International Journal of Frontiers in Engineering Technology
ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

Big Data Crawling and Mining Based on Internet Recruitment Website

Jinting Wan, Fuqiang Wang*,Zhaoling Li and Hao Zhang

School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266000, China
**Corresponding author e-mail: wffw22@163.com*

ABSTRACT. *The recruitment information of computer related majors on the Internet recruitment website in 2019 has been crawled by means of data processing with the web crawler program in this paper mainly and the Visualization data has been presented. By using data analysis technology, the related needs and working conditions of big data industry and computer industry are displayed by visual means, which provides information support for college education and teaching, employment choice of college students and docking between schools and enterprises.*

KEYWORDS: *Big Data; Crawing; Mining; Internet Websit; Recruitment*

1. Introduction

With the operation of information technology in all aspects of national governance and economic operation, a large amount of data is produced, and the explosive development of Internet technology has made the total amount of data produced in the past two years exceed the total amount of data recorded in human history. In 2017, the development of the big data industry was written into the government work report, and big data began to appear not only in the strategy of enterprises, but also in the planning of the government, which can be said to be the darling of the Internet world.[1]

At present, many governments and international organizations have recognized the important role of big data, and have taken the development and use of big data as an important grasp to seize the commanding point of a new round of competition, and implemented the big data strategy. Big data industry development has a high of enthusiasm. Using big data to analyze related industries, mining and analyzing big data through big data technology, combining with the field of computer industry, it plays a reference role in reality and excavates the deep value of data, and is the future information science and technology development trend.

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

The intelligent recommendation for the analysis and statistics of recruitment information data is mainly "Tiger Employment Network ", but the data management is difficult because the update is not timely, and the data efficiency is low because of the traffic problem of network users[2]. Therefore, through the massive data mining of recruitment information, this paper uses Python and big data technology[3] to crawl effective data to merge and analyze, put forward suggestions and make appropriate predictions, and provide information support for college students' campus learning and technical mastery. Let more college students understand the trend of talent demanded, thereby to improve the efficiency of job seekers and recruiters.

2. Research Methodology

In this paper, a practical software platform (system) is designed and implemented to realize data visualization and provide valuable information for college students, such as: technical needs, hot areas, salary distribution and so on.

The research method mainly adopts the big data crawling of the recruitment website, the visual display of the data after cleaning, and the literature research, the case data analysis, the observation comparison and so on. In order to maintain the validity and timeliness of the data, the automatic updating strategy is adopted for the data crawling of the recruitment website, and the crawling time interval is set to ensure the reference of the data mining analysis.

2.1 Data Sources For Research

The Internet recruitment website is characterized by a large amount of data, many positions, open compensation and unlimited areas, etc. It can provide one-stop professional human resources services to large companies and rapidly developing small and medium-sized enterprises, including online recruitment, newspaper recruitment, campus recruitment, headhunting services, recruitment outsourcing, enterprise training and talent evaluation. Today, one of the Internet recruitment sites, Zhaopin.com, has an average daily view of more than 100 million people, with a daily average of more than 2.55 million online positions, with a large amount of time-limited data to crawl.

At present, the existing domestic authoritative recruitment websites are mainly pull hook network, Boss direct recruitment, direct recruitment,51 job and so on, as shown in table 1[4]. According to research data, 51 job has the highest share, reaching 30% of the total network recruitment market, while Zhaopin recruitment is the second, accounting for 25% of the total in 2018. These websites are characterized by timely updating of recruitment data, large user views and comprehensive recruitment information. Considering the above factors and the degree of anti-climbing, this paper chooses 51 job website as the target to crawl.[5].

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

Table. 1 Main recruitment websites in China

Recruitment website	Web site	Revenue
Lagou web site	https://www.lagou.com	5%
BossDirect employment	https://www.zhipin.com	20%
Zhaopin recruitment	https://www.zhaopin.com	25%

2.2 Research on Technical Routes

Recruitment information massive data crawling is mainly divided into web crawler and data visualization analysis module. Crawling process: first select 51 job website as the main crawling website of massive data, write algorithm to realize data crawling, then carry on preliminary screening and cleaning to crawling data, store the cleaned effective data in the designed database, Then use visualization tools and related functions to realize data visualization. Finally, data analysis or prediction can be made moderately, and the flow chart is shown in Figure 1.

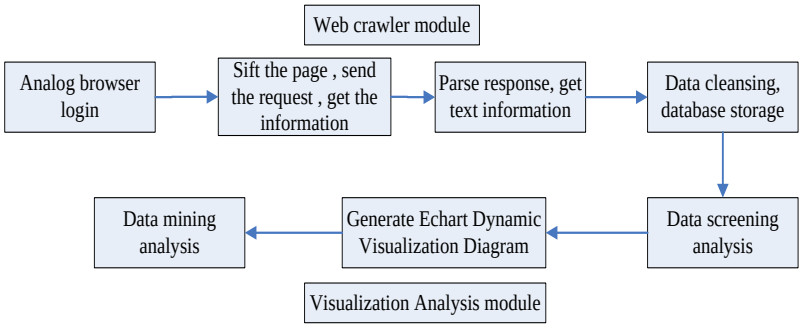

Figure.1 Flowchart for Mass Data Research

3. Research Process

3.1 Data Crawling and Preprocessing

In this paper, big data information acquisition is mainly to achieve the 51 job website content acquisition, data screening and storage, the main research steps are as follows.

(1) The list in a web page is traversed by writing a script to obtain the requested html content and related content using a BeautifulSoup,urllib library.

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

(2) Screening data according to requirements. First, the appropriate function method is selected for the preliminary filtering of a string, and then the string is segmented and reorganized to select the required information. For example, change the crawling information "average salary/year" to "average salary/month ".

(3) Create a temporary POJO class, create new objects, place objects in a List container, connect to the database, write all the data in the List to the database, and submit their own information to the database using the event processing mechanism.

(4) The last filtered data is stored asynchronously in the Mysql database.

Effective data in this paper mainly crawl "Java"、" artificial intelligence "," big data "," algorithm" four categories related to computer related majors in Shanghai and other 15 major domestic cities salary and demand data.

3.2 Design of Data Sheets

The design of database table needs to reflect a total industry demand around relevant parameters and urban information location. This paper uses relational mysql database and navicat database management tools to realize data storage. The data table in this paper mainly includes JavaCollege (Java undergraduate demand), JavaMaster (Java master graduate demand), Bigdata(big data industry demand) and AI(artificial intelligence industry demand) according to the demand analysis and professional characteristics. Bigdata table property fields include id, position, address, minWage, maxWage, id of them as primary key, Data types for other properties are set to char types, minWage and maxWage, respectively.

3.3 Data Visualization

The data visualization is displayed by Javaweb engineering. The file structure mainly includes the mapper layer of extracting data and the pojo layer of temporary storage data, the business logic layer of the middle service, and the servlet layer of responding to the data request.

1)Mapper layer: the data acquisition is realized by mybatis framework, which encapsulates the JDBCutil, application according to the database connection program and related sql statements.

2)service layer: service layer is the data processing layer, such as processing the distribution data of each city in each industry, the data can be placed in a defined class or container, and finally returned to the servlet. The servlet will be returned through the response function.

3)servlet layer: its function is a coordinated, content return role.

4) front end: the front end uses the jQuery frame and the echarts frame to realize the data chart effect display.

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

3.4 Data Mining

1)Beyond 200 Urban Information Tables

Now the following description is an example in Table 2.

Table. 2 City information table of beyond 200 recruitment information release

Index	1	2	3	4	5	6	7	8	9
Addr	Shang hai	Shen zhen	Guang zhou	Bei jing	Nan jing	Hang zhou	Cheng du	Wu han	Su zhou
Count	721	650	485	306	296	260	222	222	129

After crawling massive data, it is found that Java related positions are published in 275 cities in the country, of which 18 are in more than 50 cities and 10 in more than 100 cities, all of which are first-tier cities. In cities with job requirements exceeding 120, North Guangzhou-Shenzhen demands ranked the top four, Shanghai's highest demand reached 721 and the lowest demand is in Suzhou in Table I. City information table of beyond 200 recruitment information release. The following Figure 2 regional job demands distribution map shows the top 50 cities for visual display, the demand is more eastern coastal cities.

Figure. 2 Regional position demand distribution

2) Industry Demand Comparison

Industry demands comparison selected Beijing, Shanghai, Shenzhen, Guangzhou and other four China's computer industry technology and the fastest-growing areas to do data screening mining analysis. Through the above Figure 3, it can be found that the computer industry as a whole has the highest demand for Java、 big data, the second for python and algorithms, and the least for artificial intelligence. According to the multi-dimensional analysis of different cities, it is found that the demand for artificial intelligence, Java、 big data and algorithms in Shanghai is

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

higher than that in the other three cities, and the proportion of talent demands in four different fields is as high as 30% or more.

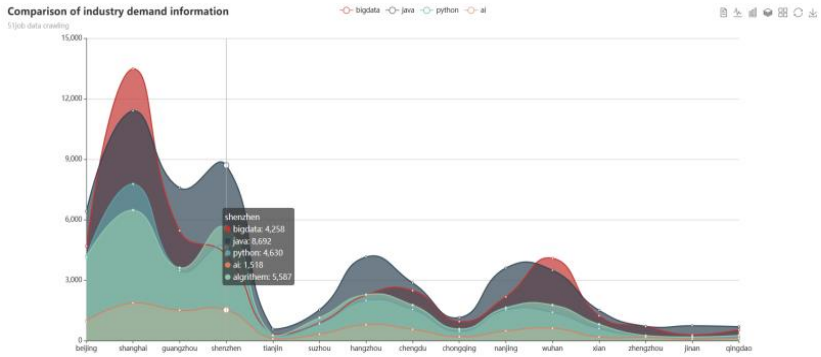

Figure. 3 The Comparison of Industry Demand Information

3)Salary distribution

The ultimate factor in recruitment is the salary situation, which is also the symbol of the value of labor force, the degree of attention to talents and the degree of prosperity of regional development. For the study of salary distribution, Beijing, Shanghai, Guangzhou, Shenzhen and Nanjing were selected as examples in Figure 4, and 20 records were extracted in these five cities, and the minimum wage, maximum wage and average wage of Java positions were used for technical analysis.

The highest salary for java positions in Shenzhen is 37 K, much higher than in other cities, while Guangzhou's minimum wage for Java positions is less than 10 K; Shanghai is the highest in terms of average salary, so Shenzhen and Shanghai are more selective and cost-effective if they selective and cost-effective.

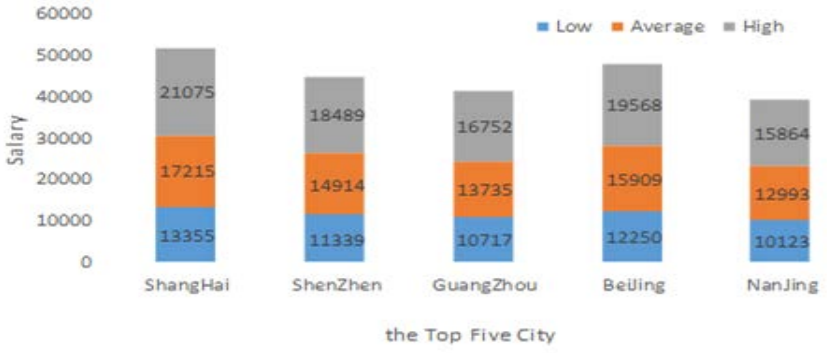

Figure. 4 The Salary distribution of the top five cities

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

4)Education Statistics

Education is one of the important indexes for enterprises to recruit talents. Up to now, China has basically entered the stage of popularization of high-grade education. The data shows that the number of undergraduate graduates has exceeded 8 million per year, and the pressure of employment competition for graduates is getting higher and higher, but high education has absolute competitiveness in job search, so the number of graduate students who choose to study master's degree is increasing every year. Through the recruitment information top20 of the city to the Java position education data crawling found: undergraduate and college education is the main goal of enterprise recruitment, and the proportion of undergraduate is twice as high as 66.6%. Therefore, If you don't have a college degree or above, you wo n' t be able to apply for a position in the software java related industries.

5)Graduate/undergraduate pay ratio

In order to facilitate the vertical and horizontal comparison and multi-dimensional analysis of the overall data, the data mining research is carried out in Beijing, Shanghai, Guangzhou and Shenzhen, the top four first-tier cities. Among them, through Figure 5 below, Figure 6 graduate / undergraduate students in artificial intelligence major and Java major comparison thermometer chart to show master's / undergraduate industry salary comparison.

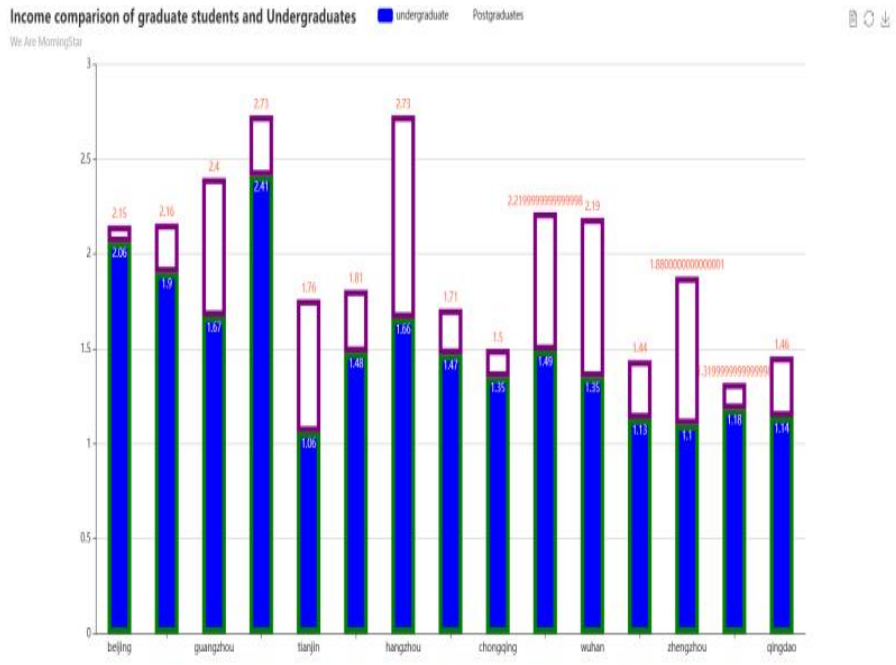

Figure. 5 Postgraduate/undergraduate artificial intelligence thermomete

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

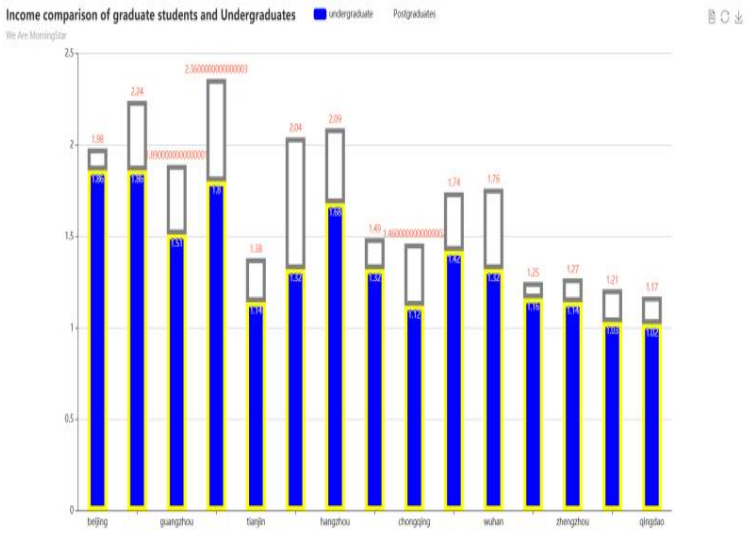

Figure. 6 Chart of Java Thermometers for Graduate/Benish Students

From the chart, it is clear that whether the artificial intelligence major or the Java major, the overall salary of graduate students is a lot higher than that of undergraduate students, which is the output benefit of higher education and higher education. The salary gap between graduate students and undergraduates in artificial intelligence industry is larger than that in Java industry, for example, the salary gap in Hangzhou is as high as 10 K yuan.

6) Type of company

Different types of companies represent different business models and different ways of working. Among them, the management and working methods of private enterprises are more free and changeable than state-owned enterprises. Choosing the right position also needs to start with the type of company. Large data crawling found that the demand of private enterprises in the software Java industry is more than 70%. The reason is that in recent years, small and medium-sized private software enterprises have developed rapidly in the software market. They are attached to the rapid development of China's policy for small and medium-sized private enterprises and gradually become the main body of technological innovation and progress, and have made great contributions to the growth of national economy and labor employment. As a result, private enterprises are very suitable for the development of java related job seekers.

7) Welfare Words

The welfare provided by the company to the staff is one of the important indicators for job seekers to choose a job. Employee welfare is an important part of

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

salary management [6] and an important part of human resource HR management. Reasonable benefits have a vital impact on attracting and retaining talented people. Through the extraction of welfare words, it is found that five risks and one gold, mid-year bonus, performance bonus, regular physical examination are the benefits that most enterprises can provide to job seekers, and are the most basic welfare policies. On the contrary, the team building activities, no overtime, etc., which are the most concerned of job seekers, show that there is still a lot of room for improvement in the follow-up of enterprise welfare.

4. Major Technologies: Data crawling, Cleaning and Storage

4.1 Data Crawling

After determining the target website, the technology needs to clarify the website request process from the level. Information retrieval, page turning processing, detail page information analysis are the necessary steps of data crawling, all need to write programs to achieve its functions.

The crawler entry is implemented by the start_requests function, passing the retrieval conditions and obtaining the home page information and the next page link, and returning the get_next.

The get_next function mainly realizes page turning function, until the next page link is not available, all retrieval data are crawled. During this period, the crawling list page information including position name, company name, work place and details page link are passed to parse. detail page parsing function

Parse function is to achieve detail page information parsing. Analysis of specific fields uses xpath statements, including salary, company size, company type, work experience, benefits, job description and other information. Pack all the parsed data fields into a data dictionary, and each recruitment record will be saved in a dictionary, and the yield will submit the data to the item as a generator for subsequent processing.

Item.py writing: Item declare using a simple class definition syntax as well as sciapy.Field objects to indicate the metadata for each field.

4.2 Data Cleaning

The data cleaning processing part mainly uses the Numpy based pandas tool[7]. The main functions of Pandas tools are: generating data tables, viewing information, cleaning data tables, data preprocessing, extracting data, summarizing, filtering, statistics and output. A large number of available functions and methods are the reasons why Python become one of the efficient data analysis environments, and the

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

analyzed data can be output into xlsx format and csv format. Data cleaning is mainly to find and correct the identifiable errors in the data file, including checking the consistency of the data, processing invalid values and missing values. Because massive data crawling contains a large number of Chinese characters, in order to be consistent and complete in the subsequent statistical work, the process_money function can be used to convert all of them into digital format. The program implementation of data cleansing payroll data items is showed in Figure 7.

```
def process_money(self,m):
    money = 0
    try:
        meonyList = m.split("/")
        if meonyList[1] == " month":
            v = meonyList[0][-1]
            lowStr, highStr = meonyList[0][:-1].split('-')
            if v == "thousand":
                lowMoney = int(float(lowStr) * 1000)
                highMoney = int(float(highStr) * 1000)
                money = "{}-{}".format(lowMoney, highMoney)
            elif v == "tenthousand ":
                lowMoney = int(float(lowStr) * 10000)
                highMoney = int(float(highStr) * 10000)
                money = "{}-{}".format(lowMoney, highMoney)
        else:
            lowStr, highStr = meonyList[0][:-1].split('-')
            lowMoney = int(float(lowStr) * 10000 / 12)
            highMoney = int(float(highStr) * 10000 / 12)
            money = "{}-{}".format(lowMoney, highMoney)
    except:
        pass
    return money
```

Figure. 7 Data cleaning

4.3 Data Storage

In order to ensure the integrity of data preservation, the data is stored asynchronously. Asynchronous access database can separate crawler and write database operations, do not affect each other, especially write faster, its asynchronous storage database code as shown in figure 8.

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

```python
class MysqlTwistedPipline(object):
    def __init__(self, dbpool):
        self.dbpool = dbpool

    @classmethod
    def from_settings(cls, settings):
        dbparms = dict(
            host = settings["MYSQL_HOST"],
            db = settings["MYSQL_DBNAME"],
            user = settings["MYSQL_USER"],
            passwd = settings["MYSQL_PASSWORD"],
            charset='utf8',
            cursorclass=pymysql.cursors.DictCursor,
            use_unicode=True,
        )
        dbpool = adbapi.ConnectionPool("pymysql", **dbparms)

        return cls(dbpool)

    def process_item(self, item, spider):

        query = self.dbpool.runInteraction(self.do_insert, item)
        query.addErrback(self.handle_error, item, spider)
        return item

    def handle_error(self, failure, item, spider):

        print (failure)

    def do_insert(self, cursor, item):
        insert_sql, params = item.get_insert_sql()
```

Figure. 8 Code for asynchronous storage of databases

5. Conclusion

Through the collection and data processing and analysis of recruitment information, this paper can provide data support for professional, employment and entrepreneurial choices for relevant industries, some colleges and universities, and students choosing jobs. Through the related crawler technology, the big data information is crawled and stored asynchronously to the database. Through the data mining, the multi-dimensional analysis of the computer specialty in recent years in the recruitment information bias and the future career guidance. At the same time, the phenomenon of "hot industry and scarce talents "[8-10] will still appear in the future, and the demand for some computer professionals will continue to expand, which is also a direction of computer professional development in the next few years.

Big data crawling and mining based on recruitment website provides a certain bias for computer major in talent training and employment application, but it is also necessary to clearly realize that the massive data of recruitment information is changing. With the attention and support of big data development, more scientific big data mining and analysis algorithms will emerge. So no matter how the career choice, self-construction can be in the workplace.

ISSN 2706-655X Vol.2, Issue 3: 36-47, DOI: 10.25236/IJFET.2020.020304

References

[1] http://language.chinadaily.com.cn/2017-03/17/content_28591593.htm
[2] Liu Ying, Wen Zhengjian(2018). Influencing Factors and Realization Paths of High Quality Employment for College Graduates with Bachelor's[J]. Degree Employment of Chinese University Students , vo.10,p.34-38
[3] Zhang Yan, Wu Yuquan(2020). Python - based web data crawler program design [J]; and Computer programming skills and maintenance, no.4,p.26-27
[4] Bi Yu Chen,Mei-Po Kwan(2020). Special Issue on Spatiotemporal Big Data Analytics for Transportation Applications. Transportmetrica A: Transport Science[J],no.1.
[5] Hongmei Qing(2019).Analysis on the Demand of Japanese Talents in Higher Vocational Colleges in Pearl River Delta - Based on the Information of Japanese Talents Recruitment[J]. South Vocational Education Journal, no.9,vol.6,p.27-32.
[6] Saloni, Silvia(2015). Impacts of multiple stressors during the establishment of fouling assemblages. Marine Pollution Bulletin[J] ,v. 91,p.211-221
[7] Song Y(2020). FeAture Explorer (FAE): A tool for developing and comparing radiomics models. PLOS ONE,vol.8
[8] Wang, Danxin(2020). A general location-authentication based secure participant recruitment scheme for vehicular crowdsensing. Computer Networks[J]:, v 171
[9] van Esch, Patrick(2019). Marketing AI recruitment: The next phase in job application and selection.Computers in Human Behavior[J],v.90, p.215-222
[10] Yuhao Fan(2018). Design and Implementation of Distributed Crawler System Based on Scrapy. IOP Conference Series: Earth and Environmental Science(CA), vol.108