International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Semantic Similarity Analysis for Courses in MOOC

Mingxi Zhang, Wei He, Dini Xu, Yuqing Su

College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai, China

Abstract: ***With the rapid development of Internet technology and online education, the reform of traditional education methods has been promoted. Facing the huge online education resources, it is often difficult for learners to choose the appropriate courses. Quickly analyzing the connection between courses and knowledge points in massive data is important; it can promote the accurate dissemination of knowledge and prevent learners from cold start problems due to the lack of available effective information in the early stage of online education platform. In this paper, we propose a curriculum knowledge point similarity analysis system based on latent semantic analysis. Firstly, the introduction text and knowledge points of the course are used to build a "course-knowledge point" binary network. Then, the weight of the network is allocated based on TF-IDF model. Finally, the latent semantic analysis is carried out based on the weight matrix, the association score between the curriculum and knowledge points is calculated, and the top k relevant knowledge points of the curriculum are returned. Experiments on real-world public data sets show the effectiveness and accuracy of the system.***

Keywords: Latent Semantic Analysis, TF-IDF, Bipartite network, Similarity

1. Introduction

With the continuous deepening of educational informatization and the rapid development of the Internet, the reform of traditional education methods has been promoted. Online education has become a new important research and application direction formed by the integration of computers into the field of traditional education. In recent years, there have been a number of excellent online education platforms, through which students can choose courses of interest. For example, the massive open online course (MOOC) is a representative kind of online learning platform [1]. With the help of developed video technology and network technology, MOOC platform provides a large number of online courses for the public, which alleviates the problem of lack and uneven distribution of educational resources to a certain extent. However, the increase of platforms leads to "information overload", which makes it difficult for learners to choose appropriate courses. Therefore, how to quickly find learners' courses in a large number of courses and complete the high correlation between knowledge points and courses, so as to promote the accurate dissemination of knowledge points is particularly important.

The similarity analysis of course knowledge points is to complete the high correlation between knowledge points and courses, promote the accurate dissemination of knowledge points, and complete downstream tasks such as course recommendation. Quickly analyzing the connection between courses and knowledge points in massive data is important, it can solve prevent learners from cold start problems due to the lack of available effective information in the early stage of online education platform. It also has many applications in the field of educational recommendation, such as curriculum management [2], cognitive diagnosis [3], knowledge tracking [4], etc.

The existing course recommendation work can be roughly divided into content-based, collaborative filtering method, hybrid recommendation method and deep learning method [5]. The content-based method uses multiple attributes of the course to construct learners' preferences [6], mining learners with similar courses to share and discover predicted courses based on collaborative filtering method [7], recommending courses based on hybrid recommendation method combined with content and collaborative filtering, and considering the interaction between users and course information based on in-depth learning method [8], The prediction model is used to predict students' behavior.

The difficulty of knowledge point curriculum similarity analysis is to tap the potential relationship between curriculum information and knowledge points. The "course-knowledge point" binary network regards knowledge points and courses as nodes, and the relationship between knowledge points and course text information as relationships. However, it is difficult to capture the potential relationship between nodes in this way. In this regard, we introduce latent semantic analysis algorithm to solve this problem. On the basis of LSA algorithm, many scholars have carried out extensive application and research. For example, literature [9] applies LSA algorithm to recommendation system; Literature [10] applies LSA algorithm to news sensitive information tracking. Its advantages make it widely used in the fields of text classification and text retrieval [11].

In this paper, based on the knowledge points and the context information of the course, the "course knowledge points" binary network is constructed, the relationship between the knowledge points and the course context information is fully considered, the TF-IDF model [12] is used to calculate the correlation between the knowledge points and the course text, and the potential correlation between the course and the knowledge points is mined in combination with the latent semantic analysis algorithm, Analyze the given course text and return the results of top k relevant knowledge points, so as to design and implement a course knowledge point similarity analysis system based on latent semantic analysis. The flow of the system is shown in Figure 1. The main contributions of this paper are as follows:

1) Based on latent semantic analysis model, a course-concept similarity analysis system based on latent semantic analysis model is proposed. The algorithm can map knowledge points and courses to the same concept space, which greatly reduces the dimension of the concept space and makes the semantic relationship clearer than the original matrix.

2) Based on the relationship between course text and knowledge points, we use heterogeneous information to construct curriculum knowledge binary network, and use TF-IDF model to distribute the weight of the network.

3) We conducted experiments on a real data set and found that for a given course text, the system can return relevant knowledge points with high confidence, which proves the effectiveness and accuracy of our proposed system.

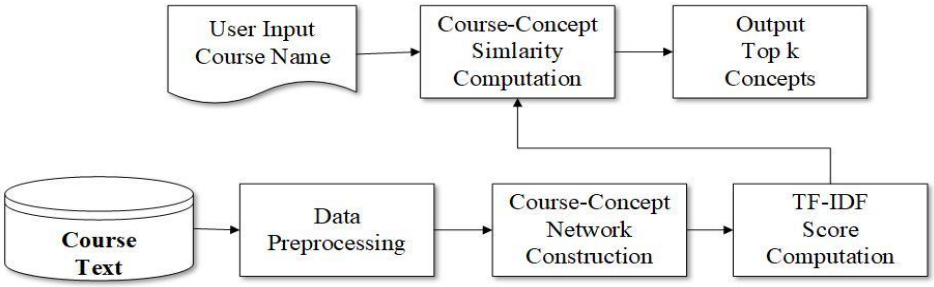

Figure 1: System framework

2. Related Work

In recent years, course recommendation with course analysis can be roughly divided into content-based, collaborative filtering, hybrid recommendation and deep learning.

In content-based recommendation, literature [13] describes an automatic personalized recommendation method, which aims to provide online automatic recommendation for active learners without explicit feedback from learners. It mainly represents the course by loading offline modules and online modules. Literature [14] uses multiple attributes to represent learners' preferences to reduce data sparsity and cold start problems and increase the diversity of ecological annotation lists.

In collaborative filtering, literature [15] proposes the framework of e-learning recommendation system, which encourages learners to cooperate with each other through peer learning and social learning, and uses the learning materials of excellent learners for representation. Reference [16] introduces a collaborative educational data mining tool based on association rule mining, which allows teachers with similar courses to share, discover and predict courses.

In the hybrid recommendation method, literature [17] uses the maximum likelihood to predict the learners' ability in the students' display feedback data, and determines the appropriate difficulty level for the course materials. Finally, the single parameter characteristic function is used to model and represent the course materials.

In the course recommendation method based on deep learning, literature [18] proposed a method of predicting students' final grades from the log data stored in the education system by using recurrent neural network (RNN). Literature [19] proposed a neural network model (bprn) of Bayesian personalized ranking network, which is used for curriculum representation.

At present, the disadvantage of traditional recommendation methods is that both content-based and collaborative filtering uses shallow models for prediction, which is difficult to effectively learn the deep-seated interactive information between users and courses. However, the deep learning model can mine the hidden models in the data, and the structure of the learning model is flexible.

3. Network Construction

3.1. Course-Knowledge network construction

The relationship between courses and knowledge points can be directly described as a "course knowledge point" network by the binary network, which is recorded as the network $G = (V, E)$, the node set $V = V_c \cup V_k$, represents the set of two types of nodes between courses and knowledge points, and E is the weighted relationship edge set between courses and knowledge points. The context information of knowledge points and course text is modeled as nodes in the graph, and the relationship between entities and entity text is a set of edges in the graph.

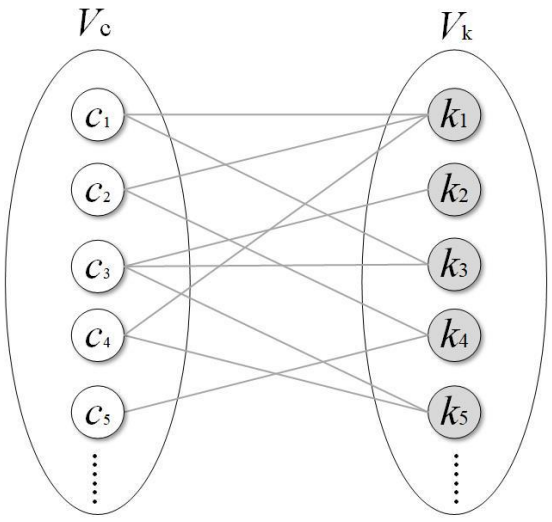

Figure 2: Course-Knowledge bipartite network

An example of the course knowledge point network is shown in Figure 2. Course C_1 is connected with knowledge points k_1 and k_3, but not with k_2, k_4 and k_5, indicating that knowledge points k_1 and k_3, are connected with course C_1. Course C_3 is also connected to documents k_1 and k_3, which indicates that course C_1 may have a potential connection with course C_3.

3.2. Weight distribution

TF-IDF model is an information retrieval model widely used

in practical applications such as search engines to evaluate the importance of a word to a document set or one of the documents in a corpus. Word frequency (TF) is the number of times words appear in the document. In order to achieve standardization, the common practice is to take the ratio between the frequency and the total number of words in the document. f_e Is the number of occurrences of knowledge point e in course d, and is the sum of the number of text words in course d. The TF value of knowledge point E in course d is calculated as follows:

$$TF_e = \frac{f_e}{M_e} \tag{1}$$

Inverse document frequency (IDF) indicates the importance of a given word. Its main idea is that if a feature item appears frequently in one text and low in other texts, it shows that this feature item has good category differentiation ability and should be given high weight. $|N|$ is the total number of courses in the data set, $|Q_e|$indicating the number of courses containing knowledge points e in the course data set. If a certain knowledge point appears frequently in one course and rarely in other courses, it is considered that this knowledge point has a good ability to distinguish categories. The IDF value and TF-IDF of knowledge point e are calculated as follows:

$$IDF_e = \lg \frac{|N|}{(|Q_e|+1)} \tag{2}$$

$$TFIDF_e = TF_e \times IDF_e \tag{3}$$

Next, based on the results of correlation score calculation between knowledge points and courses, a weight matrix p_{ed} as shown in Figure 3 is constructed, the two node types of knowledge points and courses are used as the rows and columns of the matrix and the values of the matrix $tfidf_{i,j}$, it represents the TF-IDF weight from each knowledge point e to all courses d.

$$P_{ed} = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{pmatrix} tfidf_{1,1} & 0 & tfidf_{1,3} & 0 & 0 \\ tfidf_{2,1} & 0 & 0 & tfidf_{2,4} & 0 \\ 0 & tfidf_{3,2} & 0 & tfidf_{3,4} & tfidf_{3,5} \\ tfidf_{4,1} & 0 & tfidf_{4,3} & 0 & 0 \\ 0 & 0 & 0 & tfidf_{5,4} & 0 \end{pmatrix} \end{array}$$

Figure 3: TF-IDF weight matrix

4. Similarity analysis of course and knowledge points

The key of latent semantic analysis technology is to project text information into low-dimensional space, and its schematic diagram is shown in Figure 4.

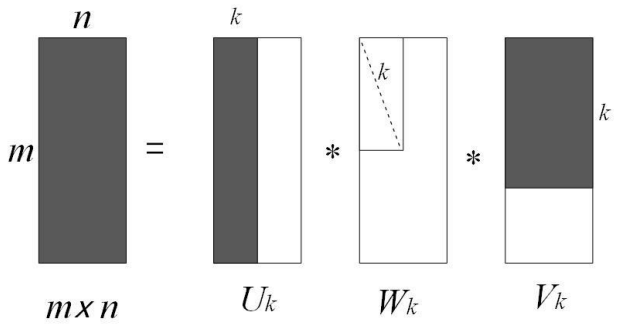

Figure 4: Schematic diagram of singular value decomposition

Assuming that the knowledge points in the course text are related, that is, there is a potential semantic structure, the singular value decomposition SVD (singular value decomposition) technology is used to reduce the dimension of the matrix. The knowledge point course matrix can be expressed as $P=(e_{i,j})_{\text{mxn}}$, and $e_{i,j}$ represents the weight of the knowledge point i in the course d_j, m is the number of knowledge points, n is the number of courses. Using SVD technology for the matrix, we can get:

$$P=UWV^T \tag{4}$$

Among them, U is the $m\times m$ order orthogonal matrix, V is the $n\times n$ order orthogonal matrix, W is the $m\times n$ order diagonal matrix, and the diagonal elements λi are the singular values of the matrix. The matrix is generally expressed approximately as follows:

$$P_k = U_k W_k V_k^T \tag{5}$$

Among them, U_k is the $m\times k$ order matrix composed of the front k column of the matrix U, W_k is the $k\times k$ order matrix composed of the front k row and front k column of the matrix W, and V_k is the $n\times k$ order matrix composed of the front column of the matrix V. The core of latent semantic analysis is the dimension reduction process. The determination of dimension k value is very key. The large value of k leads to the latent semantic space approaching the original vector space model, and the interference intensity of synonyms or polysemy is greater, and its processing performance is higher. The smaller k value will make the association model between text and feature words simpler in the latent semantic space. There is a lack of ability to recognize text and feature words, which may lead to the low accuracy of text mining and classification results. For the selection of dimension k value, the appropriate parameter value is manually determined through multiple experiments.

5. Experiments

5.1. Dataset

The dataset used in this paper is from Mocccube [20]; contains 706 courses, 38,181 videos, 114,563 knowledge points, 167,751 mapping information between courses and knowledge points, and 31,948 mapping information between videos and knowledge points. The text data of courses are obtained from the course introduction and the instructor's speech of the videos. Finally, in tag-knowledge bipartite network, we get 116,661 tags, 114,527 knowledge points, and 780,318 descriptive relationships between tags and knowledge points.

5.2. Experimental environment

The experiments run on Intel (R) Xeon (R) Bronze 3106 CPU@1.70GHz and 128GB RAM, under Windows 2012R. The development environment is IDEA 2020.2. We use Java 1.8 to implement the algorithms.

5.3. Result

In order to verify the effectiveness of our system, three randomly selected courses of the input information were used to demonstrate the effect, and both the input information and the output knowledge were translated into English.

Information of Course 1: The basic course of control engineering is a professional basic course for undergraduates in the College of machinery. Courses include: basic modeling of electromechanical system, analysis of transient response in time domain, and analysis of system stability in frequency domain. The teaching of the basic course of control engineering in Tsinghua University is mainly aimed at the analysis and correction of closed-loop control system, which provides students with a solid foundation and rich applications.

Information of Course 2: Linear algebra is one of the foundations of modern mathematics. It has extensive and profound applications in various fields of natural and social sciences such as physics, computer graphics, engineering and economics. At the same time, linear algebra is an important basic course for all majors of science and Engineering in Colleges and universities. As an important compulsory course for students majoring in non mathematical science and engineering of Tsinghua University, this course introduces the basic concepts and theories of linear algebra such as solving linear equations, matrix theory, vector space and linear transformation, and emphasizes the combination of theory and application of linear algebra.

Information of Course 3: Portuguese is the sixth largest language in the world. In Portugal, Brazil, Angola, Mozambique, Cape Verde, Guinea Bissau, Sao Tome and Principi, East Timor and Macao China and other four continents, Portuguese is the official language in the four continents of Asia, Africa, Europe and the United States. Brazilian Portuguese is slightly different from European Portuguese represented by Portugal in pronunciation, vocabulary and grammar. The Portuguese pronunciation of this course is mainly Brazilian Portuguese. Through learning this course, students can use Portuguese for daily conversation and have a solid basic knowledge of Portuguese grammar.

Information of Course 4: This course focuses on the history of human design, especially the major historical events in the past 250 years, to sort out the significant impact of design in the process of national development. In the current process of economic transformation and independent innovation in China, design plays a particularly important role in changing the image of made in China and truly becoming an innovative country.

Information of Course 5: This course is a way to understand the basic theories and methods of accounting, which helps to improve students' professional knowledge structure, enhance students' ability to understand the financial situation and operating results of enterprises, and expand students' knowledge. Through the teaching of this course, students can understand the basic accounting theories such as the definition, function, task, object and elements of accounting, and master the operation skills of setting accounting subjects and accounts, double entry bookkeeping, filling and reviewing accounting vouchers, registering account books, cost calculation, property inventory and preparing accounting statements.

Table 1 shows the top-5 results returned for the information of Courses 1-5. Course 1 is about electromechanical control system, all of the results are meet the requirement. Course 2 is about linear algebra, all of the results are also meet the requirement. Course 3 is about language of Portugal, all of the results are also meet the requirement. Course 4 is introducing the history of human design, the results: human design, history, are relevant, the remaining results like design are also related. Course 5 is Basic Accounting, Except for business English and probability theory, the rest of the results all meet the requirements. The analysis of several sets of experimental results verifies the effectiveness and accuracy of the system.

Table 1: Result of courses concept similarity analysis

Top	Course 1	Course 2	Course 3	Course 4	Course 5
1	electromechanical	modern mathematics	Portuguese	human design	Accounting
2	modeling	linear algebra	language	history	Finance
3	transient response	Engineering	translation	design	Business English
4	system stability	economics	vocabulary	nation	probability theory
5	closed-loop	matrix	grammar	Chinese	ACCA

6. Conclusion

In this paper, we propose a course knowledge point similarity analysis system based on latent semantic analysis. The system is divided into two steps. First, we use the

introduction text and knowledge points of the course to build a "course-knowledge point" binary network. Then, latent semantic analysis is carried out based on the weight matrix, the correlation score between curriculum and knowledge points is calculated, and the top k relevant knowledge points of the course are returned. Experiments on real-world public data sets have proved the effectiveness and accuracy of the system. However, latent semantic analysis cannot effectively solve the problem of text keyword ambiguity, and the text representation method is not optimized. In the future work, we will improve the stability and efficiency of the system and overcome the problem of polysemy.

Acknowledgment

This work was supported by National Natural Science Foundation of China under Grant 62002225, and Natural Science Foundation of Shanghai under Grant 21ZR1445400.

References

[1] Breslow L, Pritchard D E, Deboer J, et al. Studying Learning in the Worldwide Classroom Research into edX's First MOOC [J]. Research & Practice in Assessment, 2013, 8:13-25.
[2] Cole J, Foster H. Using Moodle: Teaching with the Popular Open Source Course Management System [M]. O'Reilly Media, Inc. 2007.
[3] Freeman S, Eddy S L, Mcdonough M, et al. Active learning increases student performance in science, engineering, and mathematics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111 (23):8410-5.
[4] Corbett A T, Anderson J R. Knowledge tracing: Modeling the acquisition of procedural knowledge [J]. User Modeling and User-Adapted Interaction, 4 (4):253-278.
[5] Xinwei Ren, Xianliang Jiang. Course recommendation for MOOC platform: A review. Journal of Ningbo University (Natural Science & Engineering Edition), 2022, 35 (1): 48-56. DOI:10.3969/j.issn.1001-5132.2022.01.008.
[6] Wen M, CP Rosé. Identifying Latent Study Habits by Mining Learner Behavior Patterns in Massive Open Online Courses [J]. ACM, 2014.
[7] Ghauth K I, Abdullah N A. Learning materials recommendation using good learners' ratings and content-based filtering [J]. Educational Technology Research & Development, 2010, 58 (6):711-727.
[8] Zhou Y, Huang C, Hu Q, et al. Personalized learning full-path recommendation model based on LSTM neural networks [J]. Information Sciences, 2018:S0020025518301397.
[9] Kontostathis A, Pottenger W M. A framework for understanding Latent Semantic Indexing (LSI) performance [J]. Information Processing & Management, 2006, 42 (1):56-73.
[10] Du W F. Research on Sentimental Lexicon Construction for Text Sentiment Analysis [D]. Harbin Institute of Technology, 2010.4.
[11] Liu T, Chen Z, Zhang B, et al. Improving text classification using local Latent Semantic Indexing [C]// Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on. IEEE, 2004.
[12] Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization [C]// International Conference on Machine Learning. 1996.
[13] Khribi, Mohamed Koutheaïr, Jemni, et al. Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. [J]. Journal of Educational Technology & Society, 2009.
[14] Sengottuvelan P, Gopalakrishnan T, Kumar R L, et al. A recommendation system for personal learning environments based on learner clicks [J]. International Journal of Applied Engineering Research, 2015, 10 (20):15316-15321.
[15] Ghauth K I, Abdullah N A. Learning materials recommendation using good learners' ratings and content-based filtering [J]. Educational Technology Research & Development, 2010, 58 (6):711-727.
[16] Aher S B, Lobo L M R J. Mining Association Rule in Classified Data for Course Recommender System in E-Learning [J]. International Journal of Computer Applications, 2012, 39 (7):1-7.
[17] Chen C M, Lee H M, Chen Y H. Personalized e-learning system using Item Response Theory [J]. Computers & Education, 2005, 44 (3):237-255.
[18] Okubo F, Yamashita T, Shimada A, et al. A neural network approach for students' performance prediction [C]// International Learning Analytics & Knowledge Conference. 2017.
[19] Li X, Li X, Tang J, et al. Improving Deep Item-Based Collaborative Filtering with Bayesian Personalized Ranking for MOOC Course Recommendation [M]. 2020.
[20] Yu J, Luo G, Xiao T, et al. MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.