DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.4, NOV. 2021

Study on big data –related position portrait based on recruitment data

Lei Liu*, Xiang Xu

School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China

ABSTRACT

This paper mainly studies big data-related position portrait construction based on recruitment data crawled from the 51job website. We first use text mining technique to classify job information and accurately obtain the texts of two important aspects: work content and job requirements. We then apply the information extraction technique to extract labels describing different aspects of positions from structured and unstructured text data, and also adopt the Kano model to obtain more labels. Finally, we construct a multi-aspect and multi-dimensional position portrait through the sunburst chart. The position portrait constructed in this paper provides multi-dimensional analysis of the requirements of big data-related jobs and can help job seekers, enterprises, universities, and even third-party training institutions know the demand for talents and quickly determine the pertinence of a candidate's resume.

KEYWORDS

Position portrait; Text classification; Label extraction; Labeling system; Kano model; Sunburst chart; Sankey diagram

1 Introduction

Due to the rapid development of the internet application and accelerating diffusion of information, big data technology has been applied widely to many aspects of human society, for example, medical health, transportation, finance, insurance, education, scientific research, e-commerce, tourism, and so on. Big data technology has a profound impact on scientific research, thinking mode, and human society. Since big data is an emerging industry, people lack basic knowledge and understanding of the employment prospects and rigidity requirements of relevant positions. Most enterprises do not yet process sufficient understanding regarding big data and their needs for big data talents. This will be detrimental to the flow and allocation of talent resources. Therefore, how to accurately summarize the current situation of the position is extremely important.

There have been many studies on data analysis and text mining by using internet recruitment data in the past few years. For example, Baumeister et al. (2020) used a content analysis technique named centering resonance analysis (CRA) to characterize required skills and competences for data specialist roles by analyzing job advertisements for data scientists and other related professionals. Monica Maceli (2015) applied Agglomerative Hierarchical Clus-

*Corresponding author: leiliu@zufe.edu.cn

tering to classify jobs based on extracting common working terms. Song et al. (2021) obtained association rules based on the Apriori algorithm and constructed a relational network diagram to explore the internal relationship of job information.

User portrait is a fast and precise data analysis tool to analyze user behavior patterns , consumption habits and other business information , which can lay the foundation for precise marketing and user experience improvement. Massanari (2010) constructed a user portrait model based on the characteristics of users' interests and hobbies, and emphasized its important role in the process of product development. Li et al. (2021a) constructed a talent portrait model from basic qualifications, knowledge requirements, tool skills and abilities, and collected online recruitment data in the archives field to explore the characteristics of talent needs. Travis (2017) proposed characteristics of user portrait, including Primary research, Realistic, Objectives, Singular, Empathy and so on.

In this paper, we follow the construction strategy of user portrait and apply the text mining technique to internet recruitment data. A comprehensive, multi-dimensional position portrait is constructed to help people have a comprehensive understanding of certain positions. To make the final results of position portraits as objective and effective as possible, this paper follows the following three principles:

Timeliness. Accounting for the irregular update of recruitment information will lead to the coexistence of new and old recruitment data, this paper makes a special filtering of the data to ensure that the data used are the latest. At the same time, position portraits should be able to make timely adjustments based on new data.

Trueness. The collected recruitment data must be true and reliable, and the corresponding positions must be real.

Comprehensiveness. Under the premise of ensuring the necessary association between the selected labels and the positions, the aspects of describing the positions should be increased as much as possible to construct a labeling system with a clear relationship and diversified dimensions.

To meet the actual needs of multi-dimensional information extraction and label extraction of related positions, we innovatively propose a two-level text classification model, which accurately classifies the topics of text blocks (such as work content, job requirements). Based on the result of classification, we could extract labels accurately from the unstructured text data. Moreover, we apply the Kano model to classify position labels from two dimensions of universality and importance to understand the similarities and differences of various positions. The Kano model was initially applied to study the relationship between product performance and user satisfaction and to classify and rank users' needs. It has also been used to find key features that affect consumer satisfaction (Kovačević & Bota, 2021), classify the needs of consumers from product reviews (Shi & Peng, 2021), and process the text of user preference extracted by the LDA model (Li et al., 2021b). The complete position portrait combined with labels from structured and unstructured data can describe various positions very well. If further combined with the text-similarity technique, accurate talent matching can be achieved. The position portrait could play an important role in the practical application of the job seekers' job recommendation and enterprise resume filtering.

2 The Collection of Positions Data

2.1 Data Source

Chnbrand, China's brand rating authority, regularly releases the China Brand Power Index

(C-BPI). From 2011 to 2021, 51job is the only internet recruitment website whose brand power has been ranked in the top two in the industry. As a comprehensive recruitment website, 51job's service agencies are all over the country. After long-term operation and accumulation, its job information in different regions and fields is sufficient. Therefore, we choose 51job recruitment data to analyze.

2.2 Data Crawling Process

By analyzing the composition of web pages and the naming rules of target web pages, the target web pages are divided into the following two categories: basic information webpage and detailed information webpage. Among them, the basic information page also shows the basic information of multiple positions, which is convenient for job seekers to browse quickly. The detailed information page is only for a specific position, for the in-depth job understanding. Because the basic information pages and the detailed information pages are links, the contents of two kinds of web pages are crawled successively to ensure the accuracy and integrity of information collected.

2.3 Data Collection Results

Since there are so many kinds of position related to big data, it is difficult to determine the selection range of job data directly. Therefore, we first collect 61242 pieces of recruitment information with "big data" as search keyword. After deleting irrelevant data and classifying the data by functional categories of positions, we obtain 29 positions related to big data, which are mainly divided into the following five categories: data collection, data analysis, data mining, database, and back-end development. To obtain more related position data, we use each position name as search keywords to crawl the website again and obtain 103953 pieces of data. The fields of the crawled text data are mainly divided into structured and unstructured data, as shown in Table 1. We adopt different processing methods for different fields.

Table 1 Crawled data fields

Field Type	Field Name	Processing Methods
Structured Data	Company Nature, Company Industry, Company Size, Work Place, Recruiting Number, Work Experience, Education Requirements (Single option)	Basic Statistical Analysis
	Position Welfare (Multiple options)	
	Position Salary (Short Text)	Convert to Numerical Type Construct Statistical Characteristics
Unstructured Data	Position Information (Long Text)	Text Mining

3 Job Information Classification

3.1 Two–level Text Classification Model

Job information is the only long text field in job data, which mainly includes two parts: job content and job requirements. These two parts are the key content to extract position labels. In order to extract labels more precisely, we need to classify these two kinds of information.

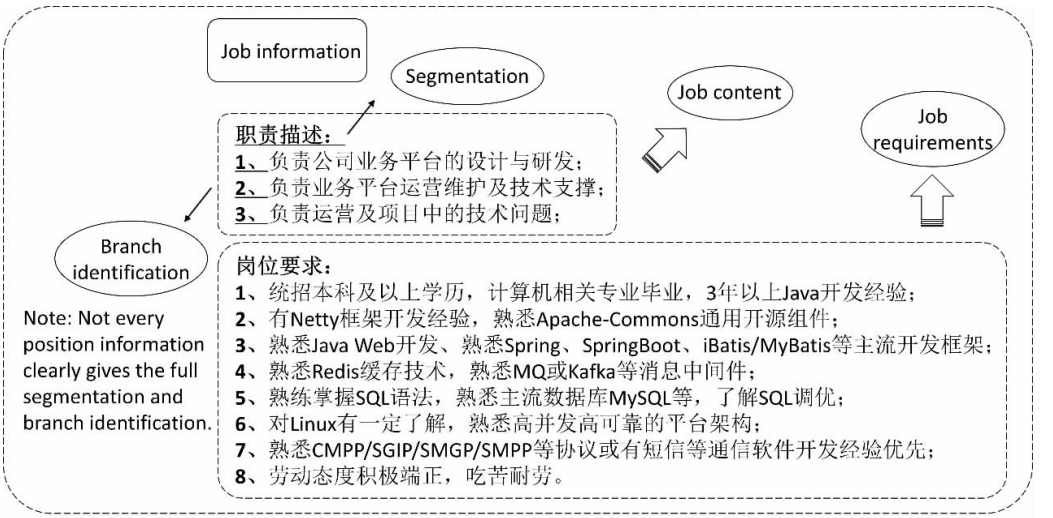

Figure 1 Position information characteristics analysis

As shown in Figure 1, the detailed description of job information is composed of headings and lists with numbers. The headings can be divided into job content and job requirements according to semantics. However, the expression of headings is not uniform. Through statistical classification, we construct a synonym dictionary about different titles of job descriptions, as shown in Table 2. This allows two categories of job information to be obtained by heading segmentation. However, we should note that the titles could be missing on some web pages. For this case, we establish a text classification model to predict the topics of the descriptions.

Table 2 Synonyms dictionary of job descriptions

Text topics	Synonym
Job content	职位描述、岗位职责、工作职责、岗位说明、基本职责
Job requirements	任职资格、岗位要求、技能要求、应聘条件、入职要求

On the basis of the determination of heading type, the process of the job information segmentation is as follows: 1. Segment job information into text blocks according to the headings, and then cut text blocks into text strings according to numbers and punctuations. 2. Topics of the text blocks are decided by the headings. The topics of text blocks with missing headings will be determined by classification models. 3. Train classification models by using text blocks with clear topics and predict the topics of text blocks with missing headings. The process of segmentation and classification is shown in Figure 2.

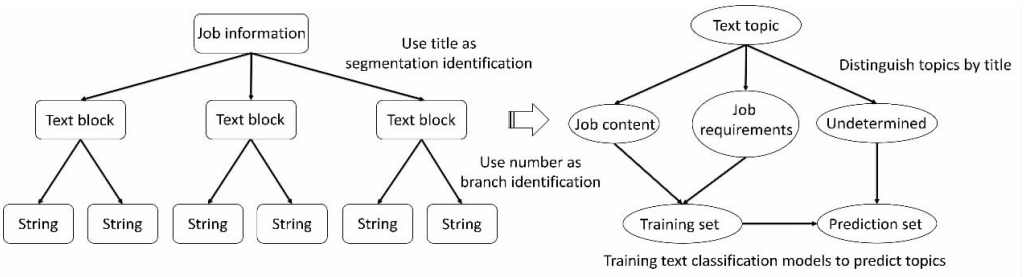

Figure 2 Segmentation and classification of job information

3.2 Training and Evaluation of Classification Model

To obtain the topics of text blocks with missing headings, we construct text classification models. Accordingly, we take text blocks with clear topics as training set and test set, and text blocks without headings as prediction set.

To vectorize the text data, we use the Python library jiebaas, the segment tool. In order to obtain an accurate model, we adopt four different representation methods, such as the One-Hot encoding, word frequency vector, TF-IDF vector, and n-grams. We combine them with five machine learning models, such as K-nearest neighbor, Naive Bayes, Logistic Regression, Linear Support Vector, and Random Forest (Manning & Schutze, 1999). The evaluation results of different combinations of models are shown in Table 3.

Table 3 Evaluation of text classification models

Text representation / Model	K–Nearest Neighbor	Naive Bayes	Logistic Regression	Linear Support Vector	Random Forest
One–Hot Encoding	96.55%	97.79%	97.48%	96.46%	98.03%
Word Frequency Vector	96.55%	97.79%	97.48%	96.43%	98.11%
TF–IDF Vector	95.28%	97.50%	98.07%	97.90%	98.13%
N–Grams	96.77%	97.46%	97.23%	96.04%	98.08%

From Table 3, we can see that the model obtained by TF-IDF vector and Random Forest can achieve the most accurate result. By inspecting the text blocks misclassified, we find that most of the mistakes are made by the website; that is, some text blocks are titled wrong headings. In summary, the Two-Level Text Classification Model proposed can classify the job content and job requirements very well. It is helpful for constructing an accurate labeling system for position portraits.

4 Label Extraction

4.1 Labeling System

A position can be well distinguished from the others by analyzing its job content and job requirements. Job seekers would pay more attention to the position treatment and basic information of the company, i.e., company nature, company size, company industry, etc. According to people's actual needs to understand the position and the data provided by the recruitment website, we construct a multi-level position portrait labeling system. The labels

of the first two levels are shown in Figure 3. In the following part of this section, we will focus on label extraction of the third level.

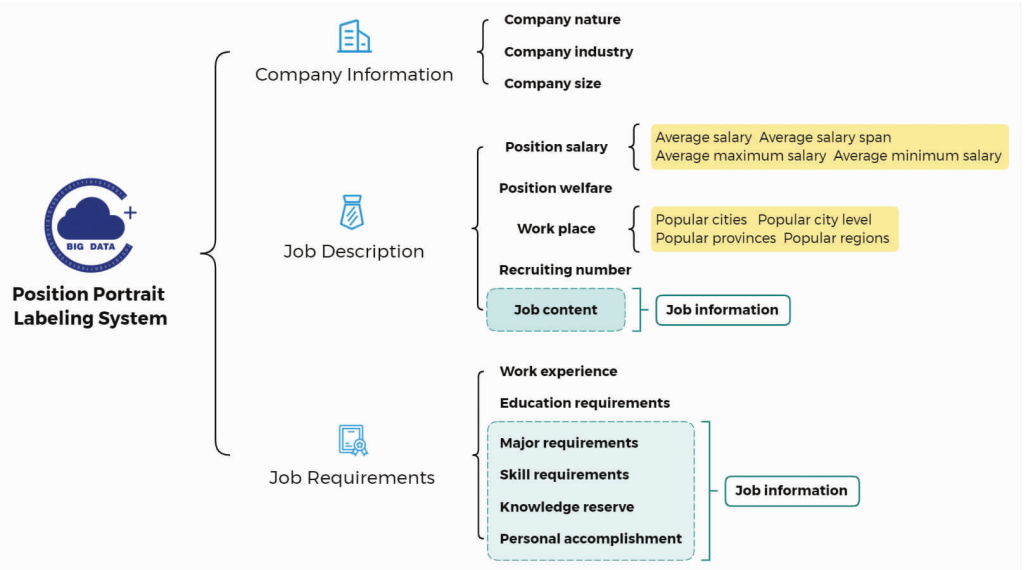

Figure 3 Position portrait labeling system

4.2 Label Extraction from Structured Text

Most third-level labels in the labeling system can be directly obtained from the corresponding field in the recruitment data. For example, for company nature, we only need to enumerate all the values appearing in the data. But for some numeric labels, extra processing is needed. For example, for position salary, we need to convert the value into a unified numeric form and use its statistical characteristics to reflect the average salary and floating range. The values of workplaces are specific to the districts and counties, so workplace is a good label for people to understand the employment situation in a certain region.

4.3 Label Extraction from Unstructured Text

After the above text processing and classification, the job description is divided into two categories, i.e., job content and job requirements. The two parts are tagged and extracted respectively. Job content is difficult to be further subdivided. But the job requirements can be subdivided into several categories, such as skill requirements, knowledge reserve, and personal accomplishment. Note that each sentence in the job requirements may include one or more than two requirements, so it is difficult to build a multi-classification model. Therefore, we first extract keyword, and then analyze the category of words, and finally classify the words in various requirements.

The traditional keyword extraction algorithms, such as TF-IDF, PageRank, and LDA, are not very suitable in the label extraction of job content and job requirements, since those models are powerless to extract the professional terms composed of several words. Therefore, we adopt a new C-value method proposed by Frantzi et al. (2000) in the automatic extraction of technical terms, which can well extract the multi-word terms (nested terms) with high occurrence frequency.

The formula for calculating the C-value is as follows

$$C-value(x)=\begin{cases}\log|x|\cdot tf_x^T & x \text{ not included by other labels}\\ \log|x|\cdot\left(tf_x^T-\frac{1}{|C_x|}\sum_{y\in C_x}tf_y^T\right) & \text{others}\end{cases}$$

wherexis a candidate label, |x| is the length of x,tf_x^T is the frequency of x appearing in the recruitment data, C_x is the candidate label set contained in the recruitment data and $|C_x|$ is the number of elements in the set The formula shows that C-value is related to the frequency of the candidate label in the C_x text, and the length of the candidate label is taken into consideration. It also tends to choose more meaningful long string labels. Some of the results of TF-IDF and C-value label extraction methods are shown in Table 4.

Table 4 Comparison of label extraction methods

TF–IDF	C–value Method
开发	需求 分析
设计	开发 工作
系统	详细 设计
项目	技术 文档
技术	设计 文档
需求	业务 需求
分析	技术 问题
产品	参与 项目
文档	文档 编写
代码	系统 设计

There are many labels with similar semantics extracted by the C-value method, which needs to be unified. We use job data to train a word vector model to realize the vectorization of label semantic information. The Euclidean Distance between word vectors is used to reflect the text-similarity between labels, and the labels are clustered by combining the Single-Pass Flow Clustering Algorithm. Unknown new labels are classified only if they are sufficiently similar to the existing categories of the model; otherwise, they become a new category. Some clustering results are shown in Table 5. Clustering can reduce the number of labels and improve the effect of label extraction.

Table 5 Label unification

Category	Synonymous Labels
项目开发	开发工作、开发任务、负责项目
系统设计	架构设计、概要设计、详细设计
文档编写	技术文档、设计文档、开发文档

The labels covered by the job requirements need to be further classified. We solve the classification problem by acquiring the knowledge representation of labels. Firstly, a regular expression is used to match the text string according to common sentence patterns. The nominal phrases are emphasized in parentheses, as shown in Table 6.

Table 6 Text matching results

Patterns	Extraction Results
List before Category behind	熟练【Mysql、Oracle、DB2、SQL Server】等【关系型数据库】
Category before List behind	熟悉【ETL 工具者】优先，如【kettle、Datax、Datastage、 Informatica】等

Further dependency parsing of short sentences is shown in Figure 4. Mining the parallelism and species relations between noun phrases, such as MySQL and Oracle are parallelism, MySQL belongs to the relational database, etc. Then, the label category is judged according to the triple reflecting the relationship, and the label can be correctly classified.

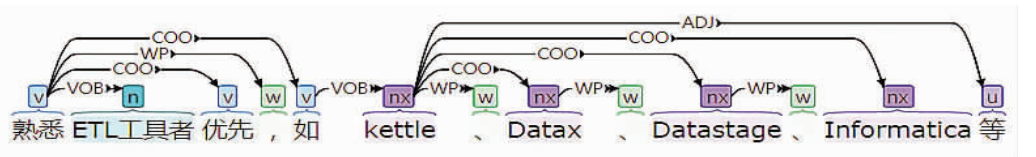

Figure 4 Visualization of the dependency syntax tree

4.4 Kano Model

The Kano model was first introduced to study the relationship between product performance and user satisfaction and to classify and rank users' needs. In this paper, we use the Kano model to classify the labels according to universality and importance. For example, for skill requirements, the KANO model can tell us which skills are commonly needed by most positions and which skills are required by only a few.

In the Kano model, there are two indicators that should be investigated, which are Label Proportion PCF_i and Label Importance PCD_i.

Label Proportion is defined to be the proportion of the occurrences of the i^{th} label to the total number of the position recruitment information. Taking into account the differences in the number of recruitment information for each position, it can be calculated by the average

$$PCF_i = \frac{1}{m}\sum_{p=1}^{m}\frac{R_i^p}{R^p},$$

where m is the total number of positions, F_i^p is the ith label of position P and R^p is the total number of recruitment information for position p. PCF_i is mainly used to characterize the universality of labels in all positions.

Label importance of the i^{th} label is defined as

$$PCD_i = \sum_{p=1}^{m}\theta_i^p/m,$$

where θ_i^p is a determining factor. If the i^{th} label is important for position p, the value θ_i^p is 1, otherwise, it is 0. The formula is as follows

$$\theta_i^p = \begin{cases} 1, R_i^p/R^p - PCF_i \geq 0 \\ 0, R_i^p/R^p - PCF_i < 0 \end{cases}.$$

From the definition of PCD_i, we can see that when the i^{th} label gets close to 1 in the job content of each position, it is important for all positions. When PCD_i gets close to 0, there may be great differences in the occurrence of the label in various positions, and it is only important for some positions.

Based on the feature definition of the Kano model, labels are divided into four categories according to the above indicators: Necessary Label, Expecting Label, Charm Label, and Exclusive Label, as shown in Figure 5.

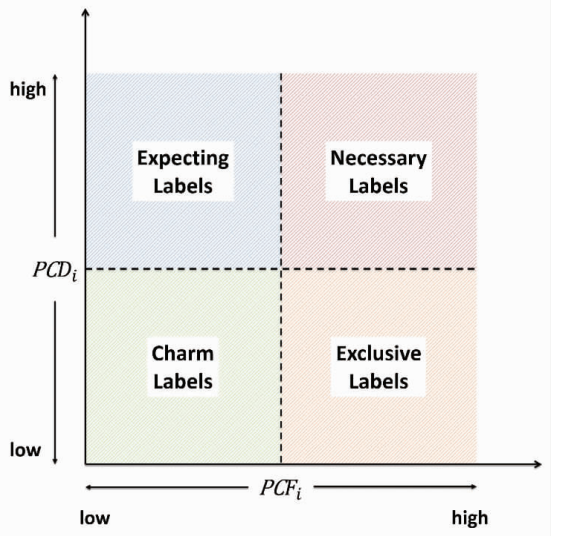

Figure 5 Labels classification of the Kano Model

Necessary Labels are frequently mentioned by most positions (higher value of PCF_i), and they are important for different positions (higher value of PCD_i). Such kinds of labels deserve more attention.

Expecting labels aren't particularly focused on compared to Necessary Labels (lower value of PCF_i). But these labels remain more evenly distributed across all positions (higher value of PCD_i), indicating that they still have some value.

Charm Labels are not common in many jobs (lower value of PCD_i), and they are not important for most positions (lower value of PCF_i). Such labels may only play a certain role in some positions and do not need to cause too much attention.

The occurrence of **Exclusive Labels** overall is higher (higher value of PCF_i), and they are only important for some positions (lower value of PCD_i). It indicates that such labels are extremely important for some positions, and they significantly make those positions different from other positions.

Through the overview of various labels, if we consider the universality of labels, the Necessary Labels and Exclusive Labels can be regarded as common labels, while the Expecting Labels and Charm Labels can be regarded as rare labels. If we consider the importance of labels, the Necessary Labels and Expecting Labels can be classified as common labels, Charm Labels, and Exclusive Labels can be classified as personalized labels.

We can apply the Kano model to classify all the labels in our labeling system into four categories. Here we take the labels of skill requirements as an example. The classification results are shown in Table 7.

From Table 7, we can see that Necessary Labels are Python, Linux and SQL. It means that those skills are generic for big data-related positions and are important for those who wish to engage in big data-related occupations. Skills of the Exclusive Labels do not widely occur, but they are important for some positions. For example, 80% of Java development engineer's

Table 7 Label classification results of skill requirements

	Skill Requirements
Necessary Labels	Python、Linux、SQL
Exclusive Labels	Java, MySQL, Oracle, Redis, Spring, Mybatis, SpringBoot
Expecting Labels	C++, C, Hadoop, shell, Spark, Hive, R, ETL, HBase, SQLServer, Caffe, Flink, Scala, BI, office, HDFS, APP, PLC, C#, SAS, Kettle, DB2, Tableau, LR, NLP, Storm
Charm Labels	TensorFlow, Pytorch, OpenCV, Excel, Web, PPT, Matlab, JavaScript, MongoDB, Kafka, SpringCloud, SpringMVC, Tomcat, JQuery, Django, Flask, HTTP, scrapy, xpath, IP, HTML, CSS, Halcon, VisionPro, GCP, ElasticSearch, Lucene

position information contains Java, and Oracle appears in 54% of ETL development engineer's position information. This indicates that the Exclusive Label skills are the core skills of some positions for which demand those skills. In addition, the Expected Labels and the Charm Labels involve more skills. The Expecting Labels are more likely to appear in more positions than the Charm Labels.

Overall, for the labels of skill requirements, the Necessary Labels are more important than the Expecting Labels, while the Expecting Labels are more important than the Charm Labels. The Exclusive Labels need special treatment. People can focus on the categories of labels depending on their needs. For universities and third-party training institutions, the Necessary Labels are instructive for teaching the most common and practical techniques and skills. For job seekers with a clear direction of employment, they can refer to more Exclusive Labels and Expecting Labels. People who want to become experts in a less popular field may pay more attention to the Charm Label.

4.5 Results

The position portraits are usually presented by the text. Whereas we visualize final position portrait labeling system by sunburst charts. Sunburst charts can well show the hierarchy of position label construction, since all the labels are tightly around circles and the importance of labels are indicated by the lengths of arcs. Sunburst charts help us understand a position in all aspects. The sunburst chart of Algorithm engineers is shown in Figure 6.

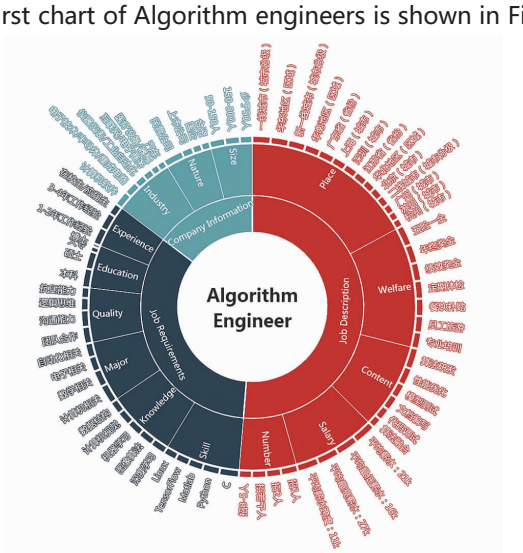

Figure 6 Position portrait labeling system of algorithm engineers

Although Sunburst charts can well show the hierarchy of position label construction, it is difficult to compare the similarities and differences between different positions by them. To better explore the differences of different positions, we plot the Sankey diagram of the labels. As an example, we select four positions belonging to the category of data mining to draw the Sankey diagram of skill requirements, as shown in Figure 7.

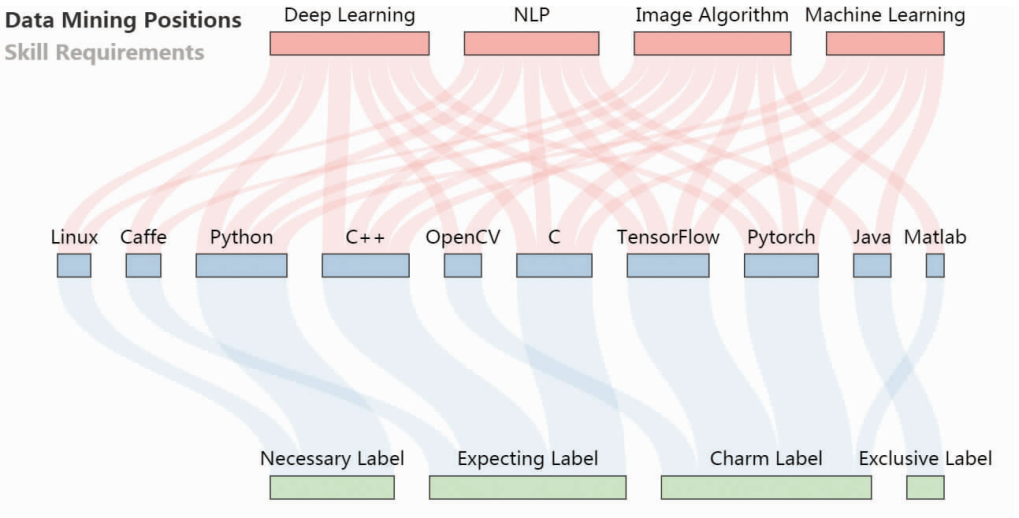

Figure 7 Skill requirements for data mining positions sankey diagram

The Sankey diagram shows the common and unique parts of the labels for various positions. We can see that different positions could share some skills, such as Python, C++, TensorFlow, etc., but there are also differences in the importance of labels to different positions demonstrated by the width of the pink connecting lines. Moreover, combined with the results of Kano Model on label classification, we can further understand the value and significance of the skills for various positions.

5 Conclusions

Taking big data-related positions as an example, we apply the techniques of web crawler, data analysis, and text mining to collect, collate and analyze the recruitment information and construct a position portrait labeling system. Based on the structured characteristics of job information, we propose a two-level text classification method, which decomposes job information into job content and job requirements. On this basis, the labels of each category of the text data are extracted. Moreover, the categories of labels are clarified through syntactic analysis, and the attributes of labels are given by using the Kano model. Finally, we construct a multi-aspect and multi-dimensional position portrait which can help job seekers, especially graduates, more intuitively understand the needs and development of the position. The position portrait will also help small and medium-sized enterprises to determine whether to add more relevant positions according to the actual situation.

In future work, we will improve the labeling system, which is supplemented by verb-object phrases. An automatic generation of job summaries could be realized and applied to question answering systems. In addition, the resume information of job seekers can also be

extracted by similar methods as we have done in this paper. In this way, we can calculate the text similarity of positions and resumes on the labels. The technique can be applied to automatically resume selecting systems and position recommendation system to provide more suitable position for job seekers.

References

Baumeister, F., Barbosa, M. W., & Gomes, R. R. (2020). What is required to be a data scientist?: Analyzing job descriptions with centering resonance analysis. *International Journal of Human Capital and Information Technology Professionals (IJHCITP), 11* (4), 21–40.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi–word terms: The C–value/NC–value method. *International Journal on Digital Libraries, 3* (2), 115–130.

Kovačević, D., & Bota, J. (2021). Consumer satisfaction with packaging materials: Kano model analysis approach. *Tehnički vjesnik, 28* (4), 1203–1210.

Li, Y., Chen, X., Mao, T., & Huang, G. (2021a, April). User portrait for archival talents based on recruitment. In *2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (pp. 116–120). IEEE.

Li, M., Zhang, J., & Pan, W. T. (2021b). Integrating Kano model, AHP, and QFD methods for new product development based on text mining, intuitionistic fuzzy sets, and customers satisfaction. *Mathematical Problems in Engineering, 2021* (5), 1–17.

Maceli, M. (2015). What technology skills do developers need? A text analysis of job listings in library and information science (LIS) from Jobs.code4lib.org. *Information Technology and Libraries, 34* (3), pp. 8–21.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing.* MIT press.

Massanari, A. L. (2010). Designing for imaginary friends: information architecture, personas and the politics of user–centered design. *New Media & Society, 12* (3), 401–416.

Shi, Y., & Peng, Q. (2021). Enhanced customer requirement classification for product design using big data and improved Kano model. *Advanced Engineering Informatics, 49* (1), 101340.

Song, Z., Yang, Y., & Guo, H. (2021). Analysis of data crawling and visualization methods for recruitment industry information. Journal of Physics: Conference Series, 1971 (1),012092.

Travis, D. (2017). *E–commerce usability: tools and techniques to perfect the on–line experience.* CRC Press.