Automatic Construction of Curriculum Knowledge Models Based on Web Recruitment Text Mining (Postprint)

Authors: Yu Yan, Chen Lei, Zhao Naixuan

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

Abstract

[Purpose/Significance] To assist university faculty and students in effectively utilizing online recruitment information, this paper proposes a curriculum knowledge model based on large-scale online recruitment text mining and its automatic construction method.

[Method/Process] This paper presents a three-level curriculum knowledge model comprising "position-course-knowledge point", employs natural language text mining techniques to achieve automatic construction of the curriculum knowledge point model, and validates and analyzes its construction process through experiments.

[Results/Conclusion] Experimental results demonstrate that the proposed model and method are highly feasible and effective, and can provide teaching and learning references for universities and students.

Full Text

Preamble

Volume 63, Issue 10, May 2019

Research on Automatic Construction of Curriculum Knowledge Model Based on Web Recruitment Text Mining

Yu Yan[1],[2], Chen Lei[1], Zhao Naixuan[1]
[1]Information Service Department, Nanjing Tech University, Nanjing 210009
[2]Computer Engineering Department, Southeast University Chengxian College, Nanjing 211816

Abstract

[**Purpose/Significance**] To help university teachers and students make full use of web recruitment information, this paper proposes a curriculum knowledge model and its automatic construction method based on large-scale web recruitment text mining. [**Method/Process**] This paper proposes a three-level curriculum knowledge model containing "position–curriculum–knowledge point," uses natural language text mining technology to achieve automatic construction of the curriculum knowledge model, and verifies and analyzes the construction process through experiments. [**Result/Conclusion**] Experimental results demonstrate that the proposed model and method are highly feasible and effective, providing teaching and learning references for universities and students.

Keywords: Web recruitment text; Curriculum knowledge model; Text mining
Classification Number: G202
DOI: 10.13266/j.issn.0252-3116.2019.10.015

Introduction

In recent years, with the rapid development of higher education and the expanding enrollment scale, the difficulty of college graduates finding jobs and enterprises recruiting talent has become a hot social issue. To some extent, the mismatch between talent cultivation in Chinese universities and social demand has created this dual dilemma. Particularly in the information age, enterprise demands for talent change rapidly, while university talent cultivation cycles are long and professional curriculum settings lag behind, resulting in student training that deviates from actual needs. Therefore, in the rapidly developing information age, quickly and accurately understanding the knowledge requirements of enterprises for recruited positions is extremely important.

With the popularization of the Internet, online recruitment has become the mainstream method for enterprise hiring. Web recruitment texts often contain specific descriptions of enterprises' professional knowledge requirements for positions, reflecting current market demands for talent expertise. Therefore, web recruitment text analysis is a technique for understanding society's knowledge requirements for talent in specific fields. This paper proposes automatically constructing a curriculum knowledge model to adapt to the characteristics of large data volumes and rapid data changes in the big data era. Finally, an empirical analysis of web recruitment texts for computer-related majors is conducted. The empirical results demonstrate the feasibility and effectiveness of the proposed model and construction process. The curriculum knowledge model can help universities continuously optimize professional curriculum systems and syllabi according to society's skill demands for talent in specific fields, providing intelligence decision support for developing professional talent training programs that meet enterprise needs. The model can also help students focus on strengthening certain professional courses and knowledge points based on their interests and

desired positions.

1 Related Research

Web recruitment text analysis typically includes two steps: recruitment entity information extraction and recruitment entity analysis. Recruitment entity information extraction refers to extracting structured recruitment entity information (such as position, skills, major, etc.) from semi-structured web recruitment texts. Based on extraction methods, these can be divided into manual and automatic approaches.

Manual methods directly extract position, required skills, and other information from web recruitment texts through human effort. For example, C. Chao and S. Shih [1] collected information from the Monster recruitment website and manually extracted recruitment positions and skills; I. Wowczko [2] manually extracted and mapped skills in recruitment advertisements; J. Y. Kim and C. K. Lee [3] manually analyzed data scientist recruitment information; D. A. Mauro et al. [4] manually extracted skills required for job types; Lü Bin et al. [5] and Li Guoqiu [6] surveyed 300 intelligence professional recruitment webpages, manually extracting intelligence professional requirements, job types, responsibilities, and functions; Xia Huosong and Pan Xiaoting [7] manually extracted talent demand information from Chinese big data enterprises; Huang Yue et al. [8] manually extracted requirements for talent positions, knowledge, and capabilities for library and information positions from web recruitment texts; Jia Dongqin and Tan Bo [9] manually extracted recruitment entities from three recruitment websites: ALA Joblist, IFLA's LIBJOBS Mailing List, and ACRL.

Obviously, manual methods cannot meet the requirements of analyzing web recruitment information in large-volume, unstructured environments. Some studies have attempted to use external resource-based, rule-based, statistical, and deep learning methods to automatically extract information from web recruitment texts. External resource-based methods utilize skill dictionaries, Wikipedia, and other resources to build professional knowledge dictionaries for information extraction. For instance, M. Sodhi and B. Son [10] built a professional core dictionary for operations research; M. Zhao et al. [11] used conventional phrases and various terms predefined by domain experts to analyze recruitment webpages; T. Xu et al. [12] downloaded skill categories and specific skills from the CSDN website, totaling 54 skill categories and 1,729 specific skills, to build a professional knowledge dictionary; Zhan Chuan [13] built a terminology dictionary for e-commerce based on existing knowledge and extracted skills with frequencies above a certain threshold from recruitment texts; Xia Lixin et al. [14] used China Education Online Career Encyclopedia, recruitment website position classifications, and paper keywords to build dictionaries for majors, positions, and knowledge points, extracting information on majors, positions, and skills. However, external resource-based methods suffer from slow updates and narrow coverage.

Rule-based methods manually construct rule templates to achieve information extraction. For example, M. Bastian et al. [15] used commas for matching to extract skill information from LinkedIn recruitment texts; Wang Zhaoyi et al. [16] used four words—"possess," "familiar with," "proficient in," and "ability"—as proximity words to build extraction rules for extracting position-required skills. Rule-based methods have issues such as overly simplistic approaches and unsatisfactory results.

Statistical methods primarily use corpus statistics on probability information of certain words to extract recruitment entities. For example, Liu Ruilun et al. [17] used word frequency statistics to extract recruitment entities; Zhang Junfeng and Wei Ruibin [18] crawled data from professional recruitment websites such as 51job and Zhaopin, using word frequency and other methods to extract recruitment entities to build recruitment dictionaries. Statistical methods also suffer from overly simplistic approaches and unsatisfactory results.

With the rapid development of deep learning, Wang Dongbo et al. [19] used deep learning models to design a platform for automatic extraction of data science recruitment entities. However, deep learning methods require large-scale manually annotated corpora as training data, and currently, there are no large-scale annotated corpora for web recruitment skill information extraction tasks.

Network entity information analysis refers to the process of analyzing extracted structured recruitment entity information. Current analysis mainly involves statistics on extracted positions, skills, majors, and other information, without fully utilizing web recruitment text information. For example, J. Y. Kim et al. [3] analyzed the requirements for data scientist positions; D. A. Mauro et al. [4] combined expert judgment to analyze 2,700 big data-related position information, evaluating the skills and proficiency requirements needed for each job type; Lü Bin et al. [5] and Li Guoqiu [6] surveyed 300 intelligence professional recruitment webpages, analyzing the demand for intelligence professionals in social organizations, as well as job types, responsibilities, and functions; Huang Yue et al. [8] analyzed the knowledge and capability requirements for big data positions from three perspectives: basic position information, job responsibilities, and qualification requirements; M. Sodhi and B. Son [10] studied differences in operations research professional skill demands across different industries; Zhan Chuan [13] analyzed the skill requirements for various e-commerce positions, overall skill demands, and specific skills required for each position; Wei Lai and Zheng Huamin [20] investigated domestic and international university library recruitment information, analyzing data librarians' professional capabilities from five aspects: statistical knowledge background, comprehensive quality, job responsibilities, professional skills, and special skills; Jia Dongqin and Tan Bo [9] analyzed foreign university library position requirements, including job quantities, job responsibility requirements, entry qualifications, and bonus qualifications; Tian Ye [21] conducted an empirical analysis of library and information professional recruitment demand data from 1,359 institutions in 2016, examining recruiting institution types, numbers, regions, educational requirements,

and position preferences; Chen Yuanyuan and Dong Wei [22] used social network analysis to analyze the employment skills of library and information graduates under social demand orientation.

Although some scholars have recognized the importance of web recruitment text analysis and conducted relevant research, current studies still have two main problems: (1) research mainly conducts statistics on skills and knowledge required for positions without further utilizing web recruitment text information; (2) analysis primarily relies on manual methods, which cannot meet the requirements of large data volumes and rapid changes in recruitment network data in the big data era.

2 Logical Structure of the Curriculum Knowledge Model

Current research mainly conducts statistics on skill knowledge points required for positions, as shown in Figure 1 Figure 1: see original paper, without further utilizing web recruitment text information. To address this problem, this paper proposes a three-level curriculum knowledge model containing "position–curriculum–knowledge point," as shown in Figure 1(b).

The curriculum knowledge model includes three objects: position, curriculum, and knowledge. A position is the sum of one or more responsibilities that an enterprise requires employees to complete and the authority granted to employees; a curriculum refers to the professional knowledge and specialized skills courses offered by universities according to training objectives; a knowledge point is the knowledge and professional skills required for a position and also the basic unit of knowledge contained in a curriculum. The curriculum knowledge model also includes two types of relationships: position–curriculum and curriculum–knowledge. There is a many-to-many relationship between positions and curricula, meaning one position requires learning several courses, and one course can be applied to several related positions. There is also a many-to-many relationship between curricula and knowledge points, meaning one curriculum contains several knowledge points, and one knowledge point can belong to several curricula. Figure 2 [Figure 2: see original paper] uses the "Big Data Engineer" position as an example to show the main courses that should be learned for this position and the main knowledge points included in each course, where the thickness of the lines between objects indicates the strength of the relationship.

3 Automatic Construction of the Curriculum Knowledge Model

Web recruitment texts typically contain information such as position, job responsibilities, and qualification requirements. Figure 3 [Figure 3: see original paper] shows an example of a web recruitment text. Figure 3(a) displays the web recruitment text in a browser, and Figure 3(b) shows the corresponding HTML text. The position describes the job title for which the enterprise is recruiting; job responsibilities describe the duties that the position needs to

undertake; qualification requirements describe the professional knowledge and skills that the position holder should possess, as well as other basic abilities.

Position information in the curriculum knowledge model can be directly extracted from the position section of recruitment web texts, and knowledge points can be extracted from the text corresponding to qualification requirements. Curriculum information is generated using a topic model, and position–curriculum relationships and curriculum–knowledge point relationships are generated based on the topic model and statistical information. Therefore, the curriculum knowledge model construction process proposed in this paper is shown in Figure 4 [Figure 4: see original paper], mainly including six steps: data crawling, position extraction, knowledge point extraction, curriculum generation, position–curriculum relationship generation, and curriculum–knowledge point relationship generation.

3.1 Data Crawling

Select appropriate recruitment websites, choose relevant majors, and use Python scripts to first obtain the URLs of web recruitment texts and push them to the Django Rest interface collection endpoint, then crawl web recruitment text information one by one based on the position URLs from the collection endpoint.

3.2 Position Extraction

To extract position information from web recruitment texts, this paper uses BeautifulSoup to convert HTML text into a tree structure, where each node corresponds to a Python object. BeautifulSoup is a Python library that can extract data from HTML or XML files. It provides navigation, search, and even modification of the parse tree through a custom parser. Therefore, this paper uses BeautifulSoup to obtain the text within the "position" tag.

3.3 Knowledge Point Extraction

Similar to position extraction, BeautifulSoup is also used to parse the text in the "qualification requirements" tag of web recruitment texts. The text then undergoes preprocessing including word segmentation, part-of-speech tagging, stop word removal, and English case conversion. Figure 5 [Figure 5: see original paper] shows a preprocessing example, where knowledge points to be extracted are marked in bold. This paper uses the jieba extension package in Python for word segmentation and part-of-speech tagging, and uses the stop word list compiled by Harbin Institute of Technology to filter out stop words.

To extract knowledge points from preprocessed text, traditional methods typically use word frequency to extract frequently occurring words in the corpus as knowledge points. However, word frequency-based methods have low extraction accuracy and often include non-knowledge-point words such as "ability" and "experience." Knowledge points have professional relevance, appearing frequently in certain majors but rarely in others. Therefore, this paper introduces

an auxiliary set containing other major collections and proposes measuring the importance of words in a major based on Auxiliary Set Importance (ASI) to extract knowledge points. The basic principle is: the higher the frequency of a word in the target set and the lower its frequency in the auxiliary set, the more likely it is to be a knowledge point of the target major.

Specifically, let TS be the target set to be analyzed, and AS be the auxiliary set containing recruitment information from other majors. The professional importance ASI(wi, TS) of a word wi in target set TS is defined as:

$$ASI(w, TS) = \frac{df(w_i, TS) + 1}{df(w_i, AS) + 1} \cdot \frac{|TS|}{|AS|}$$

where $df(w_i, TS)$ represents the number of texts containing w_i in the TS collection; $df(w_i, AS)$ represents the number of texts containing w_i in the AS collection; $|TS|$ represents the number of texts in the TS collection; and $|AS|$ represents the number of texts in the AS collection. Since skills are typically nouns, this paper selects nouns in the "qualification requirements" as candidate words, measures their professional importance, and extracts knowledge points by ranking them according to professional importance.

3.4 Curriculum Generation

This paper uses the Latent Dirichlet Allocation (LDA) model to generate implicit curriculum information. The LDA topic model [23] is a commonly used three-layer Bayesian probability model in natural language processing. The model consists of three layers: words, topics, and texts, as shown in Figure 6 Figure 6: see original paper. The model assumes that each text contains several implicit topics, and each topic contains specific words. The relationship between texts and words is reflected through implicit topics. Implicit topics are shared by all texts in the corpus, while each text has a specific topic distribution. The construction process of a text first selects a topic with a certain probability, then selects a word under this topic with a certain probability, thus generating the first word of the text. Repeating this process continuously generates the entire document.

Similarly, meeting the requirements of a position requires learning multiple courses, and each course contains several knowledge points, as shown in Figure 6(b). Therefore, this paper proposes using the LDA topic model to generate implicit curriculum information.

The LDA topic model typically uses Gibbs sampling inference methods to estimate the posterior distribution of topics, calculated as shown in Formula (2) [24]:

$$p(z_{ij} = k | z_{-ij}, w, \alpha, \beta) \propto \frac{n_{(\cdot)jk} + \alpha}{n_{(\cdot)j(\cdot)} + K\alpha} \cdot \frac{n_{i(\cdot)k} + \beta}{n_{(\cdot)(\cdot)k} + V\beta}$$

where z_{ij} represents the curriculum variable of knowledge point w_i in position d_j; $-ij$ represents excluding knowledge point w_i in position d_j; n_{ijk} represents the number of times knowledge point w_i in position d_j is assigned to curriculum z_k; $(\cdot)$ represents the sum of all counts in the corresponding dimension (position, curriculum, knowledge point); β represents the Dirichlet prior distribution of knowledge points; α represents the Dirichlet prior distribution of curricula; K represents the number of curricula; and V represents the total number of knowledge points in the collection. Once the curriculum for each knowledge point in each position is obtained, the posterior estimates of θ and ϕ in the LDA model can be obtained, calculated as shown in Formula (3) [24] and Formula (4) [24]:

$$\theta_{jk} = \frac{n_{(\cdot)jk} + \alpha}{n_{(\cdot)j(\cdot)} + K\alpha}$$

$$\phi_{ki} = \frac{n_{i(\cdot)k} + \beta}{n_{(\cdot)(\cdot)k} + V\beta}$$

where θ_{jk} represents the probability that position d_j contains curriculum z_k; and ϕ_{ki} represents the probability that curriculum z_k contains knowledge point w_i.

3.5 Position–Curriculum Relationship Generation

The correlation r is used to represent the relationship strength between position d_i and curriculum z_k, indicating the ratio of the average probability that position d_i contains curriculum z_k to the average probability that all positions contain curriculum z_k. It is defined as follows:

$$r(d_i, z_k) = \frac{\sum_{d_i \in D_i} \theta_{ik}}{\sum_{d_j \in TS} \theta_{jk}}$$

where $D_i = \{d_i | d_i \in TS\}$ represents the number of web recruitment texts in collection TS that contain position d_i. From Formula (5), it can be seen that a larger r value indicates a stronger relationship between position d_i and curriculum z_k; conversely, the relationship is weaker.

3.6 Curriculum–Knowledge Point Relationship Generation

The curriculum–knowledge point relationship indicates the main knowledge points contained in a specific curriculum. Since the LDA topic model can obtain ϕ_{ki} representing the probability of knowledge point w_i in curriculum z_k, the top several knowledge points are selected for each curriculum to generate the curriculum–knowledge point relationship, using ϕ_{ki} to represent the relationship strength of knowledge point w_i in curriculum z_k.

4 Empirical Analysis

4.1 Data Crawling

To verify the feasibility and effectiveness of the proposed method, experiments selected computer-related majors from the mainstream domestic recruitment website 51job (www.51job.com). 51job is a web recruitment service provider and one of the most influential talent recruitment websites in China. According to job functions, data was crawled from the "Computer/Internet/Communication/Electronics" category on the 51job website (crawling dates: March 19-26, 2018). To obtain the auxiliary set, data was sequentially crawled from the "Sales/Customer Service/Technical Support," "Accounting/Finance/Banking/Insurance," "Production/Operations/Purchasing/Logistics," "Biology/Pharmaceuticals/Medical/Nursing," "Advertising/Marketing/Media/Arts," "Construction/Real Estate," "Human Resources/Administration/Senior Management," and "Service Industry" categories on the 51job website (crawling dates: March 19-26, 2018). After crawling, recruitment texts with education below bachelor's degree, duplicate content, full English text, and those without qualification requirements were removed. The basic information of the final dataset is shown in Table 1 .

Table 1 Basic Dataset Information

Dataset Type	Number of Recruitment Texts
Computer/Internet/Communication/Electronics	10,000
Sales/Customer Service/Technical Support	10,000
Accounting/Finance/Banking/Insurance	10,000
Production/Operations/Purchasing/Logistics	10,000
Biology/Pharmaceuticals/Medical/Nursing	10,000
Advertising/Marketing/Media/Arts	10,000
Construction/Real Estate	10,000
Human Resources/Administration/Senior Management	10,000

4.2 Position Extraction

Through statistical analysis of word frequencies in position names from recruitment webpages, after removing words such as "development," "R&D," and "engineer" that cannot represent clear positions, the top 10 high-frequency words forming positions and the standardized position names provided in this paper are shown in Table 2 . As can be seen from Table 2, computer science technology changes very rapidly. Although there are consistently popular positions such as Java Engineer, C++ Development Engineer, and .NET Engineer, there are also new positions with rapidly growing demand, such as Front-end Development Engineer, Big Data Engineer, and Algorithm Engineer.

Table 2 Top 10 Positions in Computer-Related Majors

Position Names Containing Keywords	Standardized Position Name
Java Software Engineer, Java Engineer, Java Development Engineer	Java Engineer
Big Data Engineer, Big Data R&D Engineer, Big Data R&D Personnel	Big Data Engineer
C++ Development Engineer, C++ Software Engineer, C++ Software Development Engineer	C++ Development Engineer
.NET Development Engineer, .NET Engineer, .NET Software Development Engineer	.NET Engineer
Web Front-end Development Engineer, Front-end Development Engineer	Front-end Development Engineer
Test Engineer, Software Test Engineer	Test Engineer
Operations Engineer, System Operations Engineer	Operations Engineer
Embedded Software Engineer, Embedded Software Development Engineer	Embedded Software Engineer
IOS Development Engineer, IOS Mobile R&D Engineer	IOS Development Engineer
Algorithm Engineer, AI Algorithm Engineer, Image Processing Algorithm Engineer, NLP Algorithm Engineer	Algorithm Engineer

4.3 Knowledge Point Extraction

Knowledge point extraction is generated by ranking ASI values, with manual verification used for judgment. To avoid subjectivity and limitations of professional knowledge, knowledge websites such as Baidu Baike, Wikipedia, and Hudong Baike are used to check whether corresponding knowledge point entries exist to verify the correctness of extracted knowledge points.

Table 3 shows the top 10 words extracted using the TF method and the method proposed in this paper, along with their corresponding values, where bold words indicate non-knowledge-point words. As can be seen from Table 3, among the top 10 words from the TF method, non-knowledge-point words appear frequently in the target set and cannot effectively extract knowledge points from the target set. In contrast, the top 10 words identified using the ASI method are all skills, significantly outperforming the TF method, because high-frequency words in the target set such as “experience” and “ability” also appear frequently in the auxiliary set, making their ASI values smaller.

Table 3 Comparison of Top 10 Words Identified by Different Methods

TF Method	ASI Method
Word	Value
Python	2054.952
MySQL	2442.615
JavaScript	2194.096
Experience	1502.328
Ability	3950.752

4.4 Curriculum Generation

To generate curricula, the LDA model is used. According to common parameter setting methods [23-24], the topic model is set with $\alpha = 50/K$, $\beta = 0.01$, Gibbs sampling iteration parameter of 2000, and saving iteration parameter of 1000. The selection of curriculum number K uses perplexity calculation and expert evaluation to select the optimal value, employing five-fold cross-validation. Based on calculations, the experiment sets the curriculum number $K = 11$. Table 4 lists the top 5 knowledge points for each curriculum and the corresponding curriculum names summarized.

Table 4 Curriculum Name Generation

Curriculum	Top 5 Knowledge Points	Curriculum Name
Web HTML JavaScript CSS HTML5	Web Development	
Java J2EE Spring Framework Hibernate	Java	
C# .NET Winform Object-Oriented	C#	
Software Architecture Programming	Software Engineering	
C++ C Linux Unix	C/C++	
Database Oracle SQL MySQL Stored Procedures	Database	
Linux Bottom Layer Process Shell Communication Protocol	Linux	
TCP/IP HTTP Communication	Network Communication	
Bug Test Case Unit Testing White Box	Software Testing	
Python C++ Vision AI	Algorithm	
Data Analysis MapReduce Storage Modeling Data Mining	Data Analysis	

Among these 11 curricula, some are computer professional courses that many universities have offered for many years, such as “Java,” “C++,” “C#,” “Database,” “Linux,” “Network Communication,” “Software Testing,” and “Software Engineering.” There are also courses that have emerged with big data, such as “Web Development” and “Data Analysis.”

4.5 Position–Curriculum Relationship Generation

Based on the calculation of correlation between positions and curricula, the relationship strength between positions and curricula is obtained. Table 5 lists the relationships between positions and curricula.

Table 5 Position–Curriculum Relationships

Position	Java	C#	C/C++	Linux	Database	Web Development	Software Engineering	Software Testing	Algorithms	Data Analysis	Network Communication
Java Engineer											
Big Data Engineer											
C++ Development Engineer											
.NET Engineer											
Front-end Development Engineer											
Test Engineer											
Operations Engineer											

Position	Java	C#	C/C++	Linux	Database	Web Development	Software Engineering	Software Testing	Algorithms	Data Analysis	Network Communication
Embedded Software Engineer											
IOS Development Engineer											
Algorithm Engineer											

Note: *indicates position–curriculum correlation* $r \in [0.9, 1)$, *indicates* $r \in [1, 1.5)$, *indicates* $r \in [1.5, 3)$.

Through the results in Table 5, the main courses required for each position can be identified: "Java Engineer" uses the Java development language to complete software product program design and development, mainly responsible for designing and developing core backend business and external service interfaces for the operation platform, typically requiring learning Java, J2EE frameworks, databases, front-end development, software engineering, and other courses.

"Big Data Engineer" uses modern data warehouse technology, online analytical processing technology, data mining, and data visualization technology for data analysis to achieve commercial value. Therefore, in addition to mastering traditional database technology, big data engineers need to be familiar with the principles of distributed data storage, distributed computing, and data mining.

"C++ Development Engineer" mainly engages in C++ software programming on Windows or Linux platforms, primarily needing to master C++, Linux operating systems, network communication, and databases.

".NET Development Engineer" uses Microsoft's .NET to develop Web programs, Windows applications, and WAP wireless network applications. .NET development engineers mainly need to learn C#, databases, front-end development,

software engineering, and other courses.

"Web Front-end Development Engineer" is a relatively new profession, mainly engaged in website development, optimization, and improvement. A qualified Web front-end development engineer first needs to master various front-end development courses, and additionally needs to be familiar with traditional database knowledge, object-oriented software engineering knowledge, etc.

"Test Engineer": China's software testing profession is still in a developmental stage, with many medium and large software enterprises establishing separate testing departments that operate in parallel with development departments. As a test engineer, one needs to master main testing principles and tools, and also be familiar with mainstream operating systems and databases.

"Operations Engineer" is mainly responsible for maintaining and ensuring high availability of the entire service, while continuously optimizing system architecture to improve deployment efficiency and resource utilization. The biggest challenge facing operations engineers is large-scale cluster management issues, so operations engineers mainly need to master operating systems, network communication, and databases.

"Embedded Software Engineer" writes embedded systems. Embedded systems are application-centered, computer technology-based, and have software and hardware that can be tailored to meet strict requirements for functionality, reliability, cost, volume, and power consumption in application systems. Embedded software engineers mainly need to master C++, Linux, and network communication skills.

"IOS Development Engineer" mainly develops applications for portable terminals such as mobile phones based on the IOS system. This position mainly requires mastering C++, Linux, network communication, front-end development, and other courses.

"Algorithm Engineer" mainly researches robotics, speech recognition, image recognition, natural language processing, expert systems, etc., searching for knowledge hidden in massive data through algorithms. This position mainly requires mastering artificial intelligence principles and algorithms, data analysis, operating systems, and C++.

4.6 Curriculum–Knowledge Point Relationship Generation

Using the LDA topic model, the probabilistic relationships between the 11 curricula and knowledge points are obtained. The top 15 knowledge point words for each curriculum are selected to form the curriculum–knowledge point relationship, creating word clouds as shown in Figure 7 [Figure 7: see original paper].

From Figure 7, it can be seen that currently widely offered computer professional courses in universities need to pay attention to new market demands

and add new knowledge points. For example, early website content development for the "Web Development" curriculum was mainly static, focusing on images and text. With the development of Internet technology and the introduction of technologies and frameworks such as HTML5 and CSS3, modern web pages are more aesthetically pleasing and functionally powerful, so the curriculum needs to strengthen learning of these new technologies. "Java" is an object-oriented programming language, and J2EE is a Java platform designed for large-scale enterprise host-level computing types, simplifying application development and reducing programming requirements. Therefore, the curriculum needs to strengthen J2EE and related framework learning to meet enterprise needs. Due to the agility and code maintainability of enterprise system development, which requires involvement in some architectures, the "C#" curriculum needs to strengthen learning of software architecture and design patterns. In addition to learning the language syntax itself, the "C++" curriculum also needs to focus on its application and development on Linux and Unix systems. "Database" is the core part of various information systems such as management information systems, office automation systems, and decision support systems, and is an important technical means for scientific research and decision management. In recent years, with the rapid growth of data volume, distributed database technology has developed rapidly. Traditional relational databases have begun to develop from centralized models to distributed architectures. Non-relational databases represented by NoSQL and MongoDB have developed rapidly due to their high scalability and high concurrency advantages. In the teaching process, teachers need to closely monitor the development trends of these non-relational databases and introduce them. "Artificial Intelligence" is a branch of computer science that attempts to understand the essence of intelligence and produce intelligent machines that can react in ways similar to human intelligence. Current teaching and learning need to focus on recent hot enterprise applications such as robotics, speech recognition, image recognition, natural language processing, and expert systems.

With the rapid development of big data and the Internet, some emerging curricula have also appeared. The curriculum knowledge model also provides a basis for intelligence decision-making for the setting of syllabi and knowledge points for these new courses. For example, the "Data Analysis" curriculum organizes massive data accumulated over the years through online transaction processing of consulting systems using the unique data storage architecture of data warehouse theory, and systematically analyzes and organizes it through Spark, Hadoop big data cluster computing environments. Using various analysis methods such as data mining, it supports the creation of decision support systems to help decision-makers quickly and effectively analyze valuable information from massive data to facilitate decision-making and rapid response to external environmental changes, helping to build business intelligence. Data analysis, Spark, Hadoop, MapReduce, etc., are all knowledge points that need to be considered in this curriculum setting.

5 Conclusion

Current research mainly conducts manual analysis of skills and knowledge points required for positions in web recruitment texts without further utilizing web recruitment text information. To address existing problems in current web recruitment information analysis, this paper proposes a three-level curriculum knowledge model containing "position–curriculum–knowledge point" and achieves automatic construction of the curriculum knowledge model through natural language processing and text mining technologies. Finally, an empirical analysis of web recruitment texts for computer-related majors is conducted. The empirical results demonstrate the feasibility and effectiveness of the proposed model and construction process. Through analysis, the main professional skill demands of enterprises for these positions can be identified, providing guidance for universities in professional setting, teachers in syllabus and knowledge point setting, and students in career planning and knowledge point supplementation, thereby alleviating the dual contradictions of difficulty in finding jobs and difficulty in recruiting.

Due to the diversity of position names, such as "Java Development Engineer" and "Java Software Engineer" both representing the same meaning, current research methods mainly use main keywords like "Java" to standardize to the same position name. In future research, methods for position name standardization will be further optimized to automatically and accurately represent position information.

References

[1] CHAO C, SHIH S. Organizational and end-user information systems job market: an analysis of job types and skill requirements[J]. Inform Tech Learn Perform, 2005, 23(1): 1-15.

[2] WOWCZKO I. Skills and vacancy analysis with data mining techniques[J]. Informatics, 2015, 2(4): 31-49.

[3] KIM J Y, LEE C K. An empirical analysis of requirements for data scientists using online job postings[J]. International journal of software engineering and its application, 2016, 10(4): 161-168.

[4] MAURO D A, GRECO M, GRIMALDI M, et al. Beyond data scientists: a review of big data skills and job families[C]//Proceedings of the 2016 international forum on knowledge asset dynamics. Berlin: Springer International Publishing, 2016: 1844-1857.

[5] Lü Bin, ZHANG Tong, ZHOU Jue. Towards a universal intelligence profession and intelligence professionals for organizations—Analysis based on mining organizational recruitment webpage information (Part 1)[J]. Library and Information Service, 2009, 53(4): 19-23.

[6] LI Guoqiu, SANG Peiming. Intelligence process—the core of intelligence pro-

fession: problem domain and methodology—Analysis based on mining organizational recruitment webpage information (Part 2)[J]. Library and Information Service, 2009, 53(4): 24-27.

[7] XIA Huosong, PAN Xiaoting. Research on the relationship between big data academic research and talent demand based on Python mining[J]. Journal of Information Resources Management, 2017, 7(1): 4-11.

[8] HUANG Yue, WANG Kaifei, WANG Shanshan, et al. Investigation on data position recruitment requirements and implications for library and information discipline talent cultivation[J]. Library and Information Knowledge, 2016(6): 42-49.

[9] JIA Dongqin, TAN Bo. Analysis of foreign university library position requirements—Content analysis based on recruitment advertisements[J]. Library Development, 2018(2): 84-89.

[10] SODHI M, SON B. Content analysis of OR job advertisements to infer required skills[J]. The journal of the Operational Research Society, 2010, 9(1): 1315-1327.

[11] ZHAO M, JAVED F, JACOB F, et al. SKILL: a system for skill identification and normalization[C]//Proceedings of the twenty-seventh conference on innovative applications of artificial intelligence. Palo Alto: AAAI, 2015: 4012-4017.

[12] XU T, ZHU H, ZHU C, et al. Measuring the popularity of job skills in recruitment market: a multi-criteria approach[C]//Proceedings of the 32nd AAAI conference on artificial intelligence. Menlo Park: AAAI, 2018: 3013-3028.

[13] ZHAN Chuan. Analysis of professional talent skill requirements based on text mining—Taking e-commerce major as an example[J]. Library Tribune, 2017, 5(1): 116-123.

[14] XIA Lixin, CHU Lin, WANG Zhongyi, et al. Construction of employment knowledge demand relationship based on web text mining[J]. Library and Information Knowledge, 2016, 169(1): 94-100.

[15] BASTIAN M, HAYES M, VAUGHAN W, et al. LinkedIn skills: large-scale topic extraction and inference[C]//ACM conference on recommender systems. New York: ACM, 2014: 1-8.

[16] WANG Zhaoyi, XUE Chenjie, LIU Yulin. Analysis of e-commerce skill requirements based on proximity word analysis[J]. Journal of Information Resources Management, 2018, 11(2): 113-121.

[17] LIU Ruilun, YE Wenhao, GAO Ruiqing, et al. Research on text clustering of big data position requirements[J]. Data Analysis and Knowledge Discovery, 2017, 12(12): 32-40.

[18] ZHANG Junfeng, WEI Ruibin. Mining of data position talent demand characteristics in domestic recruitment websites[J]. Journal of Intelligence, 2018,

37(6): 176-182.

[19] WANG Dongbo, HU Haotian, ZHOU Xin, et al. Research on automatic extraction and analysis of data science recruitment entities based on deep learning[J]. Library and Information Service, 2018, 62(13): 64-72.

[20] WEI Lai, ZHENG Huamin. Comparative study on competency requirements for data librarians at home and abroad[J]. Library and Information Service, 2018, 62(10): 18-24.

[21] TIAN Ye. Investigation and analysis of domestic library and information professional demand status[J]. Library and Information Service, 2018, 62(9): 62-72.

[22] CHEN Yuanyuan, DONG Wei. Analysis of employment skills of library and information graduates under social demand orientation[J]. Library and Information Service, 2017, 61(19): 66-73.

[23] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. Journal of machine learning research, 2003, 3(1): 993-1022.

[24] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. PNAS, 2004, 101(1): 5228-5235.

Author Contributions

Yu Yan: Proposed research ideas, designed research plan, conducted data modeling, and wrote the paper.
Chen Lei: Conducted data collection and data cleaning.
Zhao Naixuan: Revised the paper.

---

of the Documentation and Information Center of the Chinese Academy of Sciences and the Department of Library, Information and Archives Management of the University of Chinese Academy of Sciences, *Library and Information Service* will launch a special issue on "Innovation and Capacity Building in Library, Information and Archives Management Education" in early September 2019 (Issue 18).

Submission themes are not limited to domestic or international library and information education, not limited to degree education levels, not limited to library and information education theories, methods, and experiences, and not limited to professional or general education courses. However, submissions must be original, innovative, and supported by your own research or practice.

Important Dates:
Intention submission deadline: April 15
Full paper completion deadline: June 1

Please indicate "Special Issue on Library and Information Education" in your submission.
Contact email: journal@mail.las.ac.cn
Submission website: www.lis.ac.cn

Department of Library, Information and Archives Management, University of Chinese Academy of Sciences
Graduate Education Office, Documentation and Information Center of the Chinese Academy of Sciences
Library and Information Service Magazine
February 26, 2019

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