Journal of Ambient Intelligence and Humanized Computing (2020) 11:539–551
https://doi.org/10.1007/s12652-018-0934-9

ORIGINAL RESEARCH

Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining

Zengzhen Shao[1,2] · Yancong Li[2] · Xiao Wang[2] · Xuechen Zhao[1] · Yanhui Guo[1]

Received: 19 May 2018 / Accepted: 28 June 2018 / Published online: 11 July 2018

Abstract
As an important knowledge visualization tool, concept map has become a research hotspot in educational data mining. Traditional concept map generation algorithms are difficult to generate concept maps quickly because of their strong reliance on experts' experience. A hybrid TA-ARM algorithm for automatic generation of concept map based on text analysis and association rule mining is proposed. The TA-ARM algorithm fully considers the association rules between concepts, uses the text classification algorithm in text analysis technology instead of manually classify the questions into concepts, and combines the association rule mining method to generate concept maps. The experimental result shows that the TA-ARM algorithm can automatically and rapidly generate the concept map, which not only reduces the impact of outside experts, but can also dynamically adjusts the concept map based on the parameters such as the threshold of confidence between test questions. The concept map generated by the TA-ARM algorithm expresses the association rules between the concepts and the degree of closeness through the associated pairs and relevant degree, and can clearly show the structural associations between concepts. The contrast experiment shows that the quality of the concept map automatically generated by the TA-ARM has a high quality and can visualize the associations between concepts and provide optimization and guidance for knowledge visualization.

Keywords Concept map · Educational data mining · Automatic generation · Text analysis · Text classification · Association rules mining

1 Introduction

With the continuous advancement of education informatization and education modernization (Cao 2016), educational data mining (Heiner and Heffernan 2014) has received extensive attention from researchers at home and abroad. In order to promote the development of educational technology, a large number of technologies related to educational data mining are constantly being proposed (Slater et al. 2018). The concept map is an effective knowledge visualization tool and the educational data mining technology on the concept map has become the current research hotspot (Markham et al. 2010). The concept map was first proposed by Dr. Novak (Novak and Gowin 1984) of Cornell University in 1984 and the concept map proposed by Novak expresses the associations between concepts through edges, and describes the associations between concepts using a near-natured language. The form of the concept map in recent years is still based on the network-based conceptual structure proposed by Novak, using nodes to represent concepts, using directed edges to represent the connections between concepts and using prepositional labels to represent the dependencies between concepts associations (Paul 2012).

Scholars both at home and abroad have conducted extensive research on concept maps and applied the concept maps to different fields, such as teaching diagnosis (Hirashima et al. 2015), knowledge building (Zhang et al. 2007) and clinical nursing (Kaddoura et al. 2016), and achieved some results. However, the early generation of concept maps was mainly handmade by experts based on their experience, not only time-consuming, but also difficult to guarantee its accuracy (Coffey et al. 2002). In recent years, automatic generation algorithms of concept maps relying on educational data

✉ Zengzhen Shao
shaozengzhen@163.com

[1] School of Data Science and Computer Science, Shandong Women's University, Jinan 250002, China

[2] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China

mining techniques have been continuously proposed. Jiang et al.(Jiang et al. 2009) proposed a method to understand the structures of hand-drawn concept maps and a kind of structure-based intelligent manipulation technique. However, the concept map needs to be manually generated by experts first. Chen et al. (2006) used text analysis techniques to automatically generate conceptual maps of the e-learning domain from the literature. But they only considered the association rules between words and did not reflect the association rules between concepts. Caputo and Ebecken (2011) use the natural language processing method of text analysis technology to generate concept maps from e-commerce web pages. They used information extraction to mine concepts and analyze their associations, but they do not fully consider the role of dynamic data in the process of concept map generation. Huang et al. (2015) proposed an algorithm to automatically generate concept maps under simulated datasets. They calculated the correlation between concepts through improved Apriori (Toivonen 2011) algorithm, but the association rules between concepts and test questions was still classified by experts manually. Atapattu et al. (2017) extracted concept maps from lecture slides and the suitability of auto-generated concept maps as a pedagogical tool. However, they only considered the content of the slides, and did not pay enough attention to the test data of students. In summary, researchers have made great achievements in the automatic generation algorithms of concept maps in recent years. However, there are common shortcomings such as the excessive reliance of experts, long time required to generate concept maps, and the lack of rational use of dynamic data such as student's answer records.

In this paper, we propose a text analysis-association rules mining (TA-ARM) algorithm based on text analysis and association rules mining, which is based on the test data of students. The TA-ARM algorithm uses text classification of text analysis technology to classify test questions into concepts to reduce the time spent manually classifying questions into concepts. At the same time, it combines student's answer records and introduces association rules mining to automatically generate the concept map. The experimental results are shown that the TA-ARM algorithm can rapidly generate a high quality concept map under the premise of reducing the labor intensity.

The remainder of this paper is organized as follows. Related literature is reviewed in Sect. 2. The explanation of the TA-ARM are discussed in Sect. 3. Section 4 conducts the computational experiments and analysis. Finally, we conclude our results and point out the future research directions in Sect. 5.

2 Literature review

Data mining methods and models can uncover hidden nuggets of information on large data sets (Booth 2007), and has been successfully applied in medicine (Li et al. 2004), finance (Cowan 2002), biology (Hirschman et al. 2002) and other fields. However, data mining has started late in the education field, compared to other fields (Romero and Ventura 2007). In the field of education, data mining can find and solve various problems in education. Through a variety of methods and models, educational data mining can be used to design better and smarter learning technologies to better inform learners and educators (Baker 2014).

As a tool of knowledge organization in educational data mining, concept maps are generally used to represent knowledge structures (Acharya and Sinha 2017). The method of generating the concept map has been widely concerned by researchers (Novak and Cañas 2007). Tseng et al. (2007) proposed a two-phase concept map generation (TP-CMC) method to construct a concept map of a course through learner's historical test records. Their dataset comes from the 104 students of junior high school in Taiwan and the domain of the examination is the Physics course and the subordination between concept and question are given directly. Bai and Chen (2008) apply fuzzy rules and fuzzy reasoning techniques to automatically construct concept maps, they use simulated questions-concepts matrix and grade matrix to calculate the relevant degrees between concepts. The association rules between the concepts that have been mined have generated concept maps. The authors are more inclined to propose the theory of the algorithm. Under the condition that they can be quickly calculated, the feasibility of the algorithm is verified using small sample data. Chen and Sue (2013) improved on the basis of Bai and Chen (2008) algorithm, used existing data sets, combined association rule mining methods, and then generated concept maps. Their methods can dynamically generate concept maps based on students' answer records, which have certain practical significance. Oppl and Stary (2017) present a tabletop interface designed to assist in the generation of concept maps, and this tool plays an active role in the collaborative construction of concept maps. The generation of concept maps requires the participation of multiple concept map builders and need to use the experience of many people to improve the quality of the concept map. Acharya and Sinha (2015) used automatic hashing and pruning algorithms to automatically generate concept maps. The algorithm proposed by them is helpful to improve the efficiency of the concept map generation and has achieved certain results in practical application. Although the above-mentioned papers have a variety of algorithms and have achieved good results, their data are clear final datasets and are rarely involved in multi-semantic data such as textual data. However, it takes a lot of time to acquire these final datasets.

In addition there are many studies related to text analysis. Lai et al. (2017) proposed a new system based on information retrieval technology that automatically creates keyword concept maps for each part of the book. By analyzing the association rules of key words in the book, a static concept map is generated. In this process, no dynamic data is involved in the calculation. Santos (2018) used natural language processing and machine learning techniques to discover the associations between concepts from text documents and ultimately generate concept maps. They used text classification techniques in natural language processing to process text and grouped a set of 497 structured abstracts from Computer Science and Software Testing areas. The experiments have shown that the proposed method can assist researchers to generate concept maps. Qasim et al. (2013) presented a cluster-based approach to semi-automatically construct concept maps from unstructured text documents. They selected a total of 65 sample documents from 2007 to 2011 from the information system domain as the dataset for the generation of concept maps and used an unsupervised clustering algorithm to extract the structural associations of the candidate terms in the documents to generate concept maps. Nugumanova et al. (2015) used to analyze the frequency of document terms based on the collection of teaching materials, build terminology and document matrix to generate the concept map between terms. They summarize the information obtained from the document and the concept map generated by their proposed algorithm has the following advantages: quickness, effectiveness, completeness and actuality. There are also many methods to mine association rules from texts such as academic articles or teaching materials and then generate concept maps. These methods deal with unstructured texts, which can effectively save the time to obtain the final datasets, and have great significance for the generation of concept maps. This paper is inspired by the above-mentioned literature, through the text analysis to obtain the final datasets, replacing the time-consuming process of the experts, and combine with the classic association rules mining method, and ultimately achieve the automatic generation of the concept map.

3 TA-ARM concept map automatic generation algorithm

The TA-ARM algorithm consists of two phases: test questions text analysis and association rules mining between concepts. As shown in Fig. 1, in the test questions text analysis phase, extracting text features from test questions first, then build a classification model and use text classification to classify the test questions into concepts. In this phase, the associations between test questions and concepts can be obtained. In the association rules mining between concepts phase, generating frequent item sets of test questions combine with answering records first, and mapping the associations between test questions in the previous phase to the relevant degree of concepts, and finally generate the concept map. The concept map can be automatically generated combined with text analysis and association rules mining method without the aid of expert experience. In addition, we use the notations in Table 1 throughout the paper.

3.1 Test questions text analysis

Text analysis (Matsumoto et al. 2017) is an important branch of traditional data mining, but it is different from traditional data mining. Common text analysis methods include text clustering (Steinbach 2000) and text classification (Cohen 2004) and so on. In the test questions text analysis phase of TA-ARM algorithm, we use text classification techniques to automatically classify test questions into concepts to replace the manual classification relying on expert experience. The process of test questions text classification is shown in Fig. 2.

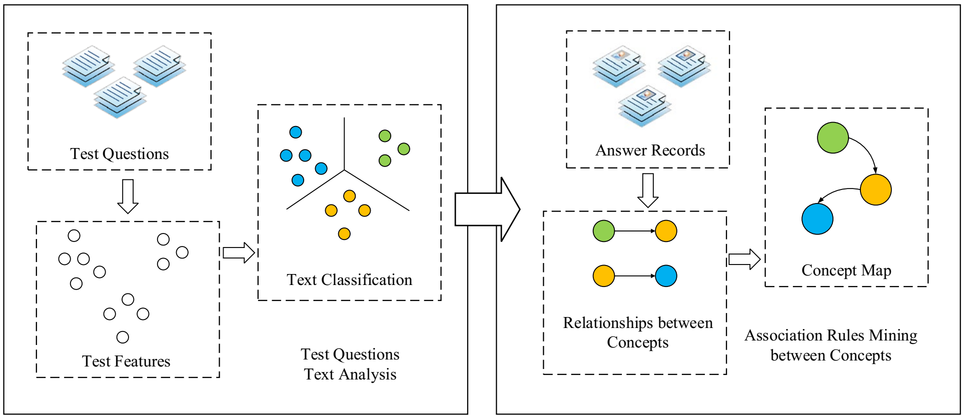

Test Questions
Test Features
Text Classification
Test Questions Text Analysis
Answer Records
Relationships between Concepts
Concept Map
Association Rules Mining between Concepts

Fig. 1 TA-ARM algorithm schematic diagram

Table 1 Notations

Symbol	Meaning	Illustration
Q	Test questions after word segmentation and stop words filtering	Q} $=(Q_1, Q_2, \dots, Q_j, \dots, Q_m)$
m	The number of test questions	
W	Text features extracted by TF-IDF	$W = (W_1, W_2, \cdots, W_j, \dots, W_m)$
W_j	The j-th text feature of test questions	$W_j = (W_{j1},\ W_{j2}, \dots, W_{ji}, \dots, W_{jr})$
r	Dimension of the text feature	
C	The concepts (that is, class labels)	$C = (C_1, C_2, \dots, C_x, \dots, C_k)$
k	The number of concepts	
QC	Questions-concepts matrix	The results classified by the k-NN classification model
G	Grade matrix	The student's answer records
n	The total number of students	
$\text{Count}(Q_a, Q_b)$	Answer records consistency	XNOR value between every two rows of grades in the grade matrix G
Min_{count}	Threshold of answer records consistency	$Min_{count} = n \times 40\%$
Conf $(Q_a \rightarrow Q_b)$	The confidence of the association rule $Q_a \rightarrow Q_b$	Conf $(Q_a \rightarrow Q_b) \in [0, 1]$
Min_{Conf}	The threshold of the confidence between test questions	
QC′	A new questions-concepts matrix constructed from QC	
Rev $(C_u, C_v)_{Q_a \rightarrow Q_b}$	The relevant degree of concepts	Rev $(C_u, C_v)_{Q_a \rightarrow Q_b} \in [0, 1]$
μ	The threshold of relevant degree of concepts	$\mu \in [0, 1]$
ε_{uv}	The total number of test questions included in concept C_u and concept C_v	

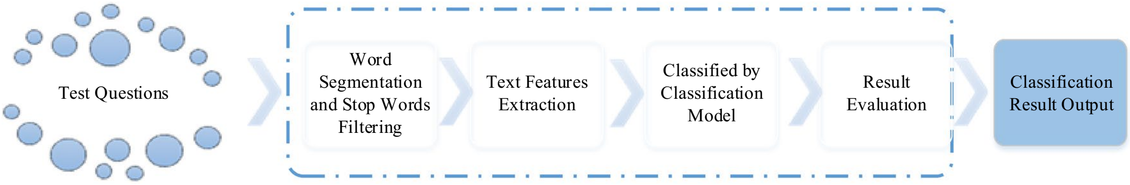

Test Questions
Word Segmentation and Stop Words Filtering
Text Features Extraction
Classified by Classification Model
Result Evaluation
Classification Result Output

Fig. 2 The process of test questions text classification

3.1.1 Word segmentation and stop words filtering

The texts of test questions are unstructured, which cannot be directly classified by the computer (Kurbatow 2015). Therefore, text preprocessing is required. The text of the Chinese test questions used in this paper is more complex than the English text, and there is no fixed interval between words, so the word segmentation is needed (Islam et al. 2008). Test questions after word segmentation contain many meaningless words, so we need to filter out meaningless words, that is, stop words filtering.

Test questions after word segmentation and stop words filtering are expressed as $Q = (Q_1, Q_2,\ \dots, Q_j, \dots, Q_m)$, where m is the number of test questions and Q_j is the j-th test question. In the next steps, the Q will be analyzed instead of the original test questions.

3.1.2 Text features extraction

Term frequency–inverse document frequency (TF-IDF) (El-Khair 2009) is a weighting function that depends on the term frequency in a given document calculated with its relative collection frequency. In this step, we choose the TF-IDF method to extract text features and transform Q into a vector space model (Melucci 2017) that can be understood by computers.

The text features extracted by TF-IDF are expressed as $W = W_1, W_2, \dots, W_j, \dots, W_m$ corresponding with Q, where W_j is the j-th text feature. Similarly, W_j can also be expressed as $W_j = (W_{j1},\ W_{j2}, \dots, W_{ji}, \dots, W_{jr})$, where W_{ji} is the weight of the feature item i is the j-th test question, and r is the dimension of the text feature. The formula for calculating the weight of a feature item W_{ji} is:

$$W_{ji} = TF_{j,i} \times IDF_i. \tag{1}$$

The $TF_{j,i}$ denotes the word frequency of text feature item i in the text of the j-th test question and IDF_i denotes the number of feature items i appearing in the whole texts, word frequency is proportional to weight, and reverse text frequency is inversely proportional to weight.

3.1.3 Classified by classification model

The Q are digitized into multidimensional vectors after text feature extraction. The text features W which extracted in the previous step can be processed by the classification model. There are many commonly used classification models, such as Rocchio (Moschitti 2003), Logistic Regression (Bertsimas and King 2017), Naive Bayes (Mccallum 1998), k-NN (Zhang et al. 2017) and SVM (Noble 2006) and so on.

Among them k-NN is a type of lazy learning model that does not require actual training (Larose 2004). Its time complexity is directly proportional to the number of test questions, and the k-NN algorithm is among the simplest of all machine learning algorithms, which is in accordance with the needs of this paper. The data trained or classified by the k-NN model are the text features W obtained from the previous step. Before the classification, the test questions text features W are divided into training samples W_{train} and samples to be classified W_{test}. The samples to be classified W_{test} are test questions that need to be manually classified into concepts by experts in traditional algorithms. Each question including W_{train} and W_{test} has a class label and the class label represents a concept. For convenience, the concepts (that is, class labels) are represented as $C = (C_1, C_2, \ldots C_x, \ldots, C_k)$, where k is the number of concepts and C_x is the x-th concept.

3.1.4 Result evaluation and classification result output

The training step of the k-NN model includes storing the text features W_{train} of the training samples and corresponding class labels. The k-NN model after training can be used to classify W_{test}. The main evaluation index of the classification model is accuracy. The higher the accuracy of the automatic classification by the k-NN model, the closer the results of the automatic classification by the k-NN model to the results classified by the experts manually.

For the convenience of the next phase of calculation, the results classified by the k-NN classification model are converted into a questions-concepts matrix QC, which is expressed as follows:

$$\mathrm{QC} = \begin{bmatrix} qc_{11} & qc_{12} & \cdots & qc_{1k} \\ qc_{21} & qc_{22} & \cdots & qc_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ qc_{m1} & qc_{m2} & \cdots & qc_{mk} \end{bmatrix},$$

where qc_{jx} indicates whether the current test question Q_j belongs to the concept C_x, $qc_{jx} \in \{0,\ 1\}$, and m is the total number of students. When $qc_{jx} = 1$ indicates that the test question Q_j belongs to the concept C_x, and $qc_{jx} = 0$ indicates that the test question Q_j does not belongs to the concept C_x. The matrix QC obtained by the k-NN model replaces the process of manually classifying test questions into concepts by experts and will be combined with association rules mining method in the next phase to automatically generate the concept map.

3.2 Association rules mining between concepts

The association rules were first proposed by Agrawal et al. (1993) in 1993 to discover meaningful associations hidden in the data. The associations discovered by the association rules mining method can be in the form of association rules or frequent item sets, expressed as A → B. A and B are disjoint item sets, where $A \cap B = \emptyset$, A is called the antecedent of the rule, and B is called the consequent of the rule. Apriori (Toivonen 2011) is one of the best algorithms for learning association rules. This paper cites the improved Apriori association rules mining algorithm proposed by Chen and Sue (2013). The algorithm proposed by Chen et al. relies on the experience of experts and combines association rule mining methods to generate concept maps, and they confirm the correctness of the generated concept map. We used their algorithms in the second phase (Association rules mining between concepts) of TA-ARM and compared the concept map generated using only their algorithms to

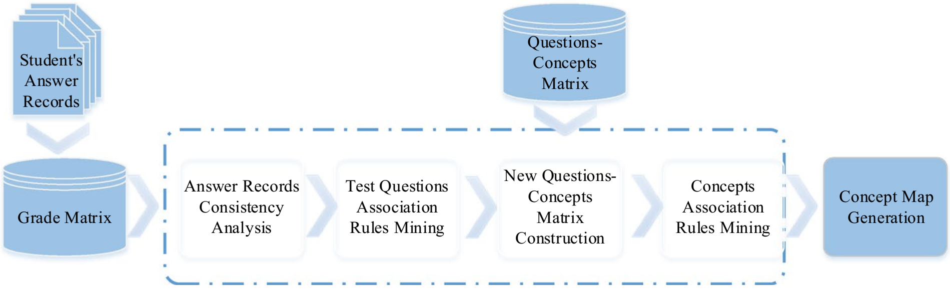

Student's Answer Records
Questions-Concepts Matrix
Grade Matrix
Answer Records Consistency Analysis
Test Questions Association Rules Mining
New Questions-Concepts Matrix Construction
Concepts Association Rules Mining
Concept Map Generation

Fig. 3 The process of association rules mining between concepts

verify the feasibility of the TA-ARM algorithm. As shown in Fig. 3, using student's answer records data to discover the association rules between test questions, and combine the questions-concepts matrix QC generated by the text analysis phase to map the associations between test questions to the associations between concepts, and automatically generate the concept map ultimately.

Before mining association rules, the student's answer records need to be digitized into a grade matrix G, expresses as follows:

$$\mathrm{G}=\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{21} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1} & g_{m1} & \cdots & g_{mn} \end{bmatrix},$$

where $g_{jy} \in \{0,1\}$, $g_{jy}=1$ means that student S_y correctly answered the question Q_j, and $g_{jy}=0$ means that student S_y erroneously answered the question Q_j, and n is the total number of students. In the next steps, we will perform association rule mining on performance matrix and problem matrix.

3.2.1 Answer records consistency analysis

In order to remove unnecessary associations between test questions, we introduce in answer records consistency. Answer records consistency is the XNOR value between every two rows of grades in the grade matrix G, which means the number of students who all answered correctly or all answered incorrectly in every two test questions. When the XNOR value is 0, it means that the student has answered one of the two test questions correctly. Otherwise, it indicates that the student has all answered correctly or all answered incorrectly in the two test questions. So we have

$$\text{Count}\ (Q_a, Q_b) = \sum_{l=1}^{n} (g_{al} \odot g_{bl}). \tag{2}$$

Using Count (Q_a, Q_b) to indicates the answer records consistency between test question Q_a and test question Q_b, and define the threshold as $Min_{count} = n \times 40\%$ for the minimum answer records consistency, where n is the number of students. The threshold Min_{count} is consistent with the threshold value set by the referenced algorithm. When $\text{Count}\big(Q_a, Q_b\big) < Min_{count}$, it indicates that the number of students who all answered correctly or all answered wrongly between the two questions is relatively rare. The association between the two questions is weak and will not be considered in the subsequent calculation.

Through the filtering of the answer records consistency with the threshold Min_{count}, the number of test questions to be calculated can be reduced, and the efficiency of generating concept maps can be improved.

3.2.2 Test questions association rules mining

Based on the grade matrix G and the Apriori (Toivonen 2011) algorithm, mining the association rules between test questions. Use the following formula to calculate the association rules between test questions:

$$\text{Conf}\ (Q_a \rightarrow Q_b) = \frac{\text{Sup}\ (Q_a, Q_b)}{\text{Sup}\ (Q_a)} \tag{3}$$

The Conf $(Q_a \rightarrow Q_b)$ in (3) denotes the confidence of the association rule $Q_a \rightarrow Q_b$, Sup (Q_a, Q_b) indicates the support of test questions Q_a and Q_b, and Sup (Q_a) denotes the support of test questions Q_a, where Conf $(Q_a \rightarrow Q_b) \in [0,1]$.

In this step, there are four kinds of association rules between questions to be considered: The student correctly answers test question Q_a and then also correctly answers test question Q_b. The student correctly answers test question Q_b and then also correctly answers test question Q_a. The student erroneously answers test question Q_a and then also erroneously answers test question Q_b. The student erroneously answers test question Q_b and then also erroneously answers test question Q_a. That is correct to correct and erroneous to erroneous. Then, summarize these four kinds of association rules into two types: correctly answered to correctly answered, and erroneously answered to erroneously answered. Calculate the confidence of two types of association rules respectively.

In order to remove unnecessary associations between test questions, the Min_{Conf} is set as the threshold of the confidence between test questions. The greater the confidence between two test questions, the closer the association between the two test questions and the greater the possibility of an association. Conversely, the weaker the confidence between the two test questions, the less likely they are associated.

3.2.3 New questions-concepts matrix construction

Construct a new questions-concepts matrix QC′ from the questions-concepts matrix QC obtained from the test questions text analysis phase using the following formula:

$$qc'_{jx} = \frac{qc_{jx}}{\sum_{h=1}^{m} qc_{hx}}, \tag{4}$$

where qc'_{jx} is the value in the new questions-concepts matrix QC′ whose position corresponds to qc_{jx} in the questions-concepts matrix QC obtained from the test questions text analysis phase, and $\sum_{h=1}^{m} qc_{hx}$ is the total degree of relevance of concept C_x in all test questions.

In the next step, the new questions-concepts matrix QC′ will be used instead of the original questions-concepts matrix QC to participate in the calculation.

3.2.4 Concepts association rules mining

The relevant degree of concepts expresses the association strength between two concepts and can be embodied as the corresponding association rule between two concepts. Combining the new question-concept matrix QC′, use the following formula to map the relevant degree of test questions to the relevant degree of concepts:

$$\mathrm{Rev}\,(C_u, C_v)_{Q_a \to Q_b} = qc_{au} \times qc_{bv} \times \mathrm{Conf}\,(Q_a \to Q_b). \tag{5}$$

The $\mathrm{Rev}\,(C_u, C_v)_{Q_a \to Q_b}$ denotes the relevant degree of test questions C_u and C_v derived from the association rule $Q_a \to Q_b$, where $\mathrm{Rev}\,(C_u, C_v)_{Q_a \to Q_b} \in [0, 1]$.

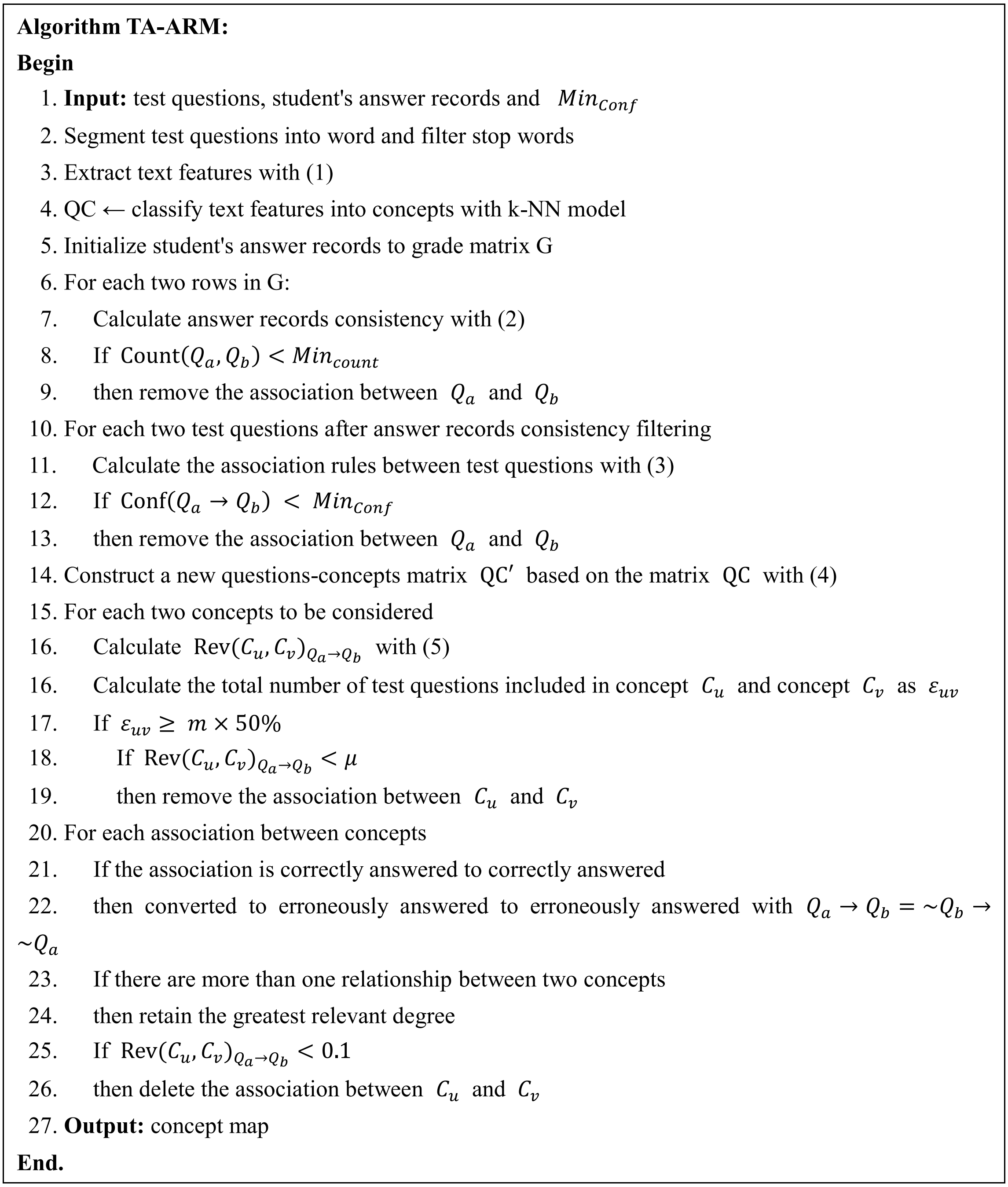
Define the threshold of relevant degree $\mu = \underset{1 \le j \le m, 1 \le x \le k \text{ and } qc_{jx} > 0}{\mathrm{Min}} qc_{jx}$, where $\mu \in [0, 1]$ and express the total number of test questions included in concept C_u and concept C_v as ε_{uv}. If $\varepsilon_{uv} < m \times 50\%$, then preserve $\mathrm{Rev}\,(C_u, C_v)_{Q_a \to Q_b}$, even though it is smaller than μ. The ε_{uv} is consistent with the value set by the referenced algorithm.

3.2.5 Concept map generation

Two types of association rules were considered in the previous steps: correctly answered to correctly answered, and erroneously answered to erroneously answered. Assume that the relevant degree of the concepts association which preserved is $\mathrm{Rev}\,(C_u, C_v)_{Q_a \to Q_b}$, and the association rule $Q_a \to Q_b$ is the erroneously answered to erroneously answered, then set a directed arrow from C_u to C_v and the weight of the edge is the relevant degree between C_u and C_v. Similarly, if the relevant degree of the concepts association which preserved is $\mathrm{Rev}\,(C_u, C_v)_{Q_a \to Q_b}$, and the association rule $Q_a \to Q_b$ is the correctly answered to correctly answered, according to the logical equality formula $Q_a \to Q_b = \sim Q_b \to \sim Q_a$, then set a directed arrow from C_v to C_u and the weight of the edge is the relevant degree between C_v and C_u. The association rules between all the concepts are standardized as a concept that should be mastered prior another concept, and the strength of the association is the degree of relevance. In order to prevent the associations between concepts from being too complex or too weak, delete associations with the relevant degree less than 0.1.

According to the association rules of the concepts obtained from the above steps, generate the associations between concepts. If there are more than one relationship between two concepts, the association with the greatest relevant degree is retained. Visualize the association rules between concepts using automatic drawing tools and generate the concept map ultimately.

The pseudo-code of the TA-ARM algorithm is as follows:



```
Algorithm TA-ARM:
Begin
Input: test questions, student's answer records and $Min_{Conf}$
Segment test questions into word and filter stop words
Extract text features with (1)
QC ← classify text features into concepts with k-NN model
Initialize student's answer records to grade matrix G
For each two rows in G:
    Calculate answer records consistency with (2)
    If $\mathrm{Count}(Q_a, Q_b) < Min_{count}$
    then remove the association between $Q_a$ and $Q_b$
For each two test questions after answer records consistency filtering
    Calculate the association rules between test questions with (3)
    If $\mathrm{Conf}(Q_a \rightarrow Q_b) < Min_{Conf}$
    then remove the association between $Q_a$ and $Q_b$
Construct a new questions-concepts matrix $\mathrm{QC}'$ based on the matrix $\mathrm{QC}$ with (4)
For each two concepts to be considered
    Calculate $\mathrm{Rev}(C_u, C_v)_{Q_a \rightarrow Q_b}$ with (5)
    Calculate the total number of test questions included in concept $C_u$ and concept $C_v$ as $\varepsilon_{uv}$
    If $\varepsilon_{uv} \geq m \times 50\%$
        If $\mathrm{Rev}(C_u, C_v)_{Q_a \rightarrow Q_b} < \mu$
        then remove the association between $C_u$ and $C_v$
For each association between concepts
    If the association is correctly answered to correctly answered
    then converted to erroneously answered to erroneously answered with $Q_a \rightarrow Q_b = {\sim}Q_b \rightarrow {\sim}Q_a$
    If there are more than one relationship between two concepts
    then retain the greatest relevant degree
    If $\mathrm{Rev}(C_u, C_v)_{Q_a \rightarrow Q_b} < 0.1$
    then delete the association between $C_u$ and $C_v$
Output: concept map
End.
```

3.3 Algorithm complexity analysis

The TA-ARM concept map automatic generation algorithm includes two phases. The k-NN algorithm is mainly used in the test questions text analysis phase. The time complexity of the k-NN algorithm is O(*m*), where *m* is the number of test questions. The Apriori algorithm is used in the phase of association rules mining between concepts. In this paper, only the frequent 2-itemsets in the test questions and concepts are considered, so the time complexity is not more than O(n^2).

4 Experiment and result analysis

4.1 Data sources and experimental environment

In order to verify the feasibility and effectiveness of the TA-ARM algorithm, this paper selects 617,940 authentic answer records from 6866 students in a large-scale examination of Computer Culture Foundation as the experimental dataset, including 90 test questions involving 9 concepts. The dataset was collected from the specialized subject undergraduate entrance simulation exam of a province in China in

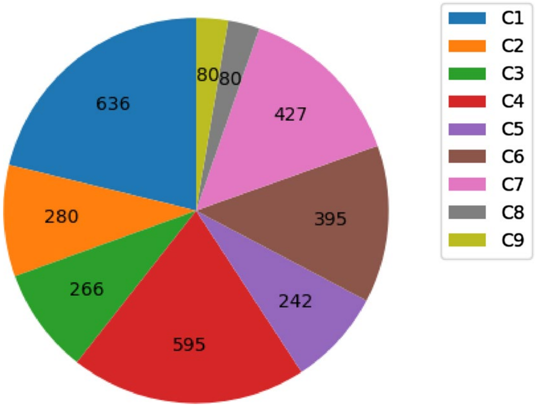


Fig. 4 The distribution of the test questions in the training samples

Table 2 The actual contents of concepts corresponding to concept labels

Concept label	Actual content
C1	Information technology and computer culture
C2	Windows7 operating system
C3	Word processing software (Word 2010)
C4	Spreadsheet system (Excel 2010)
C5	Presentation software (PowerPoint 2010)
C6	Database technology and Access 2010
C7	Computer network and web page making
C8	Digital multimedia technology
C9	Information security

December 2017. We also collected 3001 test questions with conceptual labels from many other college examinations as training samples. The distribution of the test questions in the training samples is shown in Fig. 4.

The experimental operating environment is the Windows 10 operating system. The programming language is Python 3.6, and the software development environment is PyCharm Community Edition 2018 and SQL Server 2008.

4.2 Experiment of test questions text analysis

There are three question types include test questions: multi-choice questions, true or false questions and cloze questions. Before the test questions text analysis phase, combine the options for multi-choice questions with the stem as an entire test question text, the stem of the judgment question is used as the test question text, and the correct answer of the cloze question is incorporated into the stem as the text of the entire test question. Each question has a concept label. The actual contents of concepts corresponding to concept labels are shown in Table 2. In the next steps, the questions we are referring to are all processed according to the types of questions, and all concepts are represented using concept labels.

Table 3 The classification report of the k-NN model

Concept label	Precision	Recall	F1-score	Support
C1	0.92	1.00	0.96	12
C2	1.00	0.91	0.95	11
C3	0.94	1.00	0.97	15
C4	0.93	0.93	0.93	14
C5	1.00	1.00	1.00	11
C6	1.00	0.90	0.95	10
C7	1.00	1.00	1.00	12
C8	1.00	1.00	1.00	3
C9	1.00	1.00	1.00	2
Average/total	0.97	0.97	0.97	90

There are no obvious separators in Chinese test questions, it is necessary to segment the test questions and filter the stop words. In the word segmentation step, select and use an open source tool Jieba and in the stop words filtering step, select and use the mainstream Chinese stop words list Harbin Institute of Technology stop words list.

Extracting text features W using formula (1), and each text feature W_j is a 4246-dimensional vector. Before the text classification, the test questions text features W are divided into training samples W_{train} and samples to be classified W_{test}. The number of training samples W_{train} is 3001, and the number of samples to be classified W_{test} is 90. Next, we will use the training samples W_{train} to train the k-NN model. That is, storing the feature vectors and concept labels of the training samples W_{train}.

After the k-NN classification model stores the W_{train}, then let k-NN model classifies the samples to be classified W_{test}. The model selects the 5 nearest neighbors when classifying, and this is also the default value set by the toolkit we are using. The results classified by the k-NN classification model are converted into a questions-concepts matrix QC. The abscissa represents the test questions and the ordinate represents the concepts, shown as follows:

$$
\mathrm{QC} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}.
$$

The classification report is shown in Table 3.

Table 4 Association rules between concepts obtained by the TA-ARM algorithm

Order	Association	Relevant degree
1	$C_4 \to C_9$	0.428
2	$C_4 \to C_8$	0.295
3	$C_7 \to C_8$	0.326
4	$C_3 \to C_8$	0.319
5	$C_1 \to C_8$	0.319
6	$C_1 \to C_9$	0.464
7	$C_2 \to C_8$	0.316
8	$C_2 \to C_9$	0.464
9	$C_3 \to C_9$	0.472
10	$C_5 \to C_8$	0.287
11	$C_5 \to C_9$	0.427
12	$C_6 \to C_8$	0.283
13	$C_6 \to C_9$	0.408
14	$C_8 \to C_9$	0.492
15	$C_7 \to C_9$	0.490
16	$\boldsymbol{C_2 \to C_6}$	**0.105**
17	$\boldsymbol{C_3 \to C_6}$	**0.106**
18	$\boldsymbol{C_7 \to C_6}$	**0.109**
19	$\boldsymbol{C_1 \to C_6}$	**0.105**

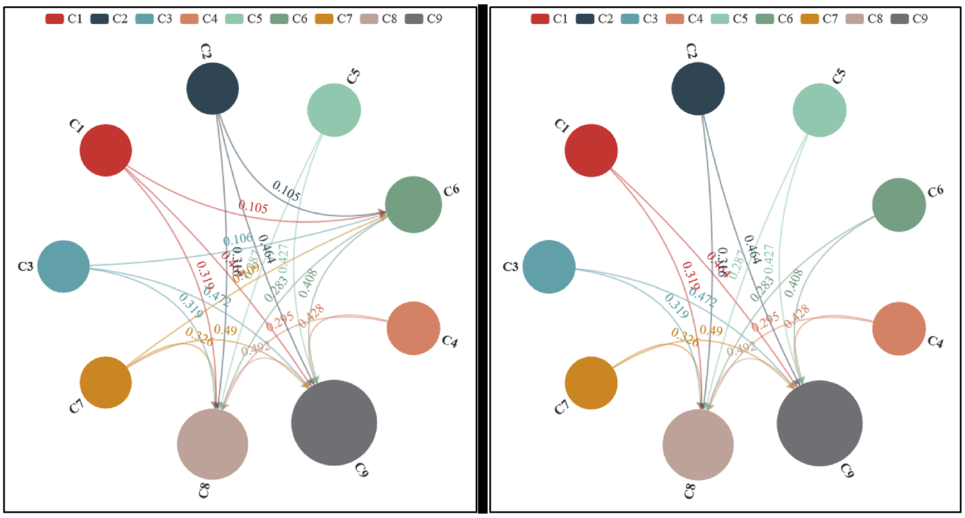
4.3 Association rules mining between concepts

Before mining association rules, the student's answer records are preprocessed as the grade matrix G, the abscissa indicates the test questions, and the ordinate indicates the students, shown as follows:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \cdots & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & \cdots & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & \cdots & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & \cdots & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & \cdots & 1 \end{bmatrix}$$

Based on the formula (2) and the grade matrix G, we can calculate the answer records consistency value Count (Q_a, Q_b) between every two questions Q_a and Q_b. Because the number of students is 6,866, we can calculate the threshold $Min_{count} = n \times 40\% = 2746.4$ and only consider association rules between test questions which meet Count $(Q_a, Q_b) \geq 2746.4$. After this step, we have retained 3660 associations between test questions.

Calculate the confidence of two types of association rules between test questions respectively and each association considers four kinds of association rules between every two questions after filtering by the threshold of answer records consistency. In order to facilitate comparison with the concept map generated by the expert manual assistance, we set the threshold of the confidence between test questions as $Min_{conf} = 0.75$, which consistent with the parameters of the algorithm proposed by Chen et al. Delete the association rules between test questions which meet Conf $(Q_a \to Q_b)$ < 0.75.After this step, we have retained 3758 association rules between test questions from correctly answered to correctly answered and 212 association rules



(a)

(b)

Fig. 5 Concept maps generated by the TA-ARM algorithm (**a**) and the algorithm proposed by Chen et al. (**b**)

Table 5 Association rules between concepts obtained by the algorithm proposed by Chen and Sue (2013)

Order	Association	Relevant degree
1	$C_4 \rightarrow C_9$	0.428
2	$C_4 \rightarrow C_8$	0.295
3	$C_7 \rightarrow C_8$	0.326
4	$C_3 \rightarrow C_8$	0.319
5	$C_1 \rightarrow C_8$	0.319
6	$C_1 \rightarrow C_9$	0.464
7	$C_2 \rightarrow C_8$	0.316
8	$C_2 \rightarrow C_9$	0.464
9	$C_3 \rightarrow C_9$	0.472
10	$C_5 \rightarrow C_8$	0.287
11	$C_5 \rightarrow C_9$	0.427
12	$C_6 \rightarrow C_8$	0.283
13	$C_6 \rightarrow C_9$	0.408
14	$C_8 \rightarrow C_9$	0.492
15	$C_7 \rightarrow C_9$	0.490

between test questions from erroneously answered to erroneously answered.

Construct a new questions-concepts matrix QC’ from the questions-concepts matrix QC obtained from the test questions text analysis phase using the following formula (4). In the next step, the new questions-concepts matrix QC′ will be used instead of the original questions-concepts matrix QC to participate in the calculation, shown as follows:

$$\mathrm{QC} = \begin{bmatrix} 0.077 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.077 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.077 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.083 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.091 & 0 & 0 & 0 & 0 \\ 0.077 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.083 & 0 & 0 \end{bmatrix}$$

Combining the new question-concept matrix QC′, use the formula (5) to map the relevant degree of test questions to the relevant degree of concepts. Calculate the threshold of relevant degree $\boldsymbol{\mu} = 1$ and calculate the total number of test questions included in concept C_u and concept C_v as ε_{uv}. The number of test questions is 90, delete the associations between concepts which meet $\varepsilon_{uv} \geq 90 \times 50\%$ and $\mathrm{Rev}(C_u, C_v)_{Q_a \rightarrow Q_b} < 1$. After this step, there are 3448 associations between concepts are retained.

If the associations between concepts are correctly answered to correctly answered, then convert the associations to erroneously answered to erroneously answered using the logical equality formula. And only keep associations between concepts with the highest relevant degree and greater than 0.1. There are 19 concepts of association rules that are retained after this step.

Visualize the association rules between concepts using automatic drawing tools and generate the concept map ultimately. Association rules between concepts obtained by the TA-ARM algorithm are shown in Table 4 and the concept map is shown in Fig. 5a.

In order to verify the feasibility and effectiveness of the TA-ARM algorithm, we use the same datasets, parameters and thresholds mentioned in the above experiment and compare the results obtained by the algorithm proposed by Chen and Sue (2013). In the algorithm proposed by Chen et al., the experts manually classify the questions into concepts firstly, and then combine association rule mining method to generate the concept map ultimately. The results obtained by the algorithm proposed by Chen et al. are shown in Table 5 and Fig. 5b.

After comparison, it can be found that the concept map generated by the TA-ARM algorithm has more than four associations compared to the concept map generated by the algorithm proposed by Chen et al. However, the relevant degree of these four associations is all less than 0.11 and relatively weak. And the other association directions and relevant degrees of the two concept maps are all the same. In this experiment, the TA-ARM algorithm required no more than 10 s to generate a concept map, which is far less than the time when the experts participate in the classification and then generate the concept map. The feasibility and effectiveness of the TA-ARM algorithm are verified.

5 Conclusions

Aiming at the limitations of the high reliance on experts and time-consuming in the current concept map generation algorithms, this paper proposes a new concept map automatic generation algorithm TA-ARM based on text analysis and association rules mining. The TA-ARM firstly uses the text classification method in text analysis technology to classify the test questions into concepts, which replaces the process of expert manual classification, and then combines the association rules mining method in current concept map generation algorithms to realize the automatic generation of concept maps.

The experiment shows that the TA-ARM algorithm has the following characteristics: (1) Low reliance on expert experience; (2) High quality concept maps and low time consuming; (3) The concept map can be dynamically adjusts the concept map based on the parameters such as the threshold of confidence between test questions. The concept map generated by TA-ARM algorithm shows directions and relevant degrees of associations between concepts. It shows the structure between concepts, and provides optimal guidance for teaching as a knowledge visualization tool.

Although the algorithm performs well in the automatic generation of concept maps, it also has some limitations: (1) Only consider the case where one test question belongs to only one concept, and did not consider the case where one test question belongs to multiple concepts; (2) When the number of classifications, that is, the number of concepts increases, the concept map generated by the TA-ARM algorithm may be significantly different from the concept map classified by the experts and then generated. In short, the TA-ARM algorithm is not suitable for the case of multiple classifications and multiple labels. In the future, we will conduct an in-depth research on these aspects.

Acknowledgements The authors are grateful to the anonymous reviewers for their constructive comments and invaluable contributions to enhance the presentation of this paper. This work was supported by National Natural Science Foundation of China (no. 71672154), Humanities and Social Sciences Foundation of the Ministry of Education of China (no. 16YJA630038), Joint Research Project between Southwest Jiaotong University and Cornell University (No. 268SWJTU15WTD01), China Postdoctoral Science Foundation (no. 2016M592697), Key Science and Technology Project of Shandong Province of China (no. 2014GGH201022), and Chang Jiang Scholars Program of China.

Appendix: Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://github.com/diligentlee/TA-ARM-data-sets.

References

Acharya A, Sinha D (2015) Construction of automated concept map of learning using hashing technique advances. Intell Syst Comput 327:567–578

Acharya A, Sinha D (2017) An educational data mining approach to concept map construction for web based learning. Inf Econ 21:41–58. https://doi.org/10.12948/issn14531305/21.4.2017.04

Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: SIGMOD Conference

Atapattu T, Falkner K, Falkner N (2017) A comprehensive text analysis of lecture slides to generate concept maps. Comput Educ 115:96–113

Bai SM, Chen SM (2008) Automatically constructing concept maps based on fuzzy rules for adapting learning systems. Expert Syst Appl 35:41–49

Baker RS (2014) Educational data mining: an advance for intelligent systems in education. IEEE Intell Syst 29:78–82

Bertsimas D, King A (2017) Logistic regression: from art to science. Stat Sci 32:367–384. https://doi.org/10.1214/16-STS602

Booth DE (2007) Data mining methods and models. Technometrics 49:500–500

Cao P (2016) On the transforming paths to future school—orientation and sustainable development of "internet + education". Educ Res 41:46–51

Caputo GM, Ebecken NFF (2011) Concept map construction applying natural language processing on text extracted from e-commerce web pages. In: Nature and biologically inspired computing, 2011 Third World Congress on, IEEE, pp 409–414

Chen SM, Sue PJ (2013) Constructing concept maps for adaptive learning systems based on data mining techniques. Expert Syst Appl 40:2746–2755

Chen NS, Kinshuk P, Wei CW, Chen HJ (2006) Mining e-learning domain concept map from academic articles. In: IEEE international conference on advanced learning technologies, pp 694–698

Coffey JW, Hoffman RR, Cañas AJ, Ford KM (2002) A concept map-based knowledge modeling approach to expert knowledge sharing. In: Boumedine M (Ed), proceedings of IKS 2002—the IASTED international

Cohen KB (2004) Natural language processing for online applications: text retrieval, extraction and categorization (review). Language 80:510–511

Cowan AM (2002) Data Mining in Finance: Advances in Relational and Hybrid Methods: Boris Kovalerchuk and Evgenii Vityaev (Eds.), Kluwer Academic Publishers, Norwell, Massachusetts, 2000, HB US $120, ISBN 0-7923-7804-0. Int J Forecast 18:155–156

El-Khair IA (2009) TF*IDF Encyclopedia of Database Systems 13:3085–3086

Heiner C, Heffernan N (2014) Educational data mining. Stud Comput Intell 1(1):467–474

Hirashima T, Yamasaki K, Fukuda H, Funaoi H (2015) Framework of kit-build concept map for automatic diagnosis and its preliminary use research & practice. Technol Enhanc Learn 10:17

Hirschman L, Park JC, Tsujii J, Wong L, Wu CH (2002) Accomplishments and challenges in literature data mining for biology. Bioinformatics 18:1553–1561

Huang X, Yang K, Lawrence VB (2015) An efficient data mining approach to concept map generation for adaptive learning. In: Industrial conference on data mining, pp 247–260

Islam A, Inkpen D, Kiringa I (2008) Applications of corpus-based semantic similarity and word segmentation to database schema matching. Vldb Journal 17:1293–1320

Jiang Y, Tian F, Wang X, Zhang X, Dai G, Wang H (2009) Structuring and manipulating hand-drawn concept maps. In: Universal communication symposium, pp 457–462

Kaddoura M, Vandyke O, Cheng B, Shea-Foisy K (2016) Impact of concept mapping on the development of clinical judgment skills in nursing students teaching & learning. Nursing 11:101–107

Kurbatow A (2015) The research of text preprocessing effect on text documents classification efficiency. In: "stability and control processes" in memory of V.i. Zubov, pp 653–655

Lai CF et al (2017) Using information retrieval to construct an intelligent E-book with. Keyword Concept Map 13:6637–6647

Larose DT (2004) k-nearest neighbor algorithm. Wiley, New York

Li L et al (2004) Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 32:71–83

Markham KM, Mintzes JJ, Jones MG (2010) The concept map as a research and evaluation tool: Further evidence of validity. J Res Sci Teach 31:91–101

Matsumoto T, Sunayama W, Hatanaka Y, Ogohara K (2017) Data analysis support by combining data mining and text mining. In: Iiai international congress on advanced applied informatics, pp 313–318

Mccallum A (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization, pp 41–48

Melucci M (2017) Vector-space model. In: Liu L, Özsu M (eds) Encyclopedia of database systems. Springer, New York, pp 1–6. https://doi.org/10.1007/978-1-4899-7993-3

Moschitti A (2003) A Study on optimal parameter tuning for rocchio text classifier lecture notes. in Computer Science 2633:420–435

Noble WS (2006) What is a support vector machine? Nat Biotechnol 24:1565–1567

Novak JD, Cañas AJ (2007) Theoretical origins of concept maps, how to construct them, and uses in education. Reflect Educ 3(1):29–42

Novak JD, Gowin DB (1984) Learning how to learn. Cambridge University Press, Cambridge

Nugumanova A et al (2015) Automatic generation of concept maps based on collection of teaching materials. In: International conference on data management technologies and applications, pp 248–254

Oppl S, Stary C (2017) Effects of a tabletop interface on the co-construction of concept maps. In: Human-computer interaction—INTERACT 2011—Ifip Tc 13 international conference, Lisbon, Portugal, September 5–9, 2011, proceedings, pp 443–460

Paul RS (2012) A review of "learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporation". Inf Soc 28:57–59

Qasim I, Jeong JW, Heu JU, Lee DH (2013) Concept map construction from text documents using affinity propagation. J Inf Sci 39:719–736

Romero C, Ventura S (2007) Educational data mining: a survey from 1995 to 2005. Expert Syst Appl 33:135–146

Santos V (2018) Concept maps construction using natural language processing to support studies selection. In: SAC 2018, Pau, France, 9–13 April 2018, ACM. https://doi.org/10.1145/3167132.3167301. ISBN 978-1-4503-5191-1/18/04

Slater S, Joksimovic S, Kovanovic V, Baker RS, Gasevic D (2018) Tools for educational data mining: a review. J Open Learn 42:85–106

Steinbach MA (2000) comparison of document clustering techniques. In: World text mining conference

Toivonen H (2011) Apriori algorithm. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1

Tseng SS, Sue PC, Su JM, Weng JF, Tsai WN (2007) A new approach for constructing the concept map. Comput Educ 49:691–707

Zhang HP, Zhou N, Chen YY (2007) Research on application of concept map in knowledge organization. Inf Sci 25(10):1570–1574

Zhang S, Li X, Zong M, Zhu X, Cheng D (2017) Learning k for kNN. Classif ACM Trans Intell Syst Technol 8:43

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.