A Study on Recruitment of Data Analyst Based on Text Mining and Visualization Technology

Fang Fang1, *, Yin Zhou1, 2
1School of Information Management, Wuhan University, Wuhan 430079, China
2School of Economics and Management, Hubei Engineering University, Xiaogan 432000, China

*Corresponding author: fang_fang@whu.edu.cn

Abstract. Nowadays, the demand for data analysts has increased dramatically. First the job requirements of data analyst are downloaded from online recruitment website ChinaHR.com and zhaopin.com, then the information of the recruitments are analyzed by text mining method such as word segmentation, word frequency statistics and visualization technology. At last we summarized the required skills for data analyst. We found that compared with Python, the company is inclined to use R for data analysis, the recruitment requirements of every type of company are different. Our findings can provide reference for schools and guide individual professionals to assess and advance their competencies.

Keywords: data analysts, text mining, career requirement

1. Introduction
With the development of information technology and the growth of global data, mankind has entered the era of big data. According to the "China Big Data Development Survey Report (2017)" released by the China Academy of Information and Communications Technology, the scale of China's big data market in 2016 was 16.80 billion yuan, with a growth rate of 45%. The growth rate of China's big data market will remain above 30% from 2017 to 2020[1].

In the era of big data, the demand for data analysis-related talents in various industries has increased dramatically, and the Internet industry's demand for data analysis-related talents has undergone greater changes. Traditional data analysis related courses are facing severe adjustments.

A large number of papers have analyzed the skill requirements that professional talents need. However, most of them use traditional research methods, such as questionnaire survey, enterprise visit and expert consultation, to analyze the skill requirements of professionals in a specific field [2].

Some scholars used recruitment information to analyze the skill requirements of enterprises for talents. Fang Hong used URL, JavaScript and other technologies to identify and extract job positions automatically [3]. Xiao Ximing used questionnaire survey, network survey and content analysis method to analyze 18 questionnaires and 274 job advertisements. Xiao investigated the demand of library and information professionals in different information professions from four aspects, such as the type of recruitment unit, position setting, job responsibilities and job ability requirements [4]. Si Li extracted the recruitment information from the recruitment websites in the United States, the United
Kingdom and Canada, then summarized the specific requirements of information professions in Europe and the United States for library and information talents from four aspects: recruitment category, job setting, knowledge and skills, and work experience [5-11]. At present, most companies are considering hiring data analysts. Because data analysis has an impact on specific groups of people, jobs and departmental structure. According to the 'Internet Talent 360° Impressions Report', data analysis positions are the seventh most sought-after positions by both supply and demand.

Recruitment information can reflect the company’s demand for talents. Therefore, investigating and analyzing the recruitment information for data analysis positions can understand the demand for data analysts in the Internet industry.

In this paper, crawler software is used to extract recruitment information of data analysts in the Internet industry from online recruitment information, and then text mining technologies such as word segmentation, word frequency statistics and visualization analysis are used to mine the demand for professional skills of data analysts.

2. Research Methodology

2.1. Text Mining

Text mining is a semi-automatic process of extracting patterns from a large amount of unstructured data, which generally consists of two main steps: transforming data sources into structured data, and using data mining techniques and tools to extract information and knowledge from the transformed data.

2.2. Research design

This paper used text mining, visualization technology and other analysis methods to analyze the data analysis of talent recruitment information. The specific processing flow is shown in Figure 1.

![Figure 1. Analysis Process of Data Analysts](image)

The process mainly includes three steps: first, data collection and data cleaning; The second is to build a dictionary to segment the collected online texts; Finally, the skill related words in the segmentation results are extracted and displayed in the form of visualization.

3. Data Collection Process

3.1. Data sources and collection pre-processing

A detailed search was conducted on the Zhaopin website (https://www.zhaopin.com) and ChinaHR (https://www.chinahr.com) website to find all possible recruitment information sources during 2017. The search keywords for the position were data analysts, and the geographical conditions were restricted to 5 cities: Beijing, Shanghai, Guangzhou, Shenzhen, and Wuhan. The degree of education was “undergraduate or above”. After searching, a total of 1068 recruitment information was found.
After using bazhuayu software (https://www.bazhuayu.com/) to crawl the 1068 information, removing duplicate information and irrelevant post information, there was only 588 unique and verified recruitment information. In this paper, 588 recruitment information was analyzed.

The terminology in the recruitment information released by each company is not uniform, and some of the recruitment information have spelling errors. For example, "Excel" is written as "excle", and "Python" is written as "pyton". First, we checked the spelling errors, then removed some information like job address, bonus, benefits and other related content in the recruitment information.

3.2. Word segmentation
This article uses the Jieba toolkit in Python to segment 588 pieces of information, combine synonyms, and remove words.

3.3. Visualization
After word segmentation, the result set is labeled according to the dictionary, and the professional skill words are identified for statistical classification, and the word cloud map is drawn with wordcloud.

4. Results analysis

4.1. Analysis of the type of recruitment unit
Judging from the crawled recruitment information, there are a total of 419 companies recruiting data analysts in the Internet field. Including JD Finance, Toutiao, 58.com, Baidu Waimai, Ele.me, Meituan Dianping, iQiyi, Sohu Media, Qihoo Technology, Xiaomi Technology, Guazi Used Cars, Hujiang Education, Duwan YY, Momo, etc. Many well-known companies. These 419 companies covered eight areas: Internet finance, social media/new media, cloud services/big data, online games/mobile games, O2O life services, e-commerce, online education, network/IT services.

4.2. Discipline Analysis of Recruitment
The Professional dictionary is based on “the catalogue of undergraduate majors” issued by the Ministry of Education in 2012, 588 recruitment information involves 46 majors and 26 categories: statistics, economics, computer science, business administration, graphic information, journalism, mathematics, sociology, psychology, linguistics, foreign languages, law, education, Pharmacy, clinical medicine, plant production, animal production, machinery, automation, industrial engineering, design, transportation, tourism management, physics, and chemistry. It seems that more and more industries begin to pay attention to the value of data analysis and introduce the position of data analyst in enterprises. In order to cultivate professional talents that meet the needs of enterprises, colleges and universities need to offer data analysis-related courses for specific majors.

4.3. Professional skills analysis
Table 1 classifies the professional skills in the recruitment information.

| Professional Skills          | Category                        | Content                                                                 |
|------------------------------|---------------------------------|-------------------------------------------------------------------------|
| Concepts and Methods         | Operating system                | Linux, UNIX, Android                                                    |
|                              | Technical development language  | R, Python, Java, Shell, Perl, Scala, C++, C                             |
|                              | Data processing and storage     | Hive, Hadoop, MySQL, Oracle, Spark, SQL Server, MapReduce, Storm, Hbase |
|                              | Visualization of statistical analysis | Excel, SPSS, SAS, Tableau, Matlab, Google Analytics, Mahout, Clementine, Weka, Eviews, Cognos |
|                              | Office software                 | Powerpoint, Word, Visio, MindManager                                   |

Table 1. Professional Skill Requirements for Data Analysts
It can be seen from Table 1 that there are many tools and technologies that can be used in the data analysis process, but the requirements of each company are different. At present, there are controversies about the application development techniques and tools that data analysts first learn. According to the number of times these tools and technologies appear in the recruitment information, use the wordcloud2 tool to get Figure 2.

In Figure 2, the larger the font of a word is, the more often the word appears. R is the most frequently used application development technology, followed by Python and Java; at the stage of data processing and storage, Hive, Hadoop, MySQL, Oracle and Spark are the most frequently used; As a specific application tool, Excel appears the most times, followed by SPSS, SAS, MATLAB, and Tableau. These tools can be used not only for work, but also for research.

![Figure 2: WordCloud of Professional Skills](image)

5. Conclusion
In the era of Big Data, the demand for data analysts in the Internet industry is gradually increasing. This paper analyzed the recruitment information of 588 and draw the following conclusions:

(1) Companies hiring data analysts in the Internet space fall into eight broad categories: Internet finance, social media/new media, cloud services/big data, online games/mobile games, O2O life services, e-commerce, online education, network/IT services

(2) From the perspective of the majors involved in recruitment, more and more industries are beginning to pay attention to data analysis, which means that universities need to offer relevant basic courses when formulating Talent Training Plan. Data analysts who work alone need to master a variety of skills, including the combination of business, data, and analysis expertise, while analysts who do not have the above knowledge need to cooperate with people of different disciplines.

(3) At present, there are many disputes over whether R or Python is used for data analysis. However, R appears more frequently than Python in recruitment information, it seems that companies are inclined to use R for data analysis in the Internet field. Those who are engaged in data analysis or students who are interested in the direction of data analysis can learn about the mainstream tools and technologies through this paper.

References
[1] China Academy of Information and Communications Technology. Research Report of China’s Big data development (2017) [EB/OL]. http://www.caict.ac.cn/kxyj/qwfb/ztbg/201703/ 20170327_2190526.htm, 2017.
[2] Dai Yue, Yu Jie. Research and Analysis of Financial Accounting Talents’s Demand and Cultivating mode [J]. Communication of Finance and Accounting, 2012(10): 158-160.
[3] Fang Hong, Lv Taizhi. Automatic Extraction of Job Listing Page Link Information in Job Web Station [J]. New Technology of Library and Information Service, 2009, 25(7): 93-96.
[4] Xiao Ximing, Li Shuo, Tian Rong. Research on the Demand of Library, Information and Archival Science Professionals from Different Information Occupations [J]. Library and Information, 2014(1): 35-40.
[5] Si Li, Jia Huan. Investigation and Analysis of the Information Professional Demand Based on the Library and Information Science Talents at Abroad [J]. Library Tribune, 2015(3): 102-108.

[6] Internet Data Center 199IT. Comprehensive Report of the Internet Talents’ Impression [EB/OL]. http://www.199it.com/archives/588636.html/, 2017.

[7] Di Chang, Xia Zhang, et al. "Location based robust audio watermarking algorithm for social TV system." In Pacific-Rim Conference on Multimedia, pp. 726-738. Springer, Berlin, Heidelberg, 2012.

[8] Di Chang, Xia Zhang, and Yue Wu. "A Multi-Source Steganography for Stereo Audio." Journal of Wuhan University (Natural Science Edition), 2013(3): 277-284.

[9] Xia Zhang, Di Chang, et al. "An Audio Steganography Algorithm Based on Air-Channel Transmitting." Journal of Wuhan University (Natural Science Edition) 57, no. 6 (2011): 499-505.

[10] Xia Zhang, Di Chang, et al. "Tree-like Dimensionality Reduction for Cancer-informatics." In IOP Conference Series: Materials Science and Engineering, vol. 490, no. 4, pp. 042028. IOP Publishing, 2019.

[11] Xia Zhang, Di Chang, et al. "A Study on Different Functionalities and Performances among Different Activation Functions across Different ANNs for Image Classification." In Journal of Physics: Conference Series, vol. 1732, no.1, p.012026. IOP Publishing, 2021.