Journal of Artificial Intelligence Practice (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2022.050115
ISSN 2371-8412 Vol. 5 Num. 1

Research on China's Postdoc Talent Profiling Based on Big Data

Yuchen Wu[a,*], Ying Xiong[b], Yi Wang[c]

Chinese Academy of Personal Science, 110100, Beijing, China
[a]18612199962@163.com, [b]15210337113@163.com, [c]richenwu@163.com
**Corresponding Author*

Keywords: Big Data, Postdoctoral Researcher, Talent Profiling

Abstract: The abundant talent information resources under the context of big data provide new opportunities for profiling of postdoctoral talents based on big data technology. The existing research on talent profiling carries out single-dimensional evaluation and measurement mostly based on limited information, and figures out some problems such as insufficient objectivity and incomplete evaluation and measurement, failing to meet China's prospective demand for labeling and selecting scientific talents in an all-round and efficient way. Talent profiling through information extraction (IE), big data association mining, artificial intelligence (AI), and other techniques can realize comprehensive and efficient labeling, selection and evaluation of talents based on the needs of different subjects, and thus can replace resumes, allowing employers to fully understand postdoctoral individuals and groups, overcome HR information asymmetry, leverage the prospective and guiding effect of talent introduction in China, and seize the initiative in global competition for talent.

1. Introduction

Talent profiling was first proposed by Alan Cooper and early talent profiling represents the prototyping of real or potential users, rather than description of real individual or regular users[1]. The profiling technique is important for mining and analyzing user information. Profiling of postdoctoral researchers (postdocs) borrows ideas from talent profiling, and uses a text information analysis and processing technology to label postdoc information and establish a postdoc user model. As information technology (IT) evolves, users are making more information public on the Internet. In this way, talent profiling in the era of big data is different from that in the early days, that is, extracting and abstracting a people-related information set from massive data and then using it to describe user characteristics[2]. Talent profiling is an important technology for mining and analyzing user information.

With the rapid development of network information technology, the explosion of information and knowledge leads us into an era of big data. Extracting valuable information from massive data has become a hot topic in the academia and industry. Everyone who uses the Internet will generate relevant data information, which is scattered across data islands, bringing great challenges to data analysis. Postdoc information mainly includes personal attributes, research achievements and

projects involved.

Specifically, postdoc talent profiling based on big data technology is to abstract, classify and label collected postdoc data by using data crawling, analysis and processing techniques, etc. According to the differences in goals, behaviors and viewpoints among postdocs, highly accurate postdoc feature identification is extracted, such as research achievements, hobbies, etc. for an all-round presentation of individual and group characteristics of postdocs, and finally display them intuitively through visualization.

2. Profiling Framework

Postdoc profiling borrows ideas from talent profiling, and uses a text information analysis and processing technology to label postdoc information and establish a postdoc user model. Postdoc information mainly includes personal attributes, research achievements and projects involved. As shown in Fig. 1, the overall framework for postdoc talent profiling mainly includes four parts: data collection, corpus processing, profiling and prototyping. Data collection is fundamental to the whole system. By building a postdoc information database, analyzing postdocs big data, processing Chinese corpus information, and generating postdoc word cloud images, key phrases, automatic abstracts, etc., the personal and national postdoc profiling prototype systems are constructed, which can display postdoc profiles. Postdoc information includes personal attributes, research achievements and projects involved.

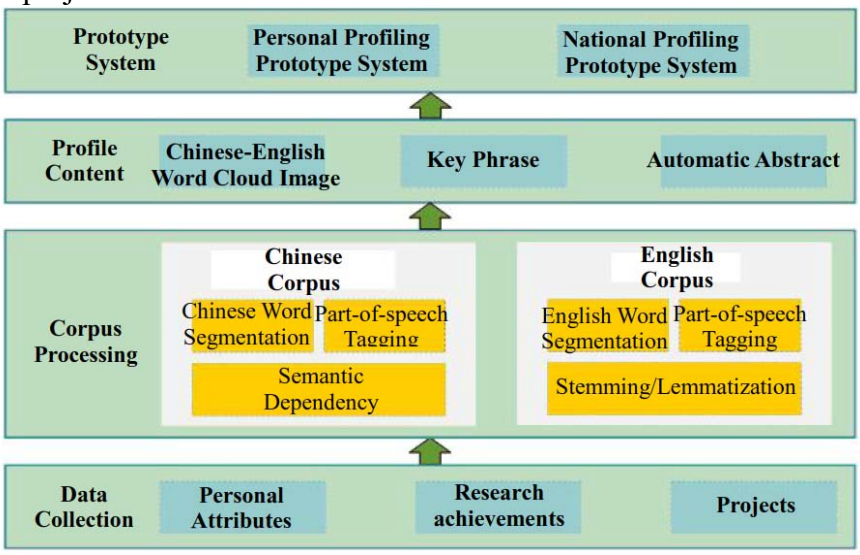

Fig. 1 Postdoc Talent Profiling Framework

3. Postdocs Database Construction

The postdoc information crawled or collected from the Internet is as varied as the data formats. Postdoc profiling is made based on structured data, which is helpful for further analysis and processing of key information. Therefore, a MySQL relational database is created by mining and collecting postdoc information to facilitate centralized management and improve data utilization. The postdoc profiling database mainly includes data collection, processing and storage.

The postdoc data collected herein are characterized by diverse sources, strong randomness and large data volume, etc. For example, postdoc population attribute data collected from the website of China Postdoctoral Science Foundation contain 32,181 records about the personnel funded by the host institutions such as schools, academies, institutes and companies in China from 2016 to 2020.

In addition, this paper studied 18,433 postdoctoral researchers who have already been enrolled as postdoctoral fellow in a postdoctoral station in China and have been funded in recent five years, as well as their relevant data.

(1) Data Collection

Postdoc information includes personal attributes, research achievements and projects involved. Personal attributes include name, major and affiliation; research achievements mainly include titles, authors, keywords, abstracts, types, downloads and citations of published papers in Chinese and English; projects involved mainly include names, participants and other information of projects that he/she participates in or heads, e.g., programs funded by the National Natural Science Foundation of China (NSFC).

Based on Python's PySpider framework, postdoc population attribute information and Chinese academic achievements were crawled from China Postdoctoral Science Foundation and CNKI. Through manual collection and batch export, postdocs' English research achievements and projects involved were collected from the Web of Science, Engineering Index, and NSFC Funded Project Information Search System under Internet-based Science Information System (ISIS).

(2) Data Processing

There are multiple formats of data crawled by Python and exported in batch, requiring further processing before storage and utilization. Among the collected data, the population attributes mainly include data unrelated to population attributes, such as serial number and grant amount. The Chinese research achievements contain some crawled data which do not belong to the object because of duplication of names. In addition, some fields, such as downloads and citations per paper, contain special characters that affect the statistical charting of the prototype system. The English research achievements also contain some collected data which do not belong to the object because of duplication of names. They are unstructured bibtex file data, which does not meet the actual demand for structured data [3].

In order to address the problems in the collected data, different data processing techniques are used to clean such data. Any fields unrelated to population attributes are deleted. By qualifying postdoc majors, the research achievements of those who have the same name are deleted. By regular expression matching, target information is extracted from the field values mixed with special characters. By writing Python scripts, unstructured English research achievements are batch-converted into structured data.

(3) Database Storage

The collected data are all structured data after processing. To facilitate management and analyze the important information by natural language processing (NLP), a relational database is designed to store data. Through SQL statements, the relational database can perform all kinds of complex query operations, which greatly facilitates the statistical analysis of data by the scientific profiling prototype system[4].

user_info, research_achievement_cn, research_achievement_ei, research_achievement_sci and research_project tables were designed for the postdoc information database, which store postdoc staff attributes, Chinese research achievements, English EI achievements, English SCI achievements and scientific research projects, respectively.

The user_info table contains fields such as id, name, province, major and affiliation, as defined in Table 1.

Table 1. User_info Field Definitions

Field Name	Field Type	Description
id	int	Record Number
postdoc_name	varchar(30)	Postdoc Name
province	varchar(30)	Province
major	varchar(30)	Major
affiliation	varchar(30)	Affiliation

The research_achievement_cn table contains fields such as id, name, title, keywords, author, number of authors, downloads, citations, year of publication and paper type, as defined in Table 2.

Table 2. Research_achievement_cn Field Definitions

Field Name	Field Type	Description
id	int	Record Number
postdoc_name	varchar(30)	Name
title	text	Title
keywords	varchar(50)	Keywords
abstract	text	Abstract
authors	varchar(30)	Author
download_count	int	Downloads
ref_count	int	Citations
year	int	Year
article_type	varchar(10)	Paper Type

The research_achievement_ei and research_achievement_sci tables contain the same fields such as id, name, title, keywords, abstract, author, citations and year of publication, as defined in Table 3.

Table 3. Research_achievement_sci Field Definitions

Field Name	Field Type	Description
id	int	Record Number
postdoc_name	varchar(30)	Name
title	text	Title
keywords	varchar(50)	Keywords
abstracts	text	Abstract
authors	varchar(30)	Author
ref_count	int	Citations
year	int	Year

The research_project table contains fields such as id, name, project name, project principal and his/her name, as defined in Table 4.

Table 4. Research_project Field Definitions

Field Name	Field Type	Description
id	int	Record Number
postdoc_name	varchar(30)	Name
project_name	varchar(50)	Project Name
is_ principal	int	Is he/she the project principal? 0 means No and 1 means Yes.
principal_name	varchar(30)	Name of Project Principal

4. Postdoc Profiling Prototype System Construction

The general fund of China Postdoctoral Science Foundation (CPSF) refers to the start-up fund for postdoctoral fellows to engage in innovative scientific research, and the number of funded postdocs is about one-third of the number of postdocs who register in the host institution that year.

It is helpful to understand the current situation of postdoc training in China by analyzing the regions and affiliations of postdocs funded every year. How to display complex statistical analysis results visually is the focus of this paper.

In response to the actual demand for data visualization, personal and national postdoc profiling prototype systems are constructed by using HTML web page building technology and Echarts visualization library, respectively. Echarts, a visualization and charting library implemented using JavaScript, contains rich visual chart types, is compatible with mainstream browsers, can display the visualization of tens of millions of data, and supports in-depth interactive operations.

(1) Overall Framework

The postdoc profiling prototype system mainly includes personal and national profiling, and its overall framework is shown in Fig. 2. Personal profiling, based on the collected personal information about research achievements and projects, shows statistical charts on 3D Chinese-English word clouds, automatic abstracts and research achievements. National profiling, based on the information about postdoc grants in recent five years, shows statistical charts on quantity, major, affiliation and region of funded postdocs nationwide, thus demonstrating the actual situation of postdoc training in China from different angles.

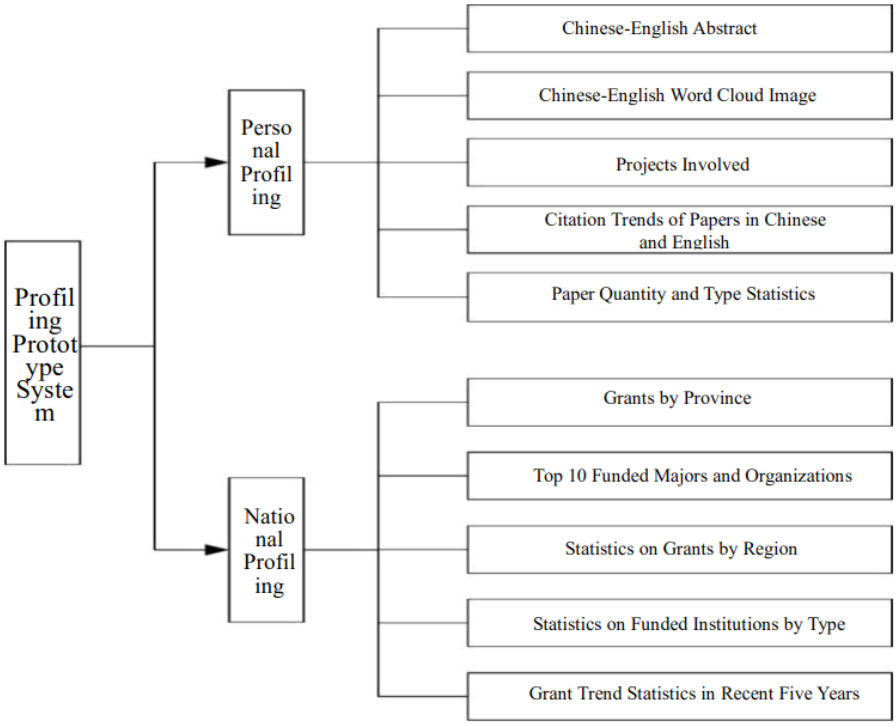

Fig. 2 Overall System Framework

(2) Functional Analysis of Profiling Prototype System

The personal and national postdoc profiling prototype systems are designed to visually display postdoc talent characteristics using a web front-end display technology through natural language processing (NLP) and statistical analysis of scattered data[5].

The personal profiling prototype system consists of eight sections, i.e., basic personal information, automatically extracted Chinese abstract, automatically extracted English abstract, annual number of papers published in recent five years, number of citations of published papers, names of research projects involved, types of papers published in recent five years, and 3D word cloud sphere. This system can display the above-mentioned information about the target postdoc, in which automatic Chinese-English abstracts are mainly automatically extracted via NLP from

collected papers in Chinese and English. The number, types and citation trend of papers published are mainly displayed by making query and statistical analysis into charts. By rotating 3D word cloud sphere clockwise, all extracted label information shows up.

Though the national postdoc profiling prototype system, some general information about nationwide grants for postdocs can be found on a yearly basis, mainly including seven sections, i.e. statistics on grants by region, statistics on funded postdocs by region, statistics on grants by region in recent five years, statistics on affiliations of funded postdocs by type, top 10 funded host institutions, top 10 funded postdoc majors, and detail maps on funded postdocs by province, municipality, autonomous region, and municipality directly under the Central Government.

The personal and national profiling prototype systems are still at the initial stage of displaying data gained from statistical analysis. Subsequently, they need further improvement in terms of real-time data linkage and profile update, etc.

(3) Implementation

In this paper, the personal and national profiling prototype systems are constructed by HTML (HyperText Markup Language), CSS (Cascading Style Sheet), JavaScript, Echarts and database technologies.

HTML, called HyperText Markup Language, is a markup language that contains a set of tags, in which tags are used to describe the contents of a webpage and combined into an HTML file, i.e., a web interface[6]. In this paper, the tag is used to divide the webpage into different partitions. Each of these partitions is further divided into sub-partitions by several tags. The entire webpage is partitioned into independent blocks in strict accordance with the design layout. For the personal and national profiling prototype systems, the webpage is divided into three blocks, i.e., left, middle and right, which are named div_left, div_middle and div_right respectively. div_left and div_right is partitioned into three subblocks, i.e., top, middle and bottom, which correspond to different types of statistical charts. div_middle is partitioned into two subblocks, i.e., top and bottom; the former is mainly used to store a descriptive text, and the latter is used to store word cloud images and maps of China. The innermost div mainly contains some basic tags in HTML, such as title, list, hyperlink, etc.

A div, called a block-level element, divides a webpage into sections. How to control the display style for each block should be solved with CSS. CSS is used to control the presentation of Web pages, including size, color and spacing. In the prototype system, each HTML page is linked to a CSS file. For a div block, webpage blocks are identified by setting id or class, in which id is used to identify specific page elements and class is used to identify a set of elements. For example, when six div blocks on the left and right sides of a webpage are used to store different types of charts, you can define a class named chart so that a chart style can be applied to all six div blocks at once by setting it in a CSS file.

JavaScript is a lightweight, interpreted and dynamic programming language[7]. It is dynamic in that a variable in JavaScript needs to be executed in an execution context, and the variable that has been assigned a value has the value defined. For the purpose of the prototype system, flexible.js, jquery.js, national_profile.js, personal_profile.js, china.js, postdoc_map.js and 3d_word_cloud.js are mainly quoted and compiled. flexible.js is mainly used for responsive web design, making webpages work on different screen sizes. jquery.js ensures that webpages are compatible with different browsers without adaptation, and also makes it easy to capture page elements and modify CSS[8]. personal_profile.js and national_profile.js mainly define block-level elements, functions and data, etc. for personal and national profiles, respectively. With this configuration, the webpage can render different types of charts based on to the data after loading. china.js is a file to define the map of China, which can display nationwide administrative boundaries and indicate province names. postdoc_map.js mainly defines the number of funded postdocs by province or city, which is

distinguished by colors in the legend. The quantity increases as the color gets darker. 3d_word_cloud.js defines the dynamic style of 3D word cloud sphere, including rotation speed, direction, loaded data source, etc.[9].

Echarts is a general big data visualization chart. Histogram, line chart, area chart and pie chart are mainly used for statistical analysis in the prototype system[10].

(4) Profiling Prototype System Application Presentation

Personal Profiling Prototype System

Taking the postdocs majoring in materials science and engineering in Peking University and majoring in management science and engineering in Tongji University as examples, the personal profiling prototype system displays their information about population attributes, research achievements and projects, as shown in Fig. 3 and Fig. 4. Chinese-English abstracts extracted by NLP are also displayed, and Chinese-English word cloud images are combined into a 3D word cloud sphere, providing a more holistic way to view word clouds by rotating the sphere. A histogram is used to count the number of papers published by the postdoc in recent five years. A time-dependent area map is used to statistically analyze the citation trend of papers. A pie chart is used to show the types of papers published, which are divided into four types: Chinese periodicals, Chinese proceedings, English periodicals and English proceedings. Text summaries and statistical charts can visually reflect the postdoc's actual research field, projects, research achievements and quality.

The comparison between the two personal profiles shows that more specific characteristics are extracted from the personal profile of the postdoc majoring in materials science and engineering. For example, the characteristics of the research field of the postdoc are reflected exactly by some technical terms. More abstract characteristics are extracted from the personal profile of the postdoc majoring in management science and engineering, so it is difficult to know the postdoc's specific research field without analyzing his/her major. The two disciplines publish papers mainly in English periodicals, indicating that the current postdoc talent evaluation attaches importance to the number of high-quality papers in English. According to the quantitative comparison of papers published in recent five years, the postdoc majoring in materials science and engineering made more achievements than the postdoc majoring in management science and engineering. The personal profiling prototype system implies the limitations in reflecting postdoc characteristics simply in terms of quantity and quality of achievements, making a system to evaluate all aspects of postdocs necessary.

National Profiling Prototype System

The national profiling prototype system displays statistics mainly based on the population attributes of funded postdocs nationwide in recent five years. Here are examples of the national profiling prototype system in 2019 and 2020, as shown in Fig. 5 and Fig. 6. The map shows the distribution of funded postdocs by province or city (except Hongkong, Macau and Taiwan), and the number of funded postdocs is distinguished by color, in which white represents 0 and dark brown represents more than 1,000. The distribution of funded postdocs shows more funded postdocs with higher quality in Beijing, Guangdong, Shaanxi and Jiangsu. The National Bureau of Statistics divides China into four regions: eastern, central, western and northeastern. The eastern region includes 10 provinces and municipalities such as Beijing, Tianjin and Shanghai; the central region includes 6 provinces such as Shanxi, Anhui and Jiangxi; the western region includes 12 provinces, municipalities and autonomous regions such as Inner Mongolia Autonomous Region, Guangxi Zhuang Autonomous Region and Ningxia Hui Autonomous Region; the northeastern region includes Liaoning, Jilin and Heilongjiang provinces.

The national profiling prototype system tracks funded postdocs by region. As can be seen in Fig. 5 and Fig. 6, the eastern region is economically developed with a larger number of provinces and

cities located there, so the number of funded postdocs seems to be more in evidence than other three regions, and that in the central and western regions is essentially flat. In order to analyze the trend of postdoc education quality by region in recent five years, the percentage of the number of funded postdocs by region to the total number of funded postdocs that year was statistically calculated. Taking 2020 National Profiling Prototype System as an example, the number of funded postdocs is fastest-growing in the eastern region, which increased from 63.47% in 2016 to 69.37% in 2020; that in the central region comes second, which increased from 11.42% to 13.68%; that in the western region shows a slight decline from 13.77% to 13.18%; and that in the northeast region shows a large decline from 8.43% to 6.71%. In general, the postdoc education quality in northeast and western regions still needs to be improved. The pie chart shows statistics on funded institutions by type, in which universities, companies, institutes and academies are listed in descending order by proportion, which shows that there are a lot of companies that have set up a postdoctoral workstation. Also, Top 10 funded majors and Top 10 funded institutions are statistically displayed. The Top 10 institutions include Sun Yat-sen University, Tsinghua University, Zhejiang University, Sichuan University, Xi'an Jiaotong University, Fudan University, University of Science and Technology of China, Shenzhen University, Peking University, and Shanghai Jiaotong University, all of which are universities, indicating that the postdoc education in China is still school-dominated. The Top 10 majors include biology, clinical medicine, materials science and engineering, chemistry, physics, mechanical engineering, environmental science and engineering, civil engineering, basic medicine and applied economics, which shows the main basic research disciplines.

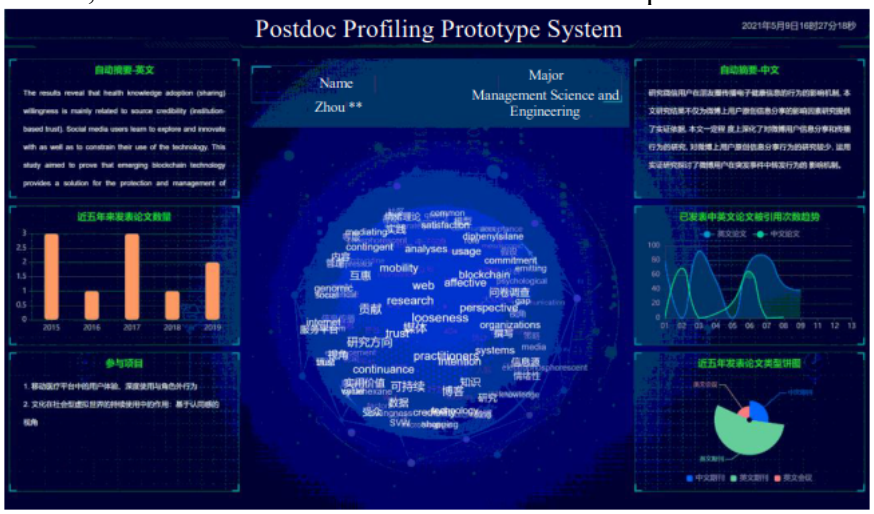

Fig. 3 Personal Profiling Prototype System Presentation (1)

By combining 2020 National Postdoc Profiling Prototype System with 2019 version, it can be found that the Top 10 funded majors and schools show only minor changes and generally remain unchanged, and the number of funded postdocs majored in biology, clinical medicine and materials science and engineering accounts for almost half of the total number of funded postdocs across all majors, which reflects the level of activity of these three majors in frontier research fields. Among the funded institutions, the proportion of postdocs in companies in 2020 has significantly increased compared with that in 2019, indicating that postdoctoral workstations in companies have stepped up to the first echelon of postdoc education in China.

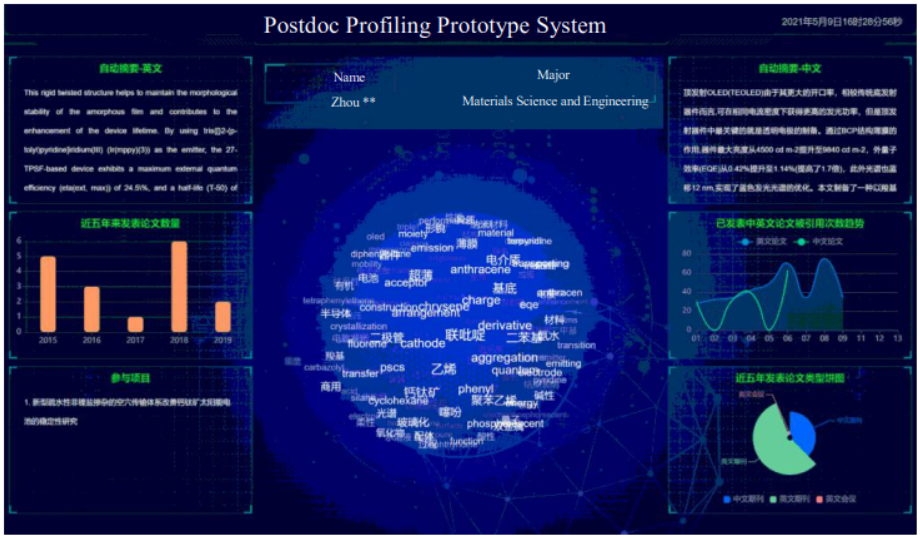

Fig. 4 Personal Profiling Prototype System Presentation (2)

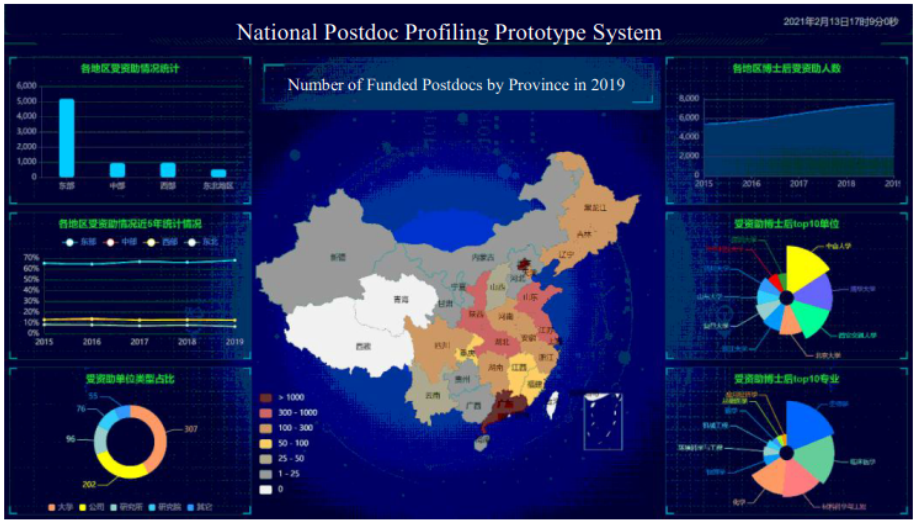

Fig. 5 2019 National Postdoc Profiling Prototype System

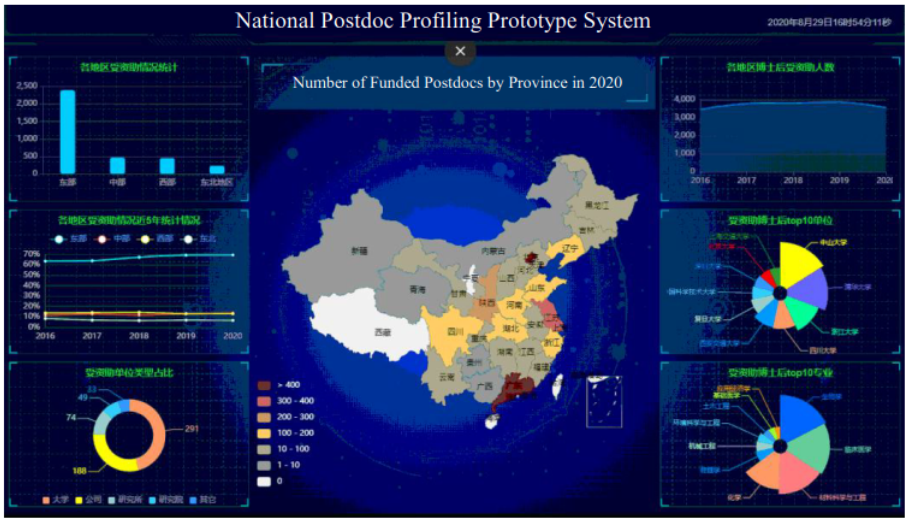

Fig. 6 2020 National Postdoc Profiling Prototype System

5. Conclusion

To address incomplete postdoc information, a postdoc talent database is established based on big data technology, and a postdoc profiling prototype system is designed to solve problems in subjects during postdoc talent evaluation, create postdoc talent data, promote multi-party collaboration on postdoc talent evaluation, enrich the diversity of subjects, and realize data sharing, interworking and integration. Firstly, the postdoc profiling framework is described. Secondly, the personal attributes, research achievements and projects of postdocs are tabularized, and the collected data are stored in a MySql database. Finally, the collected data are, by virtue of web page building technology and Echarts big data visualization charts, statistically analyzed from different perspectives to generate different types of charts, making postdoc characteristics and national postdoc education situation clear immediately.

References

[1] Wang Yuanzhuo, Jia Yantao, Liu Dawei, Jin Xiaolong and Cheng Xueqi. Open Web Knowledge Aided Information Search and Data Mining. Journal of Computer Research and Development, 2015, 52(2):456-474.

[2] Cooper A. The Inmates are Running the Asylum: Why High-tech Products Drive Us Crazy and How to Restore the Sanity. Sams Indianapolis, 2004.

[3] Li Wei, Xi Xiaotao et al. The Value, Foundation and Directions of Studies on Marketing Innovation in the Big Data Era. Science and Technology Management Research, 2014(18):181-184.

[4] Meng Wei, Wu Xuexia, Li Jing et al. Power User Potraits Based on Big Data Technology. Telecommunications Science, 2017(S1):23-28.

[5] Matei Zaharia, Reynold S. Xin, Patrick Wendell, et al. Apache spark: A Unified Engine for Big Data Processing. Communications of the Acm, 2016, 59(11):56-65.

[6] Shan Xiaohong, Zhang Xiaoyue and Liu Xiaoyan. Research on User Portrait Based on Online Review: Taking Ctrip Hotel as an Example. Information Studies: Theory & Application, 2018, 41(4):99-104,149.

[7] Eddy S R. Hidden Markov Models. Current Opinion in Structural Biology, 1996, 6(3): 361-365.

[8] Xia Jingbo, Wei Zekun, Fu Kai et al. Review of Research and Application on Hadoop in Cloud Computing. Computer Science, 2016, 43(11):6-11.

[9] Tseng H, Chang P-C, Andrew G, et al. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 2005.

[10] Hochreiter S, Schmidhuber J. Long Short-term Memory. Neural Computation, 1997, 9(8): 1735-1780.