Construction and Application of Subject Knowledge Graph for Basic Education

Tao Xu[1,2,3], Xiaqing Ma[1,2], Fengsi Wang[3,4], Chun Liu[2,4], Zixiang Zhang[1,2], Peiming Lu[2], Daojun Han[2, *]

[1] *Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, 475004, China*
[2] *School of Computer and Information Engineering, Henan University, Kaifeng, 475004, China*
[3] *Henan Technology Innovation Center of Spatio-Temporal Big Data, Henan University, Zhengzhou, 450046, China*
[4] *Henan Industrial Technology Academy of Spatio-Temporal Big Data, Henan University, Zhengzhou, 450046, China*

Abstract

In view of the problem that the distribution of knowledge points in the current knowledge system in the form of chapters can no longer meet the needs of students and teachers. This paper proposes to use SVM to identify knowledge points of tests, build a subject Knowledge Graph based on the identified knowledge points, and design and implement a subject Knowledge Graph system. This paper takes the mechanical movement of the first chapter of the eighth grade (upper) physics course as an example to conduct empirical research. By 10-fold cross-validation, the average F1 value of the algorithm used in this paper is 89.66%.

Keywords

Knowledge system, SVM, knowledge point information, subject Knowledge Graph, system

1. Introduction

In the existing teaching resources, knowledge points in the knowledge system are distributed in the form of chapter levels. However, in the face of fiercely competitive academic assessment, the difficulty of the assessment content of various tests is increasing day by day, focusing on the fusion of knowledge points. Therefore, the knowledge system distributed in the form of chapter levels can no longer meet the needs of students' learning and teachers' teaching. This paper argues that the Knowledge graph can show the complex correlation of knowledge points. At the same time, it also meets the needs of students and teachers better.

Knowledge Graph, as a form of structuring human knowledge, have attracted great attention from both academia and industry [1]. The Knowledge Graph describes the relationships between entities in real life in the form of triples. In view of the excellent form of Knowledge Graph, the research on subject Knowledge Graph has become a research hotspot of subject knowledge system in recent years. Chen et al. used TextRank to extract knowledge points from course introductions to construct a Knowledge Graph [2]. Cheng et al. used TF-IDF to retrieve teaching courseware and mine knowledge points to build a Knowledge Graph [3]. Su et al. evaluated the relationship between knowledge points by calculating the semantic similarity, PMI, and normalized Google distance between knowledge points [4]. Different from these studies, this paper believes that the relationship between knowledge points in the tests can better show the complex association of knowledge points.

Based on the above problems, this paper proposes to use the SVM [5] algorithm to identify the knowledge points of the tests, build the subject Knowledge Graph according to the identified knowledge point information, and design and implement the subject Knowledge Graph system based on test big

ISCIPT2022@7th International Conference on Computer and Information Processing Technology, August 5-7, 2022, Shenyang, China
EMAIL: txu@henu.edu.cn (Tao Xu); Maxq@henu.edu.cn (Xiaqing Ma); xxgcwfs@163.com (Fengsi Wang); liuchun@henu.edu.cn (Chun Liu); zzx1289@henu.edu.cn (Zixiang Zhang); 978726425@qq.com (Peiming Lu) * Corresponding author's email: hdj@henu.edu.cn (Daojun Han)

data. The first chapter of the physics course in grade 8 of junior middle school is empirically studied to realize the knowledge point identification, the construction of a subject Knowledge Graph and the query of knowledge points, and the construction of a subject Knowledge Graph system. The method of knowledge point recognition and the construction of a subject Knowledge Graph are evaluated.

2. Technical framework

2.1. Technology route

This paper aims to realize the identification of the knowledge points of the tests and the construction of the subject Knowledge Graph based on the teaching materials and tests, and to query the knowledge points according to the subject Knowledge Graph. The technology route is roughly divided into four steps. First, data acquisition is performed, including manual extraction of chapters, sections, and knowledge points information from the textbook and manual marking of knowledge points for the tests according to the extracted information. Then, this paper realizes the recognition algorithm of knowledge points, including preprocessing test points, extracting test points features by TF-IDF [6] method and realizing the recognition of test points by SVM classification algorithm. Next, using the chapters, sections, and knowledge points information extracted from the textbook and the tests constructs the subject Knowledge Graph. Finally, this paper uses the subject Knowledge Graph to query knowledge points.

2.2. Identification of knowledge points in tests

It takes three steps to realize the identification of the knowledge points in the tests. First, the tests data is preprocessed. Secondly, the TF-IDF method is used to extract the features of the test data. Finally, the SVM classification algorithm is used to identify the knowledge points of the tests.

2.2.1. Preprocessing of test data

In order to improve the accuracy of tests identification, it is necessary to preprocess the data. The preprocessing steps of tests data include tests data modeling and tests word frequency matrix construction.

(1) Tests data modeling

First, the pictures on the test are cleaned to obtain the tests data in plain text. When modeling the tests data, attributes such as "chapter", "test type", "knowledge point" of the tests are introduced to create the tests data model. That is, the tests data can be defined as $\mathbb{Q} =< q_i|0 < i \leq n >$, where $q_i = W, P, c, l, k$ represents the i-th tests, W represents the tests data after cleaning the picture, P represents the word segmentation result, c represents the chapter to which the tests belong, l represents the type of the tests, k represents the knowledge point of the tests, and n represents the number of tests.

(2) Constructing tests word frequency matrix

Jieba determines the association probability between Chinese characters through a Chinese thesaurus and forms phrases with high probability between Chinese characters to form word segmentation results. This paper uses Jieba to segment the test to get P_s. After completing the word segmentation, the word segmentation results are cleaned, and the irrelevant phrases P_a such as prepositions and adjectives are screened out, so that $q_i.P = P_s - P_a$. The complete set of phrases after word segmentation and screening of all tests data is denoted as $p_s = \{p_j|0 < j \leq m\}$, where m represents the number of phrases, and p_j represents the j-th phrase in the complete set of phrases. Then construct a word frequency matrix $\mathrm{M} =< f_{ij} >$ according to the complete set of phrases, and f_{ij} represents whether the j-th phrase $p_j$appears in the i-th tests.

2.2.2. TF-IDF method to extract tests features

Since the word frequency matrix obtained from the preprocessing results still has a large number of word segmentations, this paper uses the TF-IDF method to extract the features of the tests and optimize the word frequency matrix.

The TF-IDF method is often used to evaluate the importance of a word or phrase to a document set or one of the documents in a corpus. Therefore, this paper uses the TF-IDF method to extract the more important feature words from the tests corresponding to each knowledge point. TF-IDF consists of two parts, TF and IDF. TF refers to word frequency, which indicates the frequency of the word or phrase appearing in the knowledge point corresponding to the tests. The calculation of TF is shown in equation 1.

$$TF = \frac{t_c}{t_s}, \tag{1}$$

Among them, t_c represents the number of tests in which the word or phrase appears in the tests corresponding to the knowledge point, and t_s refers to the total number of tests corresponding to the knowledge point.

IDF refers to the inverse document frequency, the value of IDF is inversely proportional to the frequency of the word or phrase in the tests bank. The calculation of IDF is shown in equation 2.

$$IDF = log\left(\frac{T_s}{t_c+1}\right), \tag{2}$$

Among them, T_s represents the total number of tests in the tests bank. To avoid having a 0 in the denominator, the denominator in the equation needs to be added by 1.

Therefore, the calculation of TF-IDF is shown in equation 3.

$$TF - IDF = TF * IDF, \tag{3}$$

2.2.3. SVM classification algorithm

In this paper, the SVM algorithm is used to classify each knowledge point label into two categories so as to realize the identification of the knowledge points in the tests. The basic idea of the SVM algorithm is to find the hyperplane that can correctly divide the training set and has the largest interval. The steps to solve the SVM hyperplane are as follows:

① Define the training set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_M, y_M)\}$, where x is an n-dimensional feature vector, and the y value takes the form of 1 or -1. Then select the kernel function $H(x, z)$ and the penalty coefficient C, calculate the hyperplane when the interval is the largest, so that the closest point to the hyperplane is as far away from the hyperplane as possible. Next, use Lagrange to solve optimization problem, the result is shown in equation (4-6), and the solution obtained by the α vector is represented by α^*.

$$\min_{\alpha}\left\{\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\alpha_i\alpha_j y_i y_j H(x_i, x_j) - \sum_{i=1}^{M}\alpha_i\right\}, \tag{4}$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0, \tag{5}$$

$$0 \le \alpha_i \le C, \tag{6}$$

Among them, α is the Lagrange multiplier vector, and the C value represents the degree of penalty for misclassified points. The larger the C value, the greater the penalty for misclassified points.

② For α^* calculated by ①, select a positive component $0 < \alpha_j^* < C$ of α^* and calculate b^* as shown in equation 7.

$$b^* = y_i - \sum_{i=1}^{M}\alpha_i^* y_i H(x_i, x_j), \tag{7}$$

③ Finally, the classification decision function of SVM is obtained as shown in equation 8, and the classification hyperplane of SVM is shown in equation 9.

$$f(x) = sign\left(\sum_{i=1}^{M}\alpha_i^* y_i H(x_i, x_j) + b^*\right), \tag{8}$$

$$\sum_{i=1}^{M}\alpha_i^* y_i H(x_i, x_j) + b^* = 0, \tag{9}$$

2.3. Subject Knowledge Graph construction

The subject Knowledge Graph this paper constructed in this paper defines three concept: "chapter": Z, "section": S and "knowledge point": K, and defines the subject Knowledge Graph this paper constructed as $\mathbb{Z} =< h,r,t >=<< Z,P^1,S >,< S,P^2,K >,< K,P^3,K >>$, Where h represents the head entity, r represents the relationship, t represents the tail entity, P^1 represents the relationship between chapters and sections at the chapter level, P^2 represents the relationship between sections and knowledge points at the chapter level, and P^3 represents the relationship between knowledge points and knowledge points. In this paper, the entities "chapter", "section" and "knowledge point" are extracted from the textbook through artificial mode, and the relationship between entities "chapter" and "section", "section" and "knowledge point" is defined in the form of the chapter level of knowledge point in the textbook. The knowledge point information is identified from the tests according to this paper proposed test knowledge point identification algorithm: the co-occurrence of knowledge points and the inspection frequency of knowledge points. Then judge the dependence between knowledge points and the strength of knowledge points in the test paper. Finally, the chapters, sections, knowledge points, and their relationships are stored in the Neo4j graph database in the form of triples, and the thickness of the relationship indicates the strength of the relationship between knowledge points, and the size of nodes indicates the degree of knowledge points.

2.4. Knowledge point query

Conduct knowledge point inquiry, first connect the Neo4j database, and then judge whether the knowledge point e_1 exists in the database, if the knowledge point exists in the database, then query the relevant information when e_1 is the head entity and the tail entity, and build Knowledge Graph $G(e_1)$ and relation list $L(e_1)$ related to knowledge point, if the knowledge point does not exist in the database, the output "the entity has not been added to the database yet".

3. Application case studies

3.1. Data sources

This paper takes the mechanical movement of the first chapter of the eighth grade (upper) physics course as an example of how to conduct empirical research. Using the domestic authoritative primary and secondary education resource website-subject network (zxxk.com), the eighth grade (upper) physics course in the first chapter of the mechanical movement part of the sample tests, simulated 13740 tests.

3.2. Evaluation of knowledge point recognition technology of tests

In this paper, the SVM classification is performed on each type of knowledge point separately, so as to realize the multi-label classification of the knowledge points in the tests. First, preprocess the tests data, select the Chinese word segmentation component Jieba to segment the tests data, filter the irrelevant phrases such as adjectives and prepositions, and construct the tests word frequency matrix. Then, the TF-TDF method is used to extract the tests characteristics that can completely cover all the tests of the knowledge point. For the tests of different knowledge points, the word frequency matrix is optimized by the tests features of all knowledge points. The SVM model is trained by the optimized word frequency matrix to realize the classification of a single knowledge point in the tests and finally realize the identification of the knowledge points in the tests.

As shown in Figure 1, the SVM model was optimized by adjusting the proportion of the training set, and the model performance was evaluated by the F1 value of the test set.

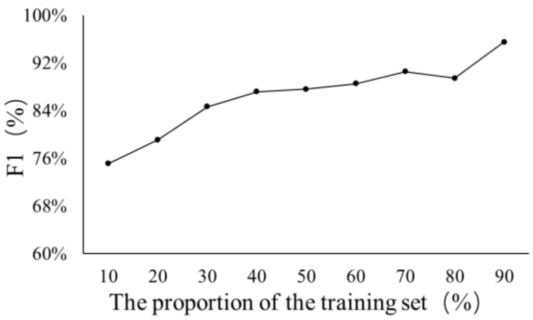

Figure 1: The change trend of F1 value with the proportion of training set

The results show that the SVM algorithm model can identify the knowledge points of the tests best when the training set ratio is 9:1, and its F1 value is 95.50%. In order to further prove the validity of the model, this paper verifies the validity of the model through ten-fold cross-validation, as shown in Figure 2, which represents the F1 value of the model in different test sets. The average value of its F1 value is 89.66%, which proves that the SVM algorithm this paper use can effectively identify the knowledge points in the test set.

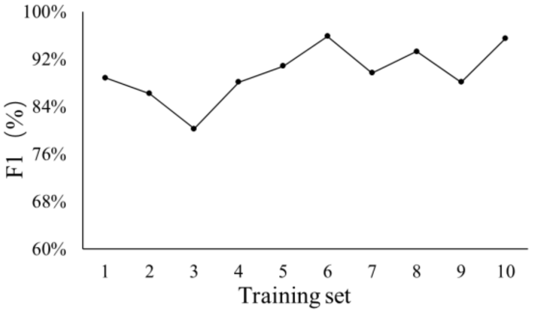

Figure 2: F1 Value of the model on different test sets

3.3. Subject Knowledge Graph construction evaluation

According to the chapter level of the knowledge points in the textbook and combined with the knowledge point information identified by the knowledge point recognition algorithm of the tests, the subject Knowledge Graph is constructed. The subject Knowledge Graph constructed in this paper contains 17 nodes: 1 "chapter" node, 4 "section" nodes, and 12 "knowledge point" nodes, and the relationships between them. As shown in Figure 3, the Knowledge Graph constructed in this paper not only retains the chapter level of knowledge points, but also includes the frequency of knowledge points in the test paper and the degree of relevance of knowledge points in the test paper.

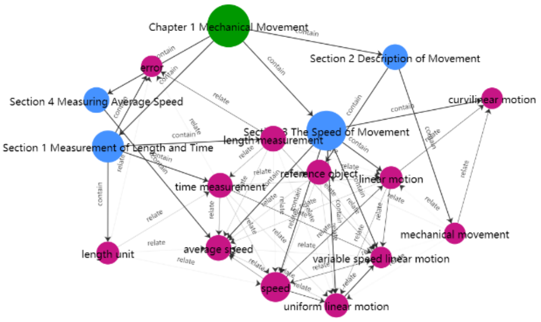
Chapter 1 Mechanical Movement
Section 2 Description of Movement
Section 4 Measuring Average Speed
Section 1 Measurement of Length and Time
Section 3 The Speed of Movement
error
curvilinear motion
length measurement
reference object
linear motion
time measurement
mechanical movement
length unit
average speed
variable speed linear motion
speed
uniform linear motion
contain
relate

Figure 3: Subject Knowledge Graph

3.4. Subject Knowledge Graph interface design and analysis

According to the method proposed in this paper, a disciplinary knowledge atlas system is constructed. The system includes three functional interfaces: test knowledge point recognition, Knowledge Graph display, and knowledge point query interface.

In the test knowledge point identification interface, enter the tests. The system will display the time of identification, the identified tests, and the identification result. It is convenient for users to learn correspondingly according to the recognition results.

In the Knowledge Graph display interface, user can clearly observe the distribution of knowledge points in chapters, the frequency of inspection in the test paper, and the dependencies between knowledge points in the tests.

In the knowledge point query interface, enter the knowledge point name. The system will display the relevant information about the knowledge point: the relationship diagram and the relationship table, which is convenient for learning about a certain knowledge point.

4. Summary and Outlook

In order to solve the problem that the knowledge points in the current knowledge system are distributed in the form of chapter levels. This paper uses the SVM algorithm to identify the knowledge points of the tests. By 10-fold cross-validation, the average F1 value of the algorithm used in this paper is 89.66%, which can effectively identify the knowledge points in the test set. The subject knowledge graph is constructed by identifying the knowledge points information and manually extracting knowledge points at the textbook chapter level. According to the method this paper proposed, this paper designed and implemented the subject Knowledge Graph system and realized three interfaces in the system: test knowledge point identification, Knowledge Graph display, and a knowledge point query interface. The next work will be carried out in two aspects: optimization of the method for identifying knowledge points in tests and adding functions such as question-and-answer search and test recommendation to the system.

5. Acknowledgements

This work was supported by the Educational Science Planning of Henan Province, China under Grant [2021YB0037]; the Open Fund of Scientific Research Laboratory of Henan University for undergraduates under Grant [No.71].

6. References

[1] Ji S, Pan S, Cambria E, et al. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[2] Xi C, Guang M, Jinjin Z, et al. A Method for Predicting Student Performance Combining Knowledge Graph and Collaborative Filtering[J]. Computer Application, 2020, 40(02):595-601.

[3] Ping C, Xun F. The Teaching Research of MPAcc Course Based on Knowledge Graph under the Background of "Golden Course" Construction——Taking the Course of "Cloud Accounting and Intelligent Financial Sharing" of Chongqing University of Technology as an Example[J]. Accounting Communications, 2019, (28):35-38.

[4] Yong S, Yong Z. Automatic Construction of Subject Knowledge Graph Based on Educational Big Data[C]//Proceedings of the 2020 The 3rd International Conference on Big Data and Education. 2020: 30-36.

[5] Cortes C, Vapnik V. Support-vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.

[6] Qaiser S, Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents[J]. International Journal of Computer Applications, 2018, 181(1): 25-29.