis System for Web-based Thematic Learning Platform. Computers & Education [J], 2007, 48 (4): PP658-679.

[11] LEE Chun-hsiung; LEE Gwo-guang; LEU Yungho. Application of Automatically Constructed Concept Map of Learning to Conceptual Diagnosis of E-learning, Expert Systems with Applications [J], 2009, 36 (2): PP1675-1684.

[12] PANJABUREE Patcharin; HWANG Gwo-jen; TRIAMPO Wannapong; SHIH Bo-ying. A Multi-expert Approach for Developing Testing and Diagnostic Systems based on the Concept-Effect Model, Computers & Education [J], 2010, 55: PP527-540.

[13] PORTER Martin F. An Algorithm for Suffix Stripping, Readings in Information Retrieval, eds. JONES Karen sparck and WILLET Peter, Morgan Kaufmann, San Francisco, 1997, PP313-316.

[14] STATHACOPOULOU Regina; MAGOULAS George D.; GRIGORIADOU Maria; SAMARAKOU Maria. Neuro-Fuzzy Knowledge Processing in Intelligent Learning Environments for Improved Student Diagnosis. Information Sciences [J]. 2005, 170 (2-4): PP273-307.