Constructing a Thematic Knowledge Graph for Clinical Medicine Curriculum: Postprint

Authors: Lu Quan, Xie Yiyu, Chen Jing, Zhang Han, Cui Haoran, Nie Shuyuan

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] From the perspective of knowledge topics, this study establishes a comprehensive curriculum knowledge system to address the issues of repetitive knowledge points across courses and "knowledge silos" in existing curriculum design and teaching, thereby enabling effective professional knowledge services.

[Method/Process] Taking core courses in clinical medicine as the research object, and based on medical education data including medical subject headings, electronic textbooks, and electronic teaching plans, this study employs the LDA model to mine knowledge topics within courses, utilizes association analysis to reveal fine-grained relationships among courses, among knowledge topics, and between courses and knowledge topics, thereby constructing a knowledge topic graph for clinical medicine curricula.

[Results/Conclusion] This study constructs a domain knowledge graph from the perspectives of the professional curriculum system and knowledge topics, which assists teaching administrators, instructors, and students in mastering the professional knowledge system, facilitates knowledge-oriented teaching activities, and advances the development of knowledge organization and services in the medical domain as well as smart medical education.

Full Text

Research on the Construction of a Clinical Medicine Course Knowledge Topic Graph

Lu Quan[1,2], **Xie Yiyu**[1], **Chen Jing**[3], **Zhang Han**[1], **Cui Haoran**[1], **Nie Shuyuan**[1] [1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072 [2]Big Data Institute, Wuhan University, Wuhan 430072 [3]School of Information Management, Central China Normal University, Wuhan 430079

Abstract

[**Purpose/Significance**] This study approaches from the perspective of knowledge topics to establish a comprehensive curriculum knowledge system, addressing the problems of overlapping knowledge points and "knowledge islands" in existing curriculum design and teaching, thereby enabling effective professional knowledge services. [**Method/Process**] Taking the core courses of clinical medicine as the research object and based on medical education data including Medical Subject Headings, electronic textbooks, and electronic lesson plans, this study employs the LDA model to mine knowledge topics within courses and utilizes association analysis to reveal fine-grained relationships among courses, among knowledge topics, and between courses and knowledge topics, thus constructing a clinical medicine course knowledge topic graph. [**Result/Conclusion**] The research constructs a domain knowledge graph from the perspective of professional curriculum systems and knowledge topics, which helps teaching administrators, instructors, and students master the professional knowledge system, carry out knowledge-oriented teaching activities, and advance knowledge organization and services in the medical field as well as the development of smart medical education.

Keywords: course-knowledge topic graph; knowledge graph; LDA; association analysis; clinical medicine

Introduction

In the process of medical education, knowledge interoperability is common, and there exists a hierarchical sequence in learning across courses and knowledge points, where new knowledge often requires supplementation from previously learned material. However, existing curriculum design in medical disciplines suffers from problems such as repetition of courses and knowledge points [1] and "knowledge islands" related to knowledge topics. This makes it difficult for teachers to effectively organize professional knowledge system teaching activities and prevents students from quickly locating content related to currently studied knowledge points. Therefore, how to avoid repetition of knowledge points, focus on key core knowledge, and establish a comprehensive disciplinary knowledge structure represents a major need for medical professionals and constitutes an important foundational educational technology problem that must be solved for scientific curriculum design, curriculum system optimization, and deepened knowledge organization and services.

1 Research Status

Curriculum system construction is an important driver for promoting educational reform and development [3], making curriculum system research a consistently popular topic in teaching reform. Many scholars have explored solutions to the problems of knowledge point repetition and "knowledge islands" in curriculum systems. Hu Wentao [4] proposed a knowledge integration model based on knowledge maps from the perspective of knowledge management theory, providing new methods for reducing knowledge internal friction, controlling knowledge integration processes, and eliminating "knowledge islands." Zheng Ning [6] used natural language processing technology to extract algorithm knowledge names and construct ontologies to identify algorithm knowledge points in online programming resources, thereby organizing massive online programming resources according to knowledge structures and solving the "knowledge island" phenomenon. Regarding curriculum system architecture and construction, Shang Wei et al. [7] constructed a TF-CDIO e-commerce professional curriculum system based on work-process-oriented curriculum development ideas and the CDIO engineering education model, incorporating the teaching factory concept. Zhou Ming et al. [8] studied the innovative construction of information management professional curriculum systems from a big data perspective, proposing to identify professional characteristics from the standpoint of big data development to build new curriculum systems. For clinical medicine specifically, Li Li et al. [9] analyzed the relationship between the value of medical big data and teaching, arguing that the application of clinical big data will transform traditional ophthalmology clinical teaching systems.

However, courses are organizational forms of knowledge topics, and knowledge topics constitute the core content of courses. Curriculum system construction and utilization must be built upon deep mining and comprehensive mastery of professional knowledge systems. R.J. Todd [10] studied how students use existing course knowledge topics to transform discovered information into personal knowledge and mapped and measured changes in students' knowledge of curriculum topics. Zhu Ke et al. [11] used topic map technology to reorganize the knowledge organization methods of individual online courses, organizing knowledge points in multi-granularity and multi-level ways to achieve semantic association and intelligent classification of online course knowledge points, providing support for personalized learning and other learning modes. E. Melis et al. [12] developed a web-based universal learning system called ActiveMath that constructs learning materials for each knowledge topic, allowing learners to select target knowledge topics while the system selects relevant materials for those topics, thereby generating entire courses for learners.

In summary, although some scholars have studied the acquisition and representation of knowledge topics when researching curriculum systems, no studies have been found that approach from the perspective of knowledge topics to form a mapping graph between curriculum systems and knowledge topics while investigating the quantitative relationships between courses and knowledge topics. Therefore, this study conducts deep mining of clinical medicine education data, uses the LDA model to mine knowledge topics in courses, employs association analysis to reveal fine-grained relationships among courses, among knowledge topics, and between courses and knowledge topics, and constructs a clinical medicine course knowledge topic graph from the perspective of professional curriculum systems and knowledge topics. This helps teaching administrators, instructors, and students master professional knowledge systems, carry out knowledge-oriented teaching activities, and advance knowledge organization and services in the medical field as well as smart medical education development.

2 Clinical Medicine Course Knowledge Topic Graph Model

The clinical medicine course knowledge topic graph model mainly includes three sub-modules: (1) Clinical medicine education data preprocessing, which collects, segments, and removes stop words from required data to obtain model input files; (2) LDA topic mining, which uses the LDA algorithm to mine knowledge topics from texts; and (3) Association calculation, which calculates associations between topic words and weights of relationships between chapters and knowledge topics based on mined knowledge topics.

2.1 Clinical Medicine Education Data Preprocessing

The preprocessing process processes raw texts from clinical medicine courses, such as medical subject headings, electronic textbooks, and electronic lesson plans, ultimately generating the data format required for LDA topic mining. The specific workflow of the clinical medicine education data preprocessing module is shown in Figure 1 [Figure 1: see original paper].

2.1.1 Data Sources and Collection This study investigated the undergraduate talent cultivation plan and curriculum system for clinical medicine (five-year program) at Wuhan University [13] and selected 14 core courses as research objects. The courses include anatomy, histology and embryology, physiology, biochemistry and molecular biology, pharmacology, pathology, pathophysiology, medical microbiology, medical immunology, clinical skills, internal medicine, surgery, obstetrics and gynecology, and pediatrics. Course materials including course introductions, electronic textbooks, electronic lesson plans, and syllabi were collected. Course chapters were divided according to the table of contents of the eighth edition textbooks published by People's Medical Publishing House, and corresponding course materials were converted to text format, yielding 385 course chapter texts.

2.1.2 Segmentation and Stop Word Removal Knowledge topic words are the basic units of this study, requiring text segmentation to obtain input files for the LDA algorithm. Python was used to crawl subject terms from subject retrieval in the Chinese Biomedical Literature Database [14] as a segmentation dictionary. Stop word lists including the "Harbin Institute of Technology Stop Word List" and "Baidu Stop Word List" were integrated and deduplicated to obtain a comprehensive stop word list. The open-source Chinese segmentation tool jieba was used for word segmentation, and segmented texts were divided in the same manner as before, yielding 385 segmented course chapter texts, with each chapter text serving as an input document for the LDA topic mining module.

2.2 LDA Topic Mining

LDA (Latent Dirichlet Allocation) is a document topic generation model proposed by D.M. Blei et al. [15-16] that includes a three-layer structure of words, topics, and documents, and can be used to identify potential topic information in document collections. LDA adopts the bag-of-words approach, treating each document as a word frequency vector. Documents are composed of several mixed topics, with each topic being a probability distribution over words. For a given document set D = $\{d_1, d_2, \cdots, d\}$, the likelihood function of document generation is obtained from a given prior Dirichlet distribution through the following process [17]:

(1) For each document d in D, obtain the multinomial vector d of topics on document d from d ~ Dirichlet(α).

(2) For each topic z, obtain the multinomial vector ϕz of vocabulary on topic z from ϕz ~ Dirichlet(β).

(3) For vocabulary wd,i in document d, generate a topic zj that follows a multinomial distribution with parameter d, and generate the probability distribution P(wd,i|zj,β) of vocabulary wd,i according to the specific topic proportion β.

For document set D, the LDA topic extraction process can be summarized as finding parameters α and β that maximize P(D|α,β) based on   and z, where:

P(D|α,β) = P(,z|α,β) = P(|α) _{i=1}^N P(zi|)P(wi|zi,β) (1)

The Gibbs sampling method [18] is used to infer the hidden variables in the above formula, thereby calculating posterior probabilities. During the Gibbs sampling process, two matrices are obtained: topic-topic word and document-topic. These are iteratively calculated with P(zi|zi,wi) in a loop, and the distribution when values converge is the corresponding topic distribution. Based on   and z, the probability of each topic in the document can be obtained, and through probability calculation, the knowledge topic words in each document can be obtained.

The LDA topic mining module mines topics from preprocessed texts based on the LDA algorithm, utilizing the latent semantic information of documents to obtain knowledge topic words. Research by Zhu Zede et al. [19] also shows that LDA-based keyword extraction methods can better avoid using common words as keywords and solve the problem of words failing to comprehensively and accurately cover document topic information, thereby improving keyword extraction accuracy. The specific process of the LDA topic mining module is shown in Figure 2 [Figure 2: see original paper].

The main purpose of this study is to obtain knowledge topic words in documents without requiring topic classification. Therefore, referencing the parameter settings of Tang Xiaobo et al. [20] for microblog hotspot mining, the number of topics extracted from each document is set to k = 10. Based on literature research [21-22] and empirical values, α and β are set to $\alpha = 50/\mathrm{k}$ and $\beta = 0.01$. The maximum number of Gibbs sampling iterations is set to 1000. Experimental results show that these parameter settings perform well in the document collection. Then, based on the obtained topic-topic word and document-topic matrices, topic words are filtered according to the following rules: for each document d in document set $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, \cdots, \mathrm{d}_{385}\}$, if topic zj has weight vj and topic word wi in topic zj has weight vi, then the final weight of topic word wi in document d is vi $\times$ vj. Words are sorted by weight, and words with weights greater than or equal to the set threshold are selected as topic words. In this experiment, setting the weight threshold to 0.008 achieves satisfactory results.

After one calculation according to the above parameters and rules, knowledge topic words meeting the conditions in each document were obtained, yielding 385 chapter-knowledge topic word documents, with each document recording the knowledge topic words of the corresponding chapter. However, since LDA algorithm-generated topic words involve a random process with slight differences in each calculation, multiple iterative experiments are needed to observe the effectiveness of the obtained topic words. The iteration method is as follows:

(1) For document d in D, let the knowledge topic word set obtained from the first calculation be w_1, and the set obtained from the second calculation be w_2. Update the knowledge topic words of d to the union of the two sets: $\mathrm{w}_1 = \mathrm{w}_1 \quad \mathrm{w}_2$.

(2) Repeat step 1 n times until no new words are added to w_1 in each document, reaching a stable state.

This study conducted 7 iterative experiments until the knowledge topic word set for each document reached a stable state. Then, 50 texts were randomly selected, and the obtained knowledge topic words were compared against the teaching syllabus for those chapters, revealing that the mined results could serve as summaries of the chapter topics.

After obtaining the knowledge topic words for each document, considering that the LDA algorithm mines some non-topic words and that most knowledge topic words should be nouns, this study combined the "Modern Chinese Verb List" to remove verbs from the extracted knowledge topic words, improving the effectiveness of LDA topic mining. Finally, knowledge topic words for each chapter were obtained, forming a many-to-many mapping matrix of clinical medicine course knowledge topics by chapter, revealing the knowledge topics contained in chapters and the chapters covered by knowledge topics. Ultimately, 1,696 unique knowledge topic words were obtained. Table 1 shows information on knowledge topic words for some chapters.

Table 1 Knowledge Topic Words for Some Chapters

Chapter	Knowledge Topic Words
Pathophysiology	Sub-health, pathogenesis, neural mechanisms, humoral mechanisms, etiology, congenital, immunity, outcome
Circulatory System Diseases	Pediatrics, blood, circulatory system, heart, congenital heart disease, artery, blood vessels, atrium, veins
Pelvis and Perineum	Pelvis, perineum, pelvic bone, pelvic wall muscles, pelvic organs, pelvic fascia, fascial spaces, anal canal
Energy Metabolism and Body Temperature	Energy, metabolism, body temperature, blood glucose, hypoxia, muscle, protein, fat, heat production
Intracranial and Intraspinal Tumors	Surgery, intracranial tumors, intraspinal tumors, glioma, meningioma, acoustic neuroma, pituitary tumor, lymphoma

2.3 Association Calculation

Based on the many-to-many mapping matrix of course knowledge topics, the co-occurrence text frequency of N knowledge topic words is counted to form an N×N co-occurrence matrix for association analysis, thereby revealing fine-grained knowledge topic associations among courses. The association calculation module mainly computes three aspects: topic-topic associations, topic-chapter associations, and chapter-chapter associations.

2.3.1 Topic-Topic Association Calculation Association analysis is a knowledge discovery method that can quantitatively describe how much the appearance of item A affects the appearance of item B [23], commonly used in transaction databases such as sales data. Applying association analysis to the medical field can mine valuable information from complex medical materials. Zhang Han et al. [24] applied association rule algorithms to analyze the collocation patterns of subject headings and subheadings for antineoplastic drugs, extracting dependency relationships among subject headings and semantic relationship combinations related to five drug categories. For example, in the indexing of a literature about drug therapy, if it includes the subject heading "disease A/drug therapy" and also includes "drug B/therapeutic application," it indicates that drug B may have therapeutic efficacy for disease A.

For the course knowledge topic data in this study, a course chapter can be viewed as a transaction T composed of an item set of multiple knowledge topic words. To obtain semantic associations among knowledge topic words, association analysis can be performed on their co-occurrence matrix to mine knowledge topic words that frequently appear together under certain support and confidence conditions [25].

Topic-topic association calculation is based on the Apriori algorithm [26]. For the 385 knowledge topic word texts and 1,696 knowledge topic words obtained from the LDA topic mining module: first, the number of texts in which each knowledge topic word appears is counted (e.g., word A appears in CA texts and word B appears in CB texts). Then, based on the co-occurrence matrix, the total number of texts where each word pair {A,B} co-occurs, denoted as CA B, is obtained. For each directed word pair {A→B}, association rules with support greater than or equal to minimum support, confidence greater than or equal to minimum confidence, and lift greater than 1 are obtained [27-28].

Support describes the probability of words A and B appearing simultaneously in all texts, calculated as: Support(A→B) = P(A B) = CA B / 385 (2)

Confidence describes the probability that texts containing word A also contain word B, calculated as: Confidence(A→B) = P(B|A) = CA B / CA (3)

Lift describes the degree of influence of word A on word B, where lift greater than 1 indicates positive correlation, calculated as: Lift(A→B) = P(B|A) / (P(A)P(B)) = (CA B × 385) / (CA × CB) (4)

In this study, minimum support is set to 0.002 and minimum confidence to 0.5. Based on the above algorithm, 12,055 strong association rules were obtained, describing the unidirectional association degree of one knowledge topic word to another. This study defines three types of associations between topics: basic relationship, advanced relationship, and peer relationship. Based on the association calculation results, among all obtained association rules, the following three cases exist:

(1) If word pair {A,B} has only one association rule, i.e., A→B with confidence x, it indicates that topic word A influences the appearance of topic word B. Therefore, A is defined as the basic topic of B, and B is defined as the advanced topic of A, meaning that knowledge of topic word A must be acquired before learning topic word B, and after learning topic word A, one can continue to learn topic word B.

(2) If word pair {A,B} has two association rules, i.e., A→B with confidence x and B→A with confidence y, and x > y, it indicates that topic word A has greater influence on the appearance of topic word B than vice versa. Therefore, rule B→A is discarded, and A is defined as the basic topic of B, and B as the advanced topic of A. Conversely, if x < y, then B is the basic topic of A.

(3) If word pair {A,B} has two association rules, i.e., A→B with confidence x and B→A with confidence y, and x = y, it indicates that topic words A and B have equal influence. Therefore, the two rules are merged, and A and B are defined as peer topics, meaning topic words A and B can be learned in parallel.

Based on the above rules and the three relationships among topics, after deleting and merging the 12,055 association rules, 8,933 valid associations were obtained, including 6,632 peer relationships and 2,301 basic and advanced relationships.

Table 2 shows the relationships between "etiology" and other knowledge topic words.

Table 2 Relationships Between "Etiology" and Other Knowledge Topic Words

Knowledge Topic Words	Relationship Type
Pathogenesis, immunity	Advanced relationship
Molecular mechanisms, death, humoral mechanisms, neural mechanisms, histiocytic mechanisms, congenital, brain death, sub-health	Basic relationship
Outcome, symptoms	Peer relationship

2.3.2 Topic-Chapter Association Calculation Topic-chapter association reveals the weight of each topic in a chapter, i.e., the importance of knowledge topic words. The calculation is based on the TF-IDF algorithm [29]. For the 385 texts containing knowledge topic words:

First, the IDF (inverse document frequency) value of each topic word relative to the 385 chapters is calculated. For topic word i, if the total number of texts containing i is dfi, then: IDF(i) = log(385/dfi).

Then, the TF (term frequency) value of topic word i in the corresponding chapter text j after the preprocessing module is calculated and normalized according to the characteristics of this study: TF(i,j) = Ni,j / Nj, where Ni,j is the number of occurrences of topic word i in text j, and Nj is the total number of words in text j.

The TF-IDF value of topic word i in text j is: TF-IDF(i,j) = TF(i,j) × IDF(i) (5)

Additionally, this study calculates topic-course associations, which reveal the importance of knowledge topic words in courses. The 385 preprocessed chapter texts are merged by course into 14 texts, and the 385 chapter topic word texts after LDA topic mining are merged by course with duplicate topic words removed to obtain knowledge topic word texts for 14 courses. The calculation method for topic-course association is similar to that for topic-chapter association, computing the IDF and TF values for each topic word relative to the 14 courses and performing TF-IDF calculation and normalization.

2.3.3 Chapter-Course Association Calculation Chapter-course association reveals the importance of each chapter within a course. Logically, the knowledge topic words of a chapter can be viewed as a highly condensed representation of chapter content. Therefore, chapter-course association calculation is based on topic-course association calculation. For each chapter, the sum of TF-IDF values of all its topic words in the corresponding course is calculated and averaged by the number of topic words contained in the chapter. For chapter j in course c, if chapter j contains n topic words, the weight of chapter j in course c is:

w(j,c) = (_{i=0}^n TF-IDF(i,c)) / n (6)

After obtaining all chapter weights in course c, the weight of chapter j is normalized by dividing it by the sum of weights of all chapters in the course. If course c contains k chapters, the final weight of chapter j in course c is:

weight(j,c) = w(j,c) / (_{i=0}^k w(k,c)) (7)

2.3.4 Chapter-Chapter Association Calculation Since knowledge topic words are highly condensed representations of corresponding chapters, chapter-chapter association calculation is based on topic-topic association calculation. By calculating the sum of association degrees of all topic words between two chapters and averaging it based on the number of possible association rules between the two chapters, the corresponding chapter-chapter association can be represented. Assuming chapter A has x knowledge topic words, chapter B has y knowledge topic words, and z association rules appear among the knowledge topic words of the two chapters, with each rule having confidence c, the association weight calculation formula between chapter A and chapter B is proposed as:

w(A,B) = (_{i=0}^z c_i) / (x × y) (8)

3 Research Results

This study aims to construct a professional course knowledge topic graph for clinical medicine by mining and analyzing knowledge topics in course materials, thereby helping teachers and students intuitively understand important knowledge points, establish a comprehensive curriculum knowledge system, and improve teaching quality and learning effectiveness. After mining knowledge topic words in clinical medicine and associations among courses, chapters, and topics, the following were obtained: 1,696 topic words, 8,933 topic-topic associations, 4,194 chapter-topic associations, 3,308 topic-course associations, 385 chapter-course associations, and 16,120 chapter-chapter associations. A clinical medicine course knowledge topic knowledge base was constructed in database file format to store all data involved in the study.

Based on the clinical medicine course knowledge topic knowledge base, this study uses force-directed graphs to visualize the course knowledge topic graph structure [30] and utilizes the Baidu open-source tool ECharts to create the force-directed graphs. The following describes the research results from three aspects: overview of the clinical medicine course knowledge topic graph, chapter-chapter associations, and topic-topic associations.

3.1 Overview of Clinical Medicine Course Knowledge Topic Graph

The overall clinical medicine course knowledge topic graph radiates outward in three layers with the "Clinical Medicine Five-Year Program" node as the center, as shown in Figure 3 [Figure 3: see original paper]. From inner to outer, the first layer of larger nodes represents courses, the second layer nodes represent chapters, and the third layer leaf nodes represent knowledge topics, with connecting lines representing associations among courses, chapters, and topics.

Figure 3 only displays nodes with larger weights and closer associations, enabling clear and intuitive focus on core knowledge points. For example, in the *Internal Medicine* course, the four chapters "Urinary System Diseases," "Respiratory System Diseases," "Hematologic Diseases," and "Cardiology" are key chapters of the course. Important knowledge topic words in the "Cardiology" chapter include "artery," "myocardium," "heart," and "ventricle." The graph also shows that this chapter has close associations with the "Heart Failure" chapter in *Pathophysiology* and the "Electrocardiography" chapter in *Clinical Skills*. According to the partial order directionality between courses, before studying "Cardiology," students need to acquire relevant knowledge points from the "Heart Failure" and "Electrocardiography" chapters, where the topic word "myocardium" is a shared topic between the "Cardiology" and "Heart Failure" chapters, and "electrocardiogram" is a shared topic between the "Cardiology" and "Electrocardiography" chapters.

Based on the investigation of the cultivation plan and curriculum system for the clinical medicine five-year program at Wuhan University, the 14 core courses selected for this study can be divided into three categories: basic medical courses, transitional courses, and clinical medicine courses, as shown in Table 3 . These three types of courses have partial order directionality, and learning between courses follows certain logical and temporal sequences. Clinical medicine courses must be built upon the foundation of basic medical courses and transitional courses.

Table 3 Three Types of Courses

Course Type	Courses
Basic Medical Courses	Anatomy, Histology and Embryology, Physiology, Biochemistry and Molecular Biology, Pharmacology, Pathology, Pathophysiology, Medical Microbiology, Medical Immunology
Basic-to-Clinical Transitional Course	Clinical Skills

Course Type	Courses
Clinical Medicine Courses	Internal Medicine, Surgery, Obstetrics and Gynecology, Pediatrics

3.2 Chapter-Chapter Associations

For a chapter node in a course, the graph displays the names of chapters associated with it. As shown in Figure 4 [Figure 4: see original paper], for the "Respiratory System Diseases"chapter in *Internal Medicine*, chapters with strong associations include "Splanchnology," "Thorax," "Respiratory Failure," "Respiratory System," and "Stress." The distance between chapters and the central point in the graph indicates the degree of association, with closer distances representing stronger associations with the "Respiratory System Diseases" chapter. Specific information includes chapter introductions, related chapters, and key topics for that chapter.

3.3 Topic-Topic Associations

For a topic node, the graph displays the names of topics associated with it. As shown in Figure 5 [Figure 5: see original paper], for the topic "asthma," its basic topics include "thorax,""acute upper respiratory infection,""rhinitis,""bronchitis," etc., indicating that knowledge of basic topic words is required before learning this topic. The advanced topic is "pneumonia," which may trigger asthma —both are distinct yet related, and after learning "asthma," students should continue to understand "pneumonia." Peer topics are "choline," "chlamydia," and "mycoplasma," which have no partial order relationship with "asthma" and can be learned simultaneously. Specific information includes topic introductions and affiliated chapters. Chapter 12 of Pediatrics, Chapter 2 of Internal Medicine, and Chapter 5 of Pharmacology all contain this topic.

Conclusion

This study conducted deep mining of course materials from the clinical medicine five-year program at Wuhan University, obtained associations among courses, knowledge topics, and between courses and knowledge topics in clinical medicine, revealed fine-grained knowledge topic associations between courses, constructed a clinical medicine course knowledge topic knowledge base and graph, and developed a query system for visualizing the knowledge graph, accessible at http://218.197.150.149/rainbow.

At the theoretical level, this study combines text mining technology and informatics theories with professional curriculum knowledge system research, constructing a domain-specific knowledge system from the perspective of professional curriculum systems and knowledge topics, which serves as a beneficial supplement to existing knowledge graph theories. At the application level, in-depth mining of the clinical medicine course knowledge topic graph can break down knowledge barriers among multiple courses within the major, helping teaching administrators, instructors, and students master professional knowledge systems and carry out knowledge-oriented teaching activities. On one hand, the research results can provide key theoretical and technical foundations for teaching administrators to scientifically manage professional knowledge systems, systematically optimize curriculum systems, and assist in course scheduling and teaching team building. On the other hand, they can be applied in actual teaching to help teachers reasonably organize professional knowledge points, optimize teaching plans, and improve teaching quality, while helping students deeply understand fine-grained knowledge topic associations between courses, plan and systematically learn effectively, and promote the understanding, utilization, and sublimation of disciplinary knowledge among teachers and students. This research has universal significance and application value for professional talent cultivation in China, knowledge organization and services in the medical field, and smart medical education.

This study still has many limitations. For example, when calculating chapter-chapter associations, only association rules of knowledge topic word co-occurrence were considered without accounting for their distribution characteristics across the two chapters, which may not accurately reveal association weights between chapters. Therefore, the algorithm could be further improved by combining association rules with topic word distribution characteristics. Additionally, existing curriculum systems may be inaccurate and could be optimized by incorporating professional domain knowledge graphs (such as sub-graph analysis on LinkedLifeData).

References

[1] Yang Yang, Fan Yuxia, Wang Dan, et al. Discussion on systematic teaching in the curriculum reform of clinical medicine five-year undergraduate program[J]. Test Questions and Research: Teaching Forum, 2016(22): 19-19.

[2] Ministry of Education, Ministry of Health. Standards for Undergraduate Medical Education–Clinical Medicine Major (Trial)[EB/OL]. [2018-07-24]. http://www.moe.gov.cn/srcsite/A08/moe_{740}/s3864/200809/t20080916_{109605}.html.

[3] Qin Lei. Curriculum construction is an important driver for connotative development[N]. China Education Daily, 2015-06-02(10).

[4] Hu Wentao. Research on learning path graph generation technology based on knowledge graphs[D]. Beijing: Beijing University of Posts and Telecommunications, 2017.

[5] Ye Chunsen, Wang Chuanlei, Chu Jiewang. Research on knowledge integration models and mechanisms based on knowledge maps[J]. Information Studies: Theory & Application, 2009, 32(10): 52-54.

[6] Zheng Ning. Research on knowledge discovery for programming resources

problem-solving based on natural language processing[D]. Shanghai: Donghua University, 2014.

[7] Shang Wei, Pan Honghua, Guo Feipeng. Construction of TF-CDIO e-commerce professional curriculum system[J]. Research in Higher Education of Engineering, 2012(2): 146-151.

[8] Zhou Ming, Xie Jun. Research on innovative training models for information management majors from a big data perspective[J]. Research on Library Science, 2016(6): 41-46.

[9] Li Li, Bai Dayong, Zhang Chengti, et al. Discussion on optimizing ophthalmology teaching system based on big data technology[J]. Chinese Hospital Management, 2015, 35(8): 51-53.

[10] TODD R J. From information to knowledge: charting and measuring changes in students' knowledge of curriculum topic[J]. Information research international electronic journal, 2006, 11(4): 264-264.

[11] Zhu Ke, Liu Qingtang, Ye Yangmei. Research on network course knowledge organization based on topic maps[J]. e-Education Research, 2014(1): 91-96.

[12] MELIS E, ANDRES E, FRISCHAUF A, et al. ActiveMath: a generic and adaptive web-based learning environment[J]. International journal of artificial intelligence in education, 2001, 12(4): 385-407.

[13] Wuhan University School of Medicine. Clinical Medicine (Five-Year) Undergraduate Talent Cultivation Plan (2013 Edition)[EB/OL]. [2018-07-25]. http://wsm70.whu.edu.cn/content1.jsp?urltype=news.NewsContentUrl&wbtreeid=1055&wbnewsid=7811.

[14] Chinese Academy of Medical Sciences. Chinese Biomedical Literature Database[EB/OL]. [2018-11-23]. http://www.sinomed.ac.cn/zh/.

[15] BLEI D M, Ng A Y, JORDAN M I. Latent dirichlet allocation[J]. Machine learning research archive, 2003, 3: 993-1022.

[16] HOFFMAN M D, BLEI D M, BACH F. Online learning for latent Dirichlet allocation[C]//International conference on neural information processing systems. New York: Curran Associates Inc., 2010: 856-864.

[17] Tang Xiaobo, Fang Xiaoke. Research on a microblog topic detection model based on text clustering and LDA fusion[J]. Information Studies: Theory & Application, 2013, 36(8): 85-90.

[18] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the national academy of sciences of the United States of America, 2004, 101(S1): 5228-5235.

[19] Zhu Zede, Li Miao, Zhang Jian, et al. A keyword extraction method based on the LDA model[J]. Journal of South-Central Minzu University (Natural Science Edition), 2015(6): 2142-2148.

[20] Tang Xiaobo, Xiang Kun. Hotspot mining based on LDA model and microblog heat[J]. Library and Information Service, 2014, 58(5): 58-63.

[21] Hu Jiming, Chen Guo. Content topic mining and evolution based on dynamic LDA topic model[J]. Library and Information Service, 2014, 58(2): 138-142.

[22] GUO X, XIANG Y, CHEN Q, et al. LDA-based online topic detection using tensor factorization[J]. Journal of information science, 2013, 39(4): 459-469.

[23] Shi Zhongzhi. Knowledge Discovery[M]. Beijing: Tsinghua University Press, 2002.

[24] Zhang Han, Lu Zhenyu, Cui Lei. An attempt to use association rules for knowledge extraction from medical text databases—taking four antineoplastic drugs as examples[J]. New Technology of Library and Information Service, 2006, 1(9): 49-52.

[25] HAN J. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Publishers Inc., 2005.

[26] INOKUCHI A, WASHIO T, MOTODA H. An apriori-based algorithm for mining frequent substructures from graph data[C]//European conference on principles of data mining and knowledge discovery. Berlin: Springer-Verlag, 2000: 13-23.

[27] Zhong Weijin, Li Jia. Research on co-word analysis (I)—process and methods of co-word analysis[J]. Journal of Intelligence, 2008, 27(5): 70-72.

[28] Gao Jiping, Ding Yi, Pan Yuntao, et al. Implementation of multi-word co-occurrence analysis method and its application in research hotspot identification[J]. Library and Information Service, 2014, 58(24): 80-85.

[29] WU H C, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. ACM transactions on information systems, 2008, 26(3): 55-59.

[30] Song Meina, Cui Danyang, E Haihong, et al. Design and implementation of a general data visualization model[J]. Computer Applications and Software, 2017, 34(9): 38-42.

Author Contributions

Lu Quan: Proposed research questions, designed research framework, provided paper revision suggestions. Xie Yiyu: Conducted data processing, wrote main paper content and revised paper. Chen Jing: Proposed research ideas, revised paper. Zhang Han: Conducted model implementation and data processing. Cui Haoran: Conducted knowledge graph visualization. Nie Shuyuan: Conducted research and collected required data.

Note: Figure translations are in progress. See original paper for figures.