Xu, Aiting; Wu, Yuchen; Meng, Feina; Xu, Shengying; Zhu, Yuhan

Article

Knowledge and skill sets for big data professions: Analysis of recruitment information based on the latent dirichlet allocation model

Amfiteatru Economic Journal

Provided in Cooperation with:
The Bucharest University of Economic Studies

Suggested Citation: Xu, Aiting; Wu, Yuchen; Meng, Feina; Xu, Shengying; Zhu, Yuhan (2022) : Knowledge and skill sets for big data professions: Analysis of recruitment information based on the latent dirichlet allocation model, Amfiteatru Economic Journal, ISSN 2247-9104, The Bucharest University of Economic Studies, Bucharest, Vol. 24, Iss. 60, pp. 464-484, https://doi.org/10.24818/EA/2022/60/464

This Version is available at:
https://hdl.handle.net/10419/281645

KNOWLEDGE AND SKILL SETS FOR BIG DATA PROFESSIONS: ANALYSIS OF RECRUITMENT INFORMATION BASED ON THE LATENT DIRICHLET ALLOCATION MODEL

Aiting Xu[1], Yuchen Wu[2], Feina Meng[3], Shengying Xu[4], and Yuhan Zhu[5*]

[1)2)3)4)5)] *Zhejiang Gongshang University, Zhejiang, China*

Please cite this article as:
Xu. A., Wu. Y., Meng, F., Xu. S., and Zhu, Y., 2022. Knowledge and Skill Sets for Big Data Professions: Analysis of Recruitment Information Based on The Latent Dirichlet Allocation Model. *Amfiteatru Economic*, 24(60), pp. 464-484.

DOI: 10.24818/EA/2022/60/464.

Article History
Received: 16 December 2021
Revised: 17 January 2022
Accepted: 9 March 2022

Abstract

Universities, enterprises, and students are the key subjects in the talent training system of Big Data. This study used text mining, interviews, questionnaires, and other methods to analyze the characteristics and deficiencies of Chinese universities in the training of Big Data talents, the requirements of enterprises on the professional quality of big data talents, and the cognition of students on the ability of big data talents. Furthermore, this study used the theory of plan-action-inspection-action cycle to evaluate the talent cultivation and quality management system of Big Data in China. The results showed that employees with rich professional background and proficiency in multiple programming languages are more favored by enterprises in recruitment. From the perspective of students, 83% of students hope that universities will implement the multi-disciplinary training model. From the perspective of talent training, more Chinese universities should open up the mode of interdisciplinary talent training. The results, based on pair-to-body comparisons, showed that students and businesses differ in skills. Secondly, most of the graduates think that the training of talents in universities needs to be improved. Furthermore, the training system of most universities is not suitable for the versatile talents that today's enterprises need.

Keywords: Big Data professions, training system, matching degree, PDCA, LDA

JEL Classification: J2, J4, M1

* Corresponding author, **Yuhan Zhu** – e-mail: yuhan.zhu@mail.zjgsu.edu.cn

Introduction

Data are regarded as valuable objects in today's world (Rahul and Banyal, 2020). Big Data has gradually become an excellent factor for improving the core competitiveness of enterprises, particularly with the increased use of streaming data. Big Data is a burgeoning subject, ranging across information management, social science, and computer science (Rahul and Banyal, 2020; De Mauro et al., 2018). Nowadays, data are being generated at an alarming rate in all walks of life. According to reliable estimates from the International Data Corporation, more than 2.5 TB of data are created per day (Hackenberger, 2019). Furthermore, the global data volume will unprecedentedly reach 163 ZB. Note that we are currently in the era of information explosion, with the continuous emergence of huge amounts of data. However, there is a significant benefit: these data are critical in the data mining filed.

Big Data is characterized by velocity, volume, variety, variability, veracity, visualization, value, and so on (Khaloufi et al., 2018). Velocity, volume, and variety are not only the significant characteristics that distinguish Big Data from 'normal' data but also the key attributes of Big Data. In addition, variability, veracity, visualization, and value are important features that show Big Data's complexity for the collection, disposition, analysis, and benefits of Big Data (Hackenberger, 2019; Khaloufi et al., 2018). Generally, Big Data is the oil and gold of the 21st century when mined properly (De Mauro et al., 2016). The emergence and application of Big Data have pervasively impacted countless aspects of everyday life, in both positive and negative ways (De Mauro et al., 2016). Big Data analysis plays an instructive part of utmost importance in the decision making and planning of many development fields, such as social media, smart cities, education, and e-commerce (Rahul and Banyal; 2020, Hilbert, 2016).

As an information asset, Big Data analysis can extract actionable knowledge from a large amount of available digital information to achieve breakthroughs and innovations in productivity, competitiveness, performance, and so on (Hilbert, 2016). However, many companies find it difficult to fully explore the value of Big Data because of the shortage of professionals and uncertainties about how to optimally deploy Big Data analysis patterns and rapidly translate them into business value (Persaud, 2020). Having large amounts of data and diversified analysis methods does not necessarily imply greater data value or deeper insights. Faced with this dilemma, managers are eagerly searching for data scientists or people with excellent business thinking and promising value judgments (Persaud, 2020). This is because only when data scientists or business experts have a deep understanding of a company's business model can they maximize their innovation or wisdom to complete Big Data projects (Persaud, 2020).

In addition, universities are paying attention to the large-scale application of Big Data (Hilbert, 2016). Currently, more than 40 universities in the United States offer master's programs in Big Data. Data analysis has become a hot spot in American universities (Hilbert, 2016). There are 30 universities in the UK offering Big Data majors, which also focus on the use of data business, data mining, and data analysis (Persaud, 2020). In the UK, the research on data application is more extensive and the basic research of data science is more diversified (Persaud, 2020). The construction of Big Data majors in the United States and the UK is of great significance to the construction of Big Data majors in Chinese universities. In the past 3 years, 24.0% of the records and approval results of undergraduate majors in Chinese universities are directly or indirectly related to Big Data (Hilbert, 2016), which indicates that Big Data majors are popular in China. However, McKinsey predicts that the

Big Data talent gap will reach 1.5 million in the next 3–5 years. This is mainly due to the high barriers to entry in the Big Data industry, requiring both the ability to understand and use large datasets and expertise in statistics and machine learning. Therefore, enterprises need Big Data talent to rapidly determine the career requirements of Big Data work because the existing skills and responsibilities describing Big Data careers are often vague (Khaloufi et al., 2018), which is not conducive to universities and students who have a specific training direction and goal. On this basis, as the training subject of Big Data talents, colleges and universities should adjust and optimize the training scheme according to the industry demand in a planned way. Students with Big Data career aspirations should also strive to adapt to the changing industry.

1. Literature Review

1.1. The demand side

The growth of the Internet and the subsequent mass popularization of the World Wide Web have created a business that handles more than 2.5 TB of data per day. Big Data has penetrated every industry, from banks and Internet companies to the government (De Mauro et al., 2018). Data are now equivalent to oil and have strategic value. Enterprises place a high value on information and strategic resources. The desire to exploit Big Data insights is affecting the job market, forcing companies to rethink their human resource needs. Emerging Big Data professionals are rapidly making their way onto the list of Human Resources (HR) job openings. However, according to the McKinsey's Big Data talent reports, the gap of Big Data talents will reach 1.5 million in the next 3–5 years. There is a severe shortage of Big Data talents, and the mismatch between talent supply and demand will continue to rise in the future. Therefore, many scholars have conducted a plethora of research on Big Data talents.

First, from the research perspective, the recruitment information of Big Data jobs obtained from text mining is often used to determine the knowledge fields, skills, and tools required for Big Data software engineering and development jobs to help assess the knowledge level of Big Data professionals and technical personnel (Gurcan and Cagiltay, 2019). This also helps to improve the efficiency of a company's personnel recruitment and provides them with reliable talent evaluation criteria. In addition, some scholars have designed a vocational matching degree test system according to the specific needs of the Big Data industry for talents, which can directly reflect the matching degree of the actual ability of college students and industry requirements and further reveal the difference between higher education results and the needs of the Big Data industry (De Mauro et al., 2018). On this basis, some scholars have thoroughly analyzed the recruitment requirements of different positions in the Big Data industry and then designed a job classification system applicable to job seekers with different skills, providing certain references for their career planning (Lee and Choi, 2019).

Second, from the viewpoint of the research subject, enterprises as the demand side are the main research subject in the existing research. Text mining and analysis are conducted on recruitment information displayed on a company's website to acquire the knowledge and skills required for each job position. Meanwhile, senior executives are interviewed by companies to obtain more detailed talent screening criteria (Persaud, 2020; Fulthorp and D'Eloia, 2015). Fresh graduates, the main source of Big Data talents, are another important research subject. Some researchers also start with students and enterprises; they send questionnaires to college students, asking them to evaluate their employability and skill

levels, compare students' self-evaluation with enterprise evaluation, and determine whether students' cognition of their job-seeking ability is biased (Lee et al., 1995).

Third, from the perspective of research methods, probabilistic topic modeling technology can be used to explain the knowledge fields and tools that need to be mastered for Big Data positions and to classify job demands for further research (Gurcan and Cagiltay, 2019). Some scholars used CPSP to classify and define entries with unclear definitions and then refined classification. The field experts were then consulted on whether the word aggregation results obtained at each stage of the CPSP were reasonable. If the word aggregation or definition was not reasonable, it was subdivided again. Finally, a knowledge and skill model related to the Big Data industry was obtained (Persaud, 2020). Some scholars used the traditional questionnaire survey method to obtain data and analyzed the needs of students and enterprise subjects with statistical knowledge to obtain the cognitive differences between the two subjects in the ability of Big Data talent.

On the basis of this, we can see that most of the existing studies begin with the main body of the enterprise and combine the evaluation and opinions of the students. Text mining of recruitment information to identify the knowledge and skills required for Big Data positions, which can not only help enterprises conduct more efficient recruitment activities but also help students identify their deficiencies in time. However, text mining of recruitment data alone is insufficient for obtaining comprehensive information. To compensate for this deficiency, our research will adopt two forms of survey methods, namely, text mining and questionnaire and interview, on the demand side of "enterprises" for Big Data talents and use a multitopic modeling analysis method for obtaining job characteristics to obtain the requirements of enterprises for Big Data talents in a more three-dimensional and comprehensive manner.

1.2. The supply side

1.2.1. Universities

Concerns have been raised by educators regarding the knowledge and skills that are required for information technology (IT) professionals to function effectively in dynamic environments, as well as how the university curriculum must be revised to meet the changing needs of the profession (Nelson, 1991; Yaffe, 1989). Therefore, in addition to job postings, educational programs have been used to assess the fit between academic preparation and industry expectations.

From the perspective of research purposes, the results of comparing the curriculum content covered by education programs with the needs of the industry are usually used to optimize the talent training programs of universities to improve the matching between majors and job positions (Gasmi and Bouras, 2017; Persaud, 2020). From the viewpoint of the research subjects, most of the research subjects are mainly university educators, who evaluate the rationality of the university curriculum from their perspective (Lee et al., 1995). Simultaneously, part of the literature also investigated the satisfaction of students with the courses of a specific single major (Das and Ara, 2014; Mourshed et al., 2013). In terms of research methods, some scholars identify and discuss key issues facing the university curriculum through open forums (Lee et al., 1995). Given the subjectivity of forum comments, some scholars made some improvements by sending questionnaires to students

and educators to obtain quantitative data and judging the relevance of professional degrees in getting a job and how to improve skills training courses (Das and Ara, 2014).

We can see that most of existing studies evaluate the curriculum setting of a single major in universities from the perspective of students or educators. Based on this, we made a horizontal comparison of the proportion of various courses in training programs of different majors in universities and sent questionnaires to graduates to understand whether the curriculum setting in universities is reasonable from the perspective of the course audience.

1.2.2. Students

As Big Data reserve talents, students' voice is also worthy of attention, but there are few research studies on individual students regarding this research topic at present. Some scholars assessed educational needs by issuing questionnaires and analyzed them from the perspectives of IT students and end-user personnel. Their study found a difference in skill recognition between the two (Nelson, 1991). In most studies, the student subject is usually used as a secondary subject for comparative analysis with one of the two subjects: universities or enterprises (Manju Das and Velmurugan, 2019; Lee et al., 1995). Some studies compared students' self-evaluation to enterprise evaluation and discovered that students' cognition of their skills was biased (De Mauro et al., 2018). Studies evaluating school curricula from the perspective of students reflect a problem: school curricula do not always adequately prepare students for the job market, and some aspects of the curriculum are outdated by the time they begin working.

Based on this, we issue a questionnaire to college graduates to obtain their satisfaction with the college curriculum plan and provide a basis for the adjustment of the college training plan. Furthermore, we crawled forums to obtain datasets of the college students' cognitive status of Big Data vocational skills and compared them with the needs of enterprises.

1.3. The third-party relationships

Enterprises are the main demand. The highest demand for high-quality graduate students comes from enterprises. Therefore, the recruitment requirements of enterprises for data science and Big Data positions are also the development goals and training guidance of Big Data talents (Gurcan and Cagiltay, 2019). Enterprises play a proper and significant role in a data science and Big Data talent cultivation and quality management system. In this system, universities play the role of "producer." They take social demand as the guide and set up corresponding training programs to cultivate high-quality students who meet job market demands (Das and Ara, 2014; Gasmi and Bouras, 2017). Graduate students are the "main force" of high-quality Big Data talents in the eyes of enterprises, and they are also the "barometer" for evaluating educational outcomes in the eyes of universities (Lee and Choi, 2019). In this system, companies communicate their needs to universities and students through job advertisements, universities adjust their curriculum, and students adjust their skill sets accordingly.

In the data science and Big Data talent cultivation and quality management system, enterprises, universities, and students cannot be separated from each other. However, few studies have taken these three subjects as research objects (Manju Das and Velmurugan,

2019). In addition, existing studies have found that universities' perception of the importance of skills is often different from that of enterprises, students are dissatisfied with universities training programs, and there is also a "skills gap" between the practical skills of applicants and employers' expectations of them (Mourshed et al., 2013; Metilda and Neena, 2016). There is a lack of overlap between students' field of education and their assigned job placement (Arpat et al., 2021). Therefore, businesses, students, and universities should be asked to assess their perceptions of the personal qualities and skills/competencies required in workplaces. Students' evaluation of courses should also be considered, to help universities establish a complete talent cultivation plan, comprehensively considering students' aspirations. Only in this manner can we fully consider the current situation, the transmission of requirements and location information, and the resulting practical problems. On this basis, we can establish a sustainable operation of the Big Data talent training and delivery system (Mehrolia and Alagarsamy, 2019).

This study will begin with three subjects: enterprises, universities, and students. First, we use job advertisements to create skill demand profiles of employers. Second, the rationality of the setting is judged on the basis of the cultivation schemes of different universities, and a specific judgment is made on the basis of the requirements of enterprises and students' evaluation results of colleges and universities. Finally, we combine three lines of personnel training and delivery that frequently operate in parallel despite being intricately linked: enterprises, students, and universities.

2. Research methodology

The talent quality management and promotion system, which is based on "university–student–enterprise," follows the plan–do–check–action cycle theory to manage talent quality. Quality management is divided into four stages: plan, do, check, and action. The highest demand for high-quality graduate students comes from enterprises. Therefore, the recruitment requirements of enterprises for data science and Big Data positions are also the development goals and training guidance of Big Data talents. Universities play the role of "producer." They take social demand as the guide to cultivate high-quality talents who meet market demands. Graduate students are the "main force" of high-quality Big Data talents in the eyes of enterprises, and they are also the "barometer" for evaluating educational outcomes in the eyes of universities. From the perspective of "enterprises-graduate students," enterprises have gradually improved educational requirements for Big Data talents, so Big Data talents with a master's degree or higher degrees have become the preferred choice of enterprises with the advance of the Big Data era. These students have greater potential for future career development and higher salary. From the perspective of "university graduate students," graduate students are direct participants of professional training programs in universities and have the right of discourse to evaluate the effectiveness and scientific qualities of professional training programs in universities.

In this study, we investigated our research problem using three different datasets, each of which provided unique insights. The first dataset is targeted at enterprises; the second dataset is aimed at universities; and the third dataset includes the content of questionnaire surveys, interviews, and forum comments of university students. Each dataset is described in Appendix in more detail.

Recruitment advertisement and student comment datasets were analyzed through text mining. Word segmentation is the basis of information processing. "Jieba" library, a third-party library in Python, was used to implement word segmentation. On the basis of word segmentation, word frequency statistics are conducted, and the results are presented visually and comprehensively through a word cloud graph. However, analysis using the word cloud is shallow. In this study, linear discriminant analysis (LDA)-based topic modeling was performed to reveal the hot demand topic in Big Data positions. Then, words in the topic were used to analyze specific requirements for knowledge and skills. LDA can provide an efficient way to process large sets of documents by discovering abstract topics in documents and mining information hidden behind the semantics. This method has been widely used in the field of topic exploration. In the implementation of LDA-based topic modeling, topics and corresponding words were generated using the prior knowledge of the Dirichlet distribution.

In this study, we used Gensim, which is a third-party library in Python that implements LDA-based topic modeling. Another third-party library, called pyLDAvis, was then used to achieve model visualization. Many studies have proved the difficulty in the application of an LDA model in determining the optimal number of topics because this parameter is artificially specified rather than automatically generated by the model. Therefore, different values ranging from 2 to 5 were tested, with reference to several research studies. Furthermore, perplexity, which is often used to measure topic modeling, was taken into consideration to determine the appropriate number of topics. However, the number of topics is usually large when it is solely determined using the perplexity of a model. In addition, a problem can be easily created in which high similarity between topics affects topic identification. Accordingly, when determining the number of topics, the distribution of topics and the distribution of keywords within these topics should be considered to avoid overlapping topics. After thorough consideration of topic distributions and the perplexity of the model, the number of topics was finally set at 4.

3. Results

3.1. Enterprises: Demanders of High-quality Talents

3.1.1. Educational Background

Educational background is the most fundamental principle for enterprises to select data science and Big Data talents. As shown in Figure no. 1, enterprises prefer data science and Big Data talents with postgraduate qualifications. The professional backgrounds of the job seekers present significant characteristics of interdisciplinary talents. Note that enterprises prefer candidates with backgrounds in computer science, statistics, mathematics, and finance for data science and Big Data positions.

Considering the academic careers and professions provided in the word cloud (Figure no. 1), we identified the eight most in-demand combinations. As shown in Table no. S1, the data science and Big Data industry require more highly educated talents who possess in-depth systematic theoretical knowledge and technical ability. From a professional perspective, 32.4% of positions in data science and Big Data require applicants with statistical knowledge or background knowledge in these disciplines, 34.9% of these positions require applicants to have a professional background in computer science and technology. The frequency of keywords in "education background" is listed in Table no. S1.

Figure no. 1. The word cloud of educational background

3.1.2. Professional Knowledge

The depth and breadth of professional knowledge are the core index and essential reference used by enterprises to investigate data science and Big Data talents. As shown in Figure no. 2, enterprises eagerly expect more versatile and professional talents with Big Data knowledge and mathematical knowledge of Big Data analysis, Big Data mining, and optimization method to join their teams, which indicates that enterprises place a high value on business strategies and professional quality of employees and closely focus on the needs of the business frontier to find the most suitable candidates. On this basis, the keywords frequency of "professional knowledge" are counted, and the employer's preference for professional knowledge is found by comparison. Refer to Table no. S2 for details.

Figure no. 2. The word cloud of professional knowledge

3.1.3. Programming Ability

Programming ability is a necessary skill for data science and big data talents. It is also the "hard power" of talents that is subject to scrutiny by enterprises. Figure no. 3 indicates that enterprises strongly believe that data science and big data talents should be proficient and hold rich practical experiences in Python, SQL, Hadoop, and other common software. In turn, we have a frequency table that lists the programming preferred by employers, to identify the combinations of tools and technologies in high demand. Refer to Table no. S3 in the Appendix for more details.

Figure no. 3. Word cloud of programming ability

3.1.4. Work Experience and Professional Quality

Work experience and professional quality are important references for enterprises in evaluating the practical experience of data science and big data talents and a plus for the background qualification of candidates. Figure no. 4 illustrates those enterprises favor candidates with a rich internship experience, a professional sense of responsibility, and professionalism. In turn, we have a frequency table that lists the work experience preferred by the employers. Specifically, 23.6%, 10.2%, and 20.1% of positions require candidates with at least one, three or more, and two or more internships, respectively. Refer to Table no. S4 in the Appendix.

Figure no. 4. Word cloud of work experience and professional quality

3.2. Universities: Cultivators of High-quality Talents

The curriculum setting of universities, which is oriented by employment demand, attaches great importance to the improvement of practical ability and the cultivation of accomplishment. Moreover, one of the common goals of curriculum setting is to ensure the supply of high-quality talents for enterprises. Figure no. 5 depicts that different universities offer various specialized main courses. For example, big data analysis and statistics, big data mining, and machine learning are extremely common. In addition, universities pay increased

attention to the improvement of students' application ability and accumulation of practical experience in projects. All universities provide a specialty practice and professional quality training. In this manner, students can understand and improve their ability to solve practical problems independently through specialty practice. The cultivation of professional quality is helpful for students in transforming their professional roles.

Figure no. 5. Word Cloud of the main course

Furthermore, the proportion of mathematics in universities is relatively low in terms of the types of curriculum. Many universities continue to focus on the cultivation mode of talents within a single discipline, whereas only a few universities adopt the cultivation mode of interdisciplinary talents. The analysis of direct comparison of curricula is challenging. Therefore, the curriculum was divided into three courses, namely mathematics, traditional computer, and data science. The study then conducted a comparative analysis of the curriculum setting on this basis. Tables nos. S5 and S6 indicate that the proportion of mathematics in the curriculum setting is low. The fundamental reason underlying this result is that graduate education differs from undergraduate education. Moreover, most universities take a high mathematics level as a prerequisite for graduate admission. However, improving the particular abilities of students, such as data analysis and data development in graduate education, is the objective of all schools.

Therefore, the credit setting for mathematics in universities is generally less than those for data science and the traditional computer studies. The study also noted an imbalance between data science and traditional computers. The same directions for talent cultivation in computer science and technology and applied statistics were observed for big data. Traditional computer displayed a high proportion in the curriculum setting for big data in computer science and technology. The curriculum of big data in applied statistics focuses more on data science. As a result, the talent training mode of a single discipline remains as the mainstream mode in the Chinese context. Among the 16 universities, only the Harbin Institute of Technology and Nanjing University balance the curriculum setting for traditional computer and data science. Taking Nanjing University as an example, its goal for personnel training is to cultivate data engineers who are capable of building big data systems. As a result, Nanjing University set the proportion of computer and data science courses at 50.00% and 44.44%.

The direction of personnel training in universities can be categorized into elite talents in specific fields and available talents in multiple fields. As shown in Table no. S7, Peking

University offered Security Analysis, Epidemiological Statistics, Actuarial, and other related practical courses in the financial industry. The vertical and permeable teaching of Peking University continuously provides enterprises with elite talents in the financial field. The specialty teaching at Nanjing University is characterized by the horizontal extension of the subjects. Examples of these subjects are Human Resources and Social Security Statistics, Stock Investment and Fund Analysis, and other courses, which are designed to cultivate general talents with mastery of knowledge and skills in various fields.

3.3. Graduate Students: Main Force of High-quality Talents

3.3.1. Professional Skills Required for Employment

In general, graduate students majoring in data science and big data believe that professional skills, such as data management, data analysis, and data acquisition are essential skills for employment. At the same time, they believe that internship experience is the most critical requirement of the employment process followed by programming ability and professional knowledge. Educational background ranks last. Figure no. 6 indicates that graduate students believe that knowledge and skills related to big data are the most critical stepping stones for securing employment. They realize the necessity of acquiring understanding of data acquisition, processing, analysis, management, and other related fields. Moreover, they acknowledge that conducting in-depth research in a specific field is crucial for the continuous improvement of their business ability and for the receipt of excellent development opportunities in the future. During in-depth interviews with the graduate students, the study found that the majority agree with the statement that enterprises favor a master's degree more than they do a bachelor's degree. As a result, they perceive themselves as slightly ahead of their undergraduate counterparts in educational background and focus on internships, programming skills, and professional knowledge.

Figure no. 6. Word cloud for professional knowledge from the students' perspective

3.3.2. Training Programs in Universities

The students perceive the multi-disciplinary personnel training mode as more suitable for their development. Under this model, the proportions of mathematics, data science, and the traditional computer studies should be balanced. The era of big data is a new one, which requires new ideas and theories. Therefore, only interdisciplinary training can meet the new requirements of enterprises. Given that the multi-disciplinary personnel training mode is conducive for cultivating a flexible and adaptive cognition process, 83% of the students express support for this model, whereas only 17% reported otherwise. Therefore, offering various types of course in a balanced manner is an essential requirement of the mode. Figure no. 7 indicates that the graduate students surveyed set the proportion of mathematics, data science, and traditional computers at 23%, 39%, and 38%, respectively. Such a setting ratio can improve students' overall quality and adaptability to the labor market. Based on a certain level of mathematics, the setting for basic mathematics can be reduced appropriately, whereas the setting for advanced mathematics, such as advanced probability theory, can be increased.

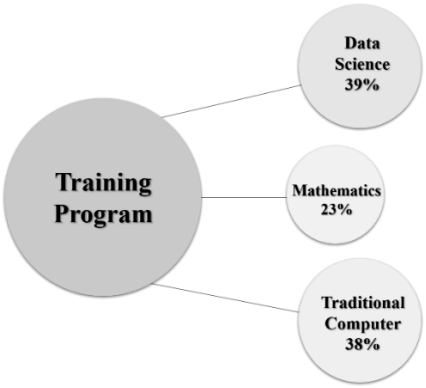

Figure no. 7. Proportion of the courses

3.3.3. Internship Experience

According to Figure no. 8, many students actively seek internship opportunities to improve their practical ability. However, only a few students obtain more than three times of internship experience. In the inherent cognition of students, the project practice in the academic competition is relatively different from the actual project practice in enterprises. Thus, a rich internship experience can help students better understand the project process and lay a solid foundation for them in completing future projects independently. To improve their ability to deal with practical problems, many students approach enterprises for internships in their spare time. A total of 14% of the students reported having no internship experience. In comparison, 49% obtained only one internship experience for various reasons, such as unreasonable courses offered by various universities. Lastly, students with two or three internships comprised 36% of the study population. Only 1% of the students obtained three or more internships. Although students consider the internship experience a top priority, very few students have a rich internship experience.

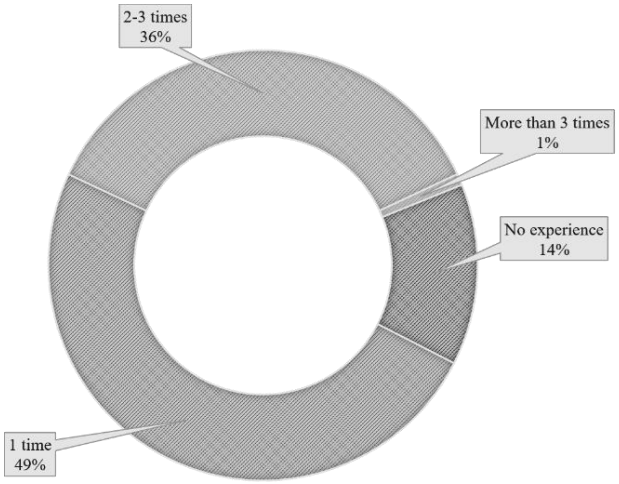

Figure no. 8. Distribution of internship experience

3.4. Mismatch between Enterprises, Graduate Students, and Universities

Without a doubt, enterprises, universities, and graduate students formulate their plans and implement them to the letter. For example, enterprises establish criteria for hiring talent in the field of big data. These criteria can be used to select the best to be included in the company. Universities conduct different talent training programs. As such, a course is formulated according to the corresponding training programs to realize the high efficiency of talent training. Moreover, individuals have different goals. For example, students aim to become data engineers, whereas others want to be data scientists. Thus, they make plans according to their own goals and realize their development according to the plan.

"Check and action" will adjust the plans of companies, universities, and graduate students accordingly. First, the enterprises will provide feedback about their specific requirements for talents to universities. According to the requirements to improve the matched degree of talents and demands, universities will adjust their training programs and courses to send more qualified talents to enterprises. At the same time, based on their acceptance of the courses arranged by universities, graduate students can improve their abilities according to the requirements. However, the study found that the assumptions of talent demand and actual talent supply in this program did not match successfully.

3.4.1. Mismatch Between Enterprises and Graduate Students

There is a gap between the expectations of the companies and the actual experience of graduate students. Many enterprises prefer to recruit graduate students with work experience. On the one hand, individuals who work in data mining and algorithm optimization are required to accumulate experience in the actual combat of projects. On the other hand, experience can reduce the cost of training. Such experienced young people are energetic, have a strong ability to accept new concepts, and can meet the changing needs of the industry. In real life, the majority of graduate students lack less internship experience, which may cause

problems in processing data of varying magnitudes or complexities. As a result, enterprises need to spend a long time training those who lack project experience. If the employee continues to fail to meet the requirements of enterprises after training, then certain damages will be incurred to enterprises and graduate students. In summary, failing to recruit satisfactory employees and finding suitable jobs are extremely easy for enterprises and graduate students, respectively, which then influence the flow of big data.

3.4.2. Mismatch Between Enterprises and Universities

Several differences are observed between enterprises and universities in terms of the evaluation standards for outstanding big data talents. During the recruitment process, enterprises pay more attention to the candidates' professional qualities and understanding about the business. Enterprises prefer candidates who are proficient in one or more programming software. However, the types of programming software are not limited to those that are commonly used in the industry. On the contrary, universities continue to attach more importance to students' achievements in professional courses. Although they may set a large proportion of credits in practice, they continue to lack a scientific evaluation system. Furthermore, universities tend to overlook the balanced development of professional knowledge and software skills of students to some extent when designing a training scheme for postgraduates. Meanwhile, the required software is relatively single and outdated, which fails to cope with the emerging trend of industry development. Thus, this setting is not conducive to the formation of the core competitiveness of big data talents.

3.4.3. Mismatch Between Universities and Graduate Students

Moreover, a difference between the direction of talent cultivation in universities and the exception of post-graduates is notable. Among the abovementioned universities, many continue to focus on the cultivation systems of single-discipline talents, whereas only a few institutions lead the initiation of the cultivation systems of multi-disciplinary talents. However, many graduate students expect to become all-rounders with knowledge and skills in various fields. Graduate students interviewed suggested that mathematics, data science, and programming should be allocated in a balanced manner in proportions of 23%, 39%, and 38%, respectively. In the IT industry, it is essential to acquire and develop skills centered on technical skills (Tokarčíková et al., 2020). Based on this allocation, students hope that the universities could add some highly efficient strategies of instruction and other industry-related courses which are high quality into the cultivating system to enable the students to deeply study their industry fields of interest and realize the perfect combination of technology and business thinking (Rosca et al., 2008).

Conclusions

The study investigates the abilities required to be competent in big data-related work, the degree of perfection of postgraduate training systems in higher education institutions, the relevant ability that students should possess and the interaction among these factors. The study innovatively starts with three unique research subjects and uses various methods, such

as text mining, interview, and questionnaire, to deeply explore information. It also uses cycle theory to evaluate the cultivation and quality management of big data systems among talents.

The study found that companies require professional knowledge and programming skills for recruitment. Furthermore, applicants must obtain the appropriate educational background (i.e., level of education and professional background), work experience (i.e., internship), and good professionalism (i.e., interpersonal skills and personal charm). Although education, work experience, and professionalism are considered important, employers place an emphasis on expertise and programming skills. In essence, the study observed that employers are interested in the following qualities of employees: multidisciplinary complex backgrounds, proficiency in one or two subjects (depth of experience and knowledge), and wide understanding of other subjects (breadth of experience and knowledge). In other words, talents proficient in big data knowledge display rich business knowledge and fast business understanding, whereas employees proficient in multiple big data processing software and programming languages can be integrated into different process lines of big data business development at any time during employment. For example, those who are proficient in SQL+Python+Hadoop are the most popular candidates for business executives at present.

Moreover, the study observed that the 16 Chinese universities ranked top in big data-related majors, such as Peking University, Tsinghua University, and Renmin University, who provided good training programs for students. However, at present, the majority of universities continue to focus on the cultivation mode of single-discipline talents, whereas only a few universities lead in developing the cultivation mode of multi-disciplinary talents. The main reason underlying this finding is that the talent training programs of big data-related majors are still in the exploration stage and require continuous improvement in combination with the cutting-edge demands of the industry. Trainees (students) believe that professional knowledge and internship experience are the trump cards for winning in the fierce competition. At the same time, 83% of students expect universities to promote a multidisciplinary talent training model. The result also demonstrates that the interdisciplinary training model for talents constitutes current trends.

By comparing the cognitive results of the three subjects, the study found that the vast majority of graduates consider that the rationality of training programs and curricula in colleges and universities require further improvement. From their perspective, universities should formulate a curriculum that balances traditional computers and big data. In terms of measuring the importance of the qualities required of big data talent, students and companies report conflicting results. This tendency reflects the cognitive gap between the expectation and reality of big data talent, as well as the fact that students lack a full and accurate interpretation of the employment needs of enterprises. In addition, single-field elites under the mainstream training mode of colleges and universities are inconsistent with the compound talents required by enterprises. These scenarios reflect the same problem, that is, there is a mismatch between information transmission and demand implementation at each link in the big data talent training and promotion system. The results provide a factual basis for the three parties to understand the current situation to serve as a reference for the parties in improving their deficiencies, further adjusting the training and promotion systems for big data talents, and improving the matching degree of big data positions.

In addition, the selection of research objects can be improved. Future studies should consider expanding the number of universities and the industry scope of the interviewed elites. In this manner, the generalizability and significance of the research results can be significantly

improved. At the same time, further studies should comprehensively analyze the content of recruitment advertisements and regularly update the list of skills that big data talents should possess according to the results of analysis to provide a basis for colleges and universities in adjusting curriculum settings in a timely manner. Students can then acquire corresponding skills to meet the needs of the industry.

Acknowledgments

This paper is supported by the Key Program of National Philosophy and Social Science Foundation of China (NO. 17ATJ001) and "Ten Thousand Talents Program" Youth Top Talent Project in Zhejiang Province (NO. ZJWR0108041). This work also supported by the characteristic & preponderant discipline of key construction universities in Zhejiang province (Zhejiang Gongshang University- Statistics) and Collaborative Innovation Center of Statistical Data Engineering Technology & Application.

References

Das, K. K. & Ara, A. 2014. Mismatch of Skills from Education to Employment: A Case Study. *IFRSA Business Review*, 4(3).

De Mauro, A., Greco, M. & Grimaldi, M. 2016. A formal definition of Big Data based on its essential features. *Library Review*, 65(3), pp.122-135. https://doi.org/10.1108/LR-06-2015-0061

De Mauro, A., Greco, M., Grimaldi, M. & Ritala, P. 2018. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, 54(5), pp.807-817. https://doi.org/10.1016/j.ipm.2017.05.004

Fulthorp, K. & D'Eloia, M. H. 2015. Managers' perceptions of entry-level job competencies when making hiring decisions for municipal recreation agencies. *Journal of Park and Recreation Administration*, 33(1), pp.57-71. https://www.proquest.com/scholarly-journals/managers-perceptions-entry-level-job-competencies/docview/1730049088/se-2?accountid=50247

Gasmi, H. & Bouras, A. 2017. Ontology-based education/industry collaboration system. *IEEE Access*, 6, pp.1362-1371. https://doi.org/10.1109/ACCESS.2017.2778879

Gurcan, F. & Cagiltay, N. E. 2019. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access*, 7, pp.82541-82552. https://doi.org/10.1109/ACCESS.2019.2924075

Hackenberger, B. K. 2019. Data by data, Big Data. *Croatian Medical Journal*, 60(3), pp.290-292. https://doi.org/10.3325/cmj.2019.60.290

Hilbert, M. 2016. Big data for development: A review of promises and challenges. *Development Policy Review*, 34, pp.135-174. https://doi.org/10.1111/dpr.12142

Khaloufi, H., Abouelmehdi, K., Beni-Hssane, A. & Saadi, M. 2018. Security model for big healthcare data lifecycle. *Procedia Computer Science*, 141, pp.294-301. https://doi.org/10.1016/j.procs.2018.10.199

Lee, D. M., Trauth, E. M. & Farwell, D. 1995. Critical skills and knowledge requirements of IS professionals: a joint academic/industry investigation. *MIS Quarterly*, pp.313-340. https://doi.org/10.2307/249598

Lee, H. & Choi, J. 2019. IT Jobs in the Era of Digital Transformation: Big Data Analytics. *Asia Pacific Journal of Information Systems*, 29, pp.717-730. https://www.earticle.net/Article/A367254

Manju Das, S. K. & Velmurugan, R. 2019. Skill for employability among MBA students. The gap between skill and employability-a study conducted among MBA students and employers. Journal of Advanced Research in Dynamical and Control Systems, 11(7), pp.1-7.

Mehrolia, S. & Alagarsamy, S. 2019. Perceptual Gap Among Corporate World, Academics and Students: Personal Qualities and Employability Competencies Of Students. *MOJEM: Malaysian Online Journal of Educational Management*, 8, pp.1-17. http://mjs.um.edu.my/ index.php/ MOJEM/article/view/21360

Mourshed, M., Farrell, D. & Barton, D. 2013. *Education to employment: Designing a system that works.* Mckinsey Center for Government. https://celt.li.kmutt.ac.th/mock/km/wp-content/uploads/ 2018/03/education-to-employment_final.pdf

Nelson, R. R. 1991. Educational needs as perceived by IS and end-user personnel: A survey of knowledge and skill requirements. *MIS Quarterly*, pp.503-525. https://doi.org/10.2307/249454

Metilda, R. M., and Neena, P.C., 2016. Gap Analysis of Employability Skills of Entry Level Business Graduates Based on Job-Fit Theory. *International Journal of Social Sciences and Management*, 3(4), pp.294-299. https://doi.org/10.3126/ijssm.v3i4.15973

Persaud, A. 2020. Key competencies for big data analytics professions: a multimethod study. *Information Technology & People*, 34(1), pp.178-203. https://doi.org/10.1108/ITP-06-2019-0290

Rahul, K. & Banyal, R. K. 2020. Data Life Cycle Management in Big Data Analytics. *Procedia Computer Science*, 173, pp.364-371. https://doi.org/10.1016/j.procs.2020.06.042

Rosca, I. G., Petrescu, V., Sârbu, R., Tomosoiu, N., Ilie, A. G. & Bucur, R. C. 2008. Quality assurance systems in higher education. *Amfiteatru Economic*, Special Issue, pp.6-12. http://www.amfiteatrueconomic.ase.ro/arhiva/pdf/special2008/revista_fulltext.pdf

Yaffe, J. 1989. MIS education: a 20th century disaster. *Journal of Systems Management*, 40(4), article no. 10. https://www.proquest.com/scholarly-journals/mis-education-20th-century-disaster/docview/ 199814018/se-2?accountid=50247

Appendix

The introduction of the dataset

The first dataset is targeted at enterprises and consists of recruitment information posted by enterprises on websites as well as information gathered from face-to-face interviews. Job advertisements are widely used in the study of job skills because these advertisements contain rich information such as job title, salary, description, company information, qualifications, and responsibilities. For this study, 51jobs, a large-scale job posting website, with search and filter options, was selected as a data source. The website was selected because it contains a huge amount of information, and the companies it covers are well known. We found that Big Data developers, software engineers, architects, consultants, analysts, experts, and so on were classified as Big Data profession when using query and filter options to obtain job posts matching the exact phrase Big Data within a title. A specific dataset was created by retrieving the job title and description for each job posting. The information in the dataset is collected by crawling the website. In this study, Python was used to crawl job recruitment information related to data science and Big Data on the 51jobs website from September 1, 2020, to September 7, 2020. A total of 8,248 enterprise-related data points were crawled, and 9,652 non-repetitive job posting data points were obtained, considering that job advertisements do not fully reflect interviewers' views on Big Data positions. For additional profiles, we interviewed 31 interviewers from different companies. The interview involved various enterprises from banking, insurance, retail, and Internet industries. Each interview lasted

20-40 min, and the interview content was recorded in real time. The interview content includes educational background, professional ability, programming foundation, and professional quality that employees in Big Data positions should have.

The second dataset is aimed at universities. With the advent of the Big Data era, the job market has put forward higher requirements for talent in Big Data. The demand for interdisciplinary talents with high-quality is increasing every day. According to data released by the Boss Zhipin Research Institute, 49.5% of enterprises set educational requirements for Big Data professions as a master's degree or above. More than 80% of enterprises require applicants to major in computer science and technology or applied statistics. To explore whether personnel training meets the expectations of candidates, this study focuses on the analysis and comparison of the postgraduate curriculum of 16 famous universities, including Peking University, Tsinghua University, Zhejiang University, National University of Defense Technology, and Renmin University of China, which rank in the top 5% in terms of these majors. Because of the great differences in the page structure of the official websites of various universities, web crawler technology cannot be used for data collection. Therefore, data are collected through manual search and download.

The third dataset is composed of questionnaire surveys, interviews, and campus forum comments of university students to have a more comprehensive understanding of students' opinions and suggestions on current training programs and the job market. On the one hand, we selected 160 recently graduated students from the 16 universities mentioned above who have worked for more than half a year and conducted questionnaire surveys and in-depth interviews on the aspects of training and employment satisfaction. The reasons for selecting these students are as follows: They still have relatively vivid memories of the learning situation and the training situation at school and have a relatively vivid understanding of the perception and satisfaction of graduate job hunting, which can ensure the reliability and representativeness of the data. On the other hand, we analyzed comments on professional training programs and employment status in campus forums (BBS) of 16 universities. After eliminating duplicates from the collected comments, a total of 2009 comments were obtained.

Table no. S1. The frequency of keywords in "education background"

Education background	**Rate %**
Statistics, Postgraduate	26.30%
Computer Science and Technology, Postgraduate	25.10%
Mathematics, Postgraduate	11.10%
Finance, Postgraduate	9.90%
Computer Science and Technology, Undergraduate	9.80%
Statistics, Undergraduate	6.10%
Mathematics, Undergraduate	5.60%
Finance, Undergraduate	3.40%
else	2.70%

As shown in Table no. S2, thirty-two percent of data science and Big Data job positions required vast knowledge of Big Data mining, Big Data acquisition, and Big Data analysis. Knowledge of database basics is also important, with 28.1% of data science and Big Data job positions requiring knowledge of databases and data warehouse. With the rapid development

of product data and user data, enterprises are placing greater emphasis on the enormous commercial value hidden behind Big Data. Of the data science and Big Data job positions, 24.2% require applicants to be proficient in machine learning, deep learning, and natural language processing. Enterprises also attach great importance to candidates' statistical basis and mathematical logic thinking. Of data science and Big Data positions, 13.0% required knowledge of mathematical basics such as mathematical models and optimization theory. With the increasing complexity of Big Data business requirements, multidisciplinary talents will gradually become the priority target of enterprises.

Table no. S2. The frequency of keywords in "professional knowledge"

Professional knowledge	Rate %
Data Mining, Data Acquisition, Data Analysis,	32.0
Database, Data Warehouse	28.1
Machine Learning, Deep Learning, Natural Language Processing	24.2
Mathematical Model, Optimization Theory, Logic	13.0
Else	2.70

Table no. S3 illustrates that enterprises require big data talent to be proficient in the mainstream software and technology of the big data industry. In this context, the study conducted further analysis using keyword indexing to identify the combinations of tools and technologies in high demand. The study identified 11 of the most in-demand triple combinations, which consist of programming languages, such as Python and Java and databases or data warehouses, such as SQL and MySQL. Table no. S3 further indicates that "SQL, Python, Hadoop" is the most popular combination followed by "SQL, Python, Java." Python and SQL hold absolute leadership positions in the big data industry. Enterprises aim to attract big data talent proficient in Hadoop, Python, Java, and other software to join their teams to create added value to enterprise development through data mining, management, analysis, and other process lines of business development in Big Data.

Table no. S3. The frequency of keywords in "programming ability"

Programming ability	Rate (%)
SQL, Python, and Hadoop	12.50
SQL, Python, and Java	12.20
SQL, Python, and Spark	10.90
Hadoop, Python, and MySQL	10.10
Hadoop, Python, and Java	9.90
Hadoop, Python, and Spark	9.80
MySQL, Python, and Java	8.90
Oracle, Python, and Java	8.70
Hbase, Python, and Java	7.10
Hadoop, Spark, and Hive	6.70
Hadoop, Spark, and Hbase	3.20

Table no. S4 demonstrates that internship experience is particularly important for enterprise recruitment. Specifically, 23.6%, 10.2%, and 20.1% of the positions require candidates with at least one, three or more, and two or more internships, respectively. Communication, management planning, professionalism, sense of responsibility, and other professional

qualities are the core elements used to evaluate the quality of candidates in corporate recruitment.

Table no. S4. The frequency of keywords in "work experience"

Work experience	Rate %	Work experience	Rate %
Internship Experience	23.6%	Job	1.0%
more than 3 times	10.2%	Analysis	1.0%
2-3 times	9.9%	Business	1.1%
Communication	8.9%	Project	1.1%
Management	6.9%	Team	0.9%
Proficiency	6.6%	Run	0.8%
Responsibility	4.1%	Energy	0.8%
Independent	3.5%	Office	0.7%
Logic	3.5%	Formulate	0.6%
Cooperation	2.7%	Advice	0.6%
Organize	2.5%	Market	0.6%
Architecture design	2.5%	Train	0.5%
Create	1.9%	One year of experience	0.5%
Data	1.1%	Deploy	0.5%
Experience	1.1%	Excellent	0.4%

Table no. S5. The proportion of postgraduate courses in Computer Science and Technology

Universities	Traditional Computer	Data Science	Basic Mathematics
Peking University	63.16%	31.58%	5.26%
Tsinghua University	50.00%	38.89%	11.11%
Zhejiang University	81.82%	18.18%	0.00%
National University of Defense Technology	65.22%	26.09%	8.70%
Beihang University	45.45%	36.36%	18.18%
Harbin Institute of Technology	50.00%	43.75%	6.25%
Shanghai Jiao Tong University	57.14%	33.33%	9.52%
Nanjing University	50.00%	44.44%	5.56%
Huazhong University of Science and Technology	57.89%	26.32%	15.79%
University of Electronic Science and Technology of China	50.00%	36.36%	13.64%
Beijing University of Posts and Telecommunications	62.50%	29.17%	8.33%

Table no. S6. The proportion of postgraduate courses in Master of Applied Statistics

Universities	Traditional Computer	Data Science	Basic Mathematics
Peking University	33.33%	50.00%	16.67%
Renmin University of China	19.05%	71.43%	9.52%
Nankai University	20.00%	66.67%	13.33%
Northeast Normal University	30.00%	55.00%	15.00%
East China Normal University	22.22%	66.67%	11.11%
Xiamen University	29.41%	58.82%	11.76%

Table no. S7. The course of Peking University and Northeast Normal University

Curriculum	The course of Peking University	The course of Northeast Normal University
Mathematical basis	Regression Analysis, Time Series Analysis, etc.	
Computer basis	Database, Data Structure, etc.	
Big Data system	Big Data Application, Big Data Development, etc.	
Industry Practice Section	Security Analysis, Epidemiological Statistics, Financial Statistical Analysis, Actuarial and applied area analysis courses	Human Resources and Social Security Statistics, Stock Investment and Fund Analysis, Educational Measurement, Securities Investment Market Analysis, etc.