# Discussions on the Realization Technologies Of Web Archiving and service Platform of National Library of China

**Ji Shiyan**
*National Library of China, China*
jisy@nlc.cn
*0000-0003-3486-6702*

**Zhao Danyang**
*National Library of China, China*
zhaody@nlc.cn
*0000-0002-8901-3590*

**Abstract – As an important component of china's public culture system, the National Library of China(NLC) web archiving program started in 2003. Based on the open source software (OSS) Heritrix, NLC started project of collecting, cataloguing and archiving the government public information , important websites and webpages at home and abroad in 2005. The NLC accumulated abundant practical experience and united libraries nationwide to carry out web archiving and service jointly. In 2018, NLC carried out technology upgrade and developed a set of "Web Archiving and Service Platform" for distributed cloud storage infrastructure. The platform adopts a distributed cloud infrastructure and supports the management and use of at least 1 million metadata data, which enables the NLC to collaborate with multiple libraries (institutes) to conduct web collection services. This paper analyzes the construction ideas, technical routes and key technologies in detail based on the analysis the strategy of web archiving and the requirements for the system platform. It is hoped to provide reference for other institutes to carry out the related work.**

**Keywords – National Library of China, Distributed system, Web archiving and service platform, Web archive**

## I. The objective of National Library of China Network Resources Preservation

### A. *Process of Project Development*

As an important part of the national public cultural service system, the National Library of China(NLC) has always attached importance to the preservation and service of web information.As early as in 2003, the National Library of China of China developed the project of Web Information Acquisition and Preservation(WICP)[1][2] and experimentally began to collect and preserve the web information in China. In 2007, the National Library of China formally joined and became a member of the IIPC(International Internet Preservation Consortium). The National Library of China has developed extensive exchange and cooperation under the IIPC framework, carried out web information preservation of china based on the international common standards to promote the process of internationalization and standardization. In 2009, National Library of China's Web Information Preservation and Protection Center was established, which was committed to the long-term preservation and protection of web information in China.NLC's website service was launched in 2012 to display the collected web information. Since 2014, it has united libraries nationwide to carry out web archiving and service jointly.In 2018, NLC carried out technology upgrate and developed a set of "web archiving and service platform" and realized standardized collecting, cataloging, wayback and service. . It has accumulatively archived more than 23,000 websites by the end of 2018, covering the government public information and important websites and webpages at home and abroad.

### B. *NLC's Strategies of Web Archiving*

In the process of large-scale acquisition, we should not only focus on the acquisition of web information, but also on prevent the accumulation of garbage information;we should also pay attention to the authority and preservation value while ensuring the comprehensiveness of web archiveing [3]. Focusing on the function orientation and service requirement of public library, theNLC adopts the mode of comprehensive acquisitionand key acquisition to realize the effective acquisitionof important web information in different fields and scope. At the same time, the continuity of the acquisition content and the effective of the collected data are fully considered in the acquisition process, and the acquisition range of the network resources is constantly revised.

Specifically, the NLC's web archiving mode is divided into two types: comprehensive archiving and thematic archiving. Comprehensive archivingis also the whole station archiving, which is the archive and preservation of all the object data of the website in the domain name. The consideration of artificial intervention, quantitative and qualitative is an essential element in the selection of the acquisition site and web information. The NLC's main responsibility is to preserve the Chinese digital memory, and select the excellent network culture resources with important preservation value for archive.

The government website is the main body of our archiving. Under the promotion and support of the «Government Information Disclosure Regulations» and the extension project of the digital library, the NLC unites the public libraries at all levels of china to carry out comprehensively archiving of governmentwebsites,covering the central level, the provincial level, the local level, the county level, the town and the following. The NLC is responsible for the archives of the central level websites, and the public libraries at all levels are responsible for archiving the respective regional government websites. For many years the archived resources provide a good resource guarantee for the public and the scholars' research, and also faithfully record the reform process of the Chinese government's institutions and provide reference for national decision-making.

Thematic archiving is to collect and preservate the related topics'web information rouding the important field and social hot spots of China, and provides the more and deeper valuable integrated information for social public. The archived resources include the thematic columns, related web pages and other types of network resources. The construction of thematic resources is an important way for digital libraries to provide resource services under the new environment. It not only lies in the preservation and inheritance of digital heritage, but also provides knowledge services after data integration for the society. At present, the NLC constructes about 60 thematic databases every year and several key themes. The service experience can be improved through the service model of technological innovation such as resource clustering, visualization and correlation analysis.

With the data capacity over 210TB, the NLC accumulatively archived more than 20,000 government websites, 2,319 national websites, 5,583 foreign websites and 276 special topics by the end of 2018. The archived web information have become an important part of the library's digital resources construction and provide important support for government decision-making, scientific investigation and the public.

## II. Web Archiving System of National Library of China

In order to meet the challenge of the web information acquisition and preservation, the NLC has initially formed a relatively complete web archiving system and effective service mechanism. The long-term, scientific and sustainable development of network resources preservation has been  realized though the standard specification, technology application, platform construction, data mining and analysis, user service experience, division of labor and cooperation.

### *A. Formulating metadata Standards for web Archiving and strengthening Resource Integration and Disclosure*

Compared with traditional digital resources, web information have many unique properties in classification, structure, representation, data characteristics and so on. The collecting and archiving web information should be described by the standardized language of objectivity and integrity for further long-term preservation and deep data mining. On the basis of researching the digital resources ' metadata specification, the NLC has established a set of relatively perfect metadata specification of web archives to realize the standardized description of website resources and thematic resources. And the metadata standard of web archives of "Digital Library Promotion Project" has been formulated in order to promote the construction and sharing of network resources. Chinese government public information online is jointly constructed by the NLCand 230 public libraries all of the country. All the libraries collect and integrate government public information of their respective administrative regions with unified metadata standards and norms. The collected web archives are orderly integrated and uniformly published.

### *B. United all public libraries to develop the web Archiving Business*

The NLC actively promotes the popularization of knowledge and business promotion of the web archiving, united all public libraries structure a hierarchical web archiving system covering the whole country with the mode of "unified dispatch, division of labor acquisition, centralized index and decentralized preservation".

The NLC developed"Internet Storage Project" all participating libraries collected the important websites and hot topic resources covering the local political, economic and cultural development. The whole websites acquisition mainly focuses on the local people's government and local organizations, the thematic resource acquisition centers on the major local cultural events, local folk customs, local cultural protection and so on. The number of libraries participating in this project is increasing year by year. In 2014, five provincial public libraries became the first joint members including the Capital Library, the Hubei Provincial Library, the Zhejiang Library, the Jilin Library and the Xinjiang Construction Corps Library . By 2018, the number of declared libraries has reached 115.

### *C. Building a Distributed Cloud Storage Management Platform to Achieve Collaborative Development of Multi-agency Business*

The NLC's web archiving based on the open source software (OSS) Heritrix [4] [5] has been collected, cataloged and preserved since 2005. The NLC has been tracking the development of technology in the field of web archiving and storage, especially focusing on storarge technology, distributed structure and cloud service architecture due to the increase ofweb information, the improvement of business efficiency and the effective use of storage space. It realizes the functional improvement of web archiving and preservation platform by developing its own software program. With the increase of joining libraries, a standardized, open and shared software platform is needed to meet the business needs of libraries and the different basic hardware environments, to support multiple libraries carry out web archiving works based on the same software platform and jointly promote the development of China's web archiving cause.

In 2018, NLC carried out technology upgrade and developed a set of "web archiving and service platform" for distributed cloud storage infrastructure.

The platform is developed based on open source software Heritrix to realize the standardized management and visual operation of web information preservation and service business and improve the work efficiency. The achived web information are stored in the distributed file system with the format of WARC [6], the qualified data can be used for further long-term preservation and used as the publishing level data for the front-end display of the website. The platform's data processing capabilities support the management and use of at least 1 million metadata data. According to the business needs, the platform is designed with eight functional modules (acquisition, cataloging, wayback, content management and release, data preservation, statistics, user management, maintenance and backup) for modular, lightweight deployment and service.

The platform adopts a distributed cloud infrastructure, which enables the NLC to collaborate with multiple libraries (institutes) to conduct web archiving services. As the central node of the platform, the NLC has the right of unified operations and management. As the parallel nodes, other libraries can install the functional modules in a customized way according to the business needs of their own. Each library can rely on its own local servers storages and network for web archiving; it can also use the network to realize the unified storage to central node of metadate.

The platform has a relatively perfect user's authority control mechanism to make the division of labor clear and ensure the every library 's resources and operations be undisturbed. As the center of the platform and the information aggregation point, the NLC manages the task of web information preservation for all nodes in the country and schedules information preservation related business based on web information preservation task with each time and each node.

## III. The Technical Realization of the web Archiving and service Platform

The platform has the characteristics of automation, distribution and cloud management which can meet the business characteristics and business management requirements. In the aspect of the technical realization of the platform, the personalized transformation is carried out under the acquisition framework of IIPC [7].

### A. *Thoughts on Platform Construction*

#### 1. *Automation, process and Modularization*

In the past, such as the deployment of the seed link, the summarize of the acquisition data, the establishment of the index file and the quality inspection of the release link, all of the operations required repeatability manual operations and interventions by staff. With the sustainable development of business, a significant increase in the amount of data needed to be processed, manual operation has been unable to meet the needs of business development. In addition, there are many differences in the level of computer operation among the staff of different libraries nationwide, and some staff do not have the computer operation skill. So, there are great difficulties in the archiving and preservation business. It seriously impedes the promotion and development of the web archiving and preservation business of librariesnationwide.

The platform should take the operator as the center to reduce the difficulty of the business operation, standardize the business process and improve the efficiency of the business. In addition, the complete process of web archiving can be divided into several reasonable, interrelated and independent business modules in the form of modularization. Staff who do not have in-depth computer knowledge can operate and complete their work by the visual operating interface of the platform.

#### 2. *Highly available, efficient, and scalable hardware architecture*

Formerly, we use multiple servers to complete the work of the web archiving. Each server deployed a Heritrix instance to run the Heritrix archiving process and completed the designated archiving task. It depends on the processing performance of the server and the number of the running Heritrix instances to determine the efficiency and frequency of the archiving. When taking on different archiving-tasks, the server's workload is quite different whike taking on different archiving tasks. After a period of monitoring, we have found that part of the server memory, such as CPU, and network bandwidth been occupied in the actual web iarchiving business. But, at the same time, there are still some servers in the idle state. It can not make full use of all hardware resources and more or less waste the resourceof network bandwidth and storage space utilization.

Therefore, it is necessary to integrate the resources of all the servers according to the actual situation of the archiving business to dynamically adjust the tasks of server configuration and acquisition service. Besides, it is necessary to give full play to the overall performance of the server cluster and deploy the server hardware, network and storage resources in a virtualized and distributed mode.

#### 3. *Distributed, shareable and scalable Storage mode*

In the past business, the archived data is saved as a file in warc format and compressed into "gz" format for final preservation. All the archived resources are stored in the online storage space of SAN. It not only limited the maximum amount of network resources captured by Heritrix software in a single acquisition, but also lead to the imbalance of storage space. During the operation of the Heritrix instance, the SAN is exclusively used by the physical server. When the capacity of the collected web information is close to or exceeds the storage space allocated by its server, the Heritrix process will be affected or even interrupted. In addition, a single server only enjoys its own storage network space and does not support server virtualization and Heritrix multi-threading virtual archiving mode.

Therefore, it is necessary to form a "cloud storage pool" that can be shared to provide reliable data storage services for national libraries and other institutions on the platform. In addition, it is necessary to expand storage space, schedule storage allocation strategy according to business requirements and then realize effective support for archiving services.

### B. *The Technical Route of choosing IIPC Open Source Software on the platform*

Due to the need to improve the automation of the original archiving method and to keep the massive data continuously, it is necessary to construct the distributed hardware architecture, the cluster batch archiving mode and the shared storage space management based on the platform. The platform adopts Heritrix, Wayback [8] and other basic archiving process framework of IIPC to guarantee the integrity and specification of the whole process of web information archiving. We revamp and develop a number of personalization features on the basis of the IIPC framework and take full advantages of open source tools.

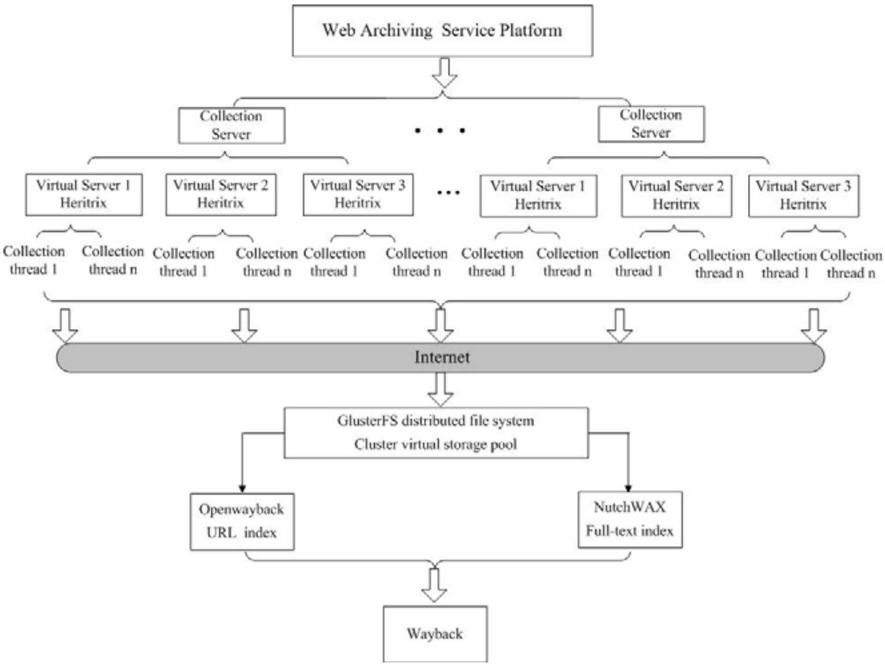

Figure 1: The Framework based on IIPC tool

We developed the administrator of a B/S mode on the basic framework to manage the distribution of local collection tasks and the automatic scheduling of collection resources,at the same time to monitor and manage the collection task of different organizations on the platform.

While expanding the acquisition server, we use virtualization technology to form a server cluster, deploy multi-node of Heritrix and apply the multi-threading collection characteristics of Heritrix to form a massive and distributed data acquisitionmodule .

We optimize the procedure of administration platform and invoke the Heritrix interface to deploy-batch collection task. We use a distributed file system (DFS) of GlusterFS as the unified organization and storage container of Heritrix collection data. We use OpenWayback and NutchWAX [9] modules to proceed URL index and full-text of the archived data stored in DFS and solve the problem that manual participation in index operation in the traditional architecture.

### C. *Implementation Technology of distributed platform Architecture based on Hadoop*

The overall framework of Web Archiving and Service Platform is a hierarchical cloud infrastructure. For ordinary users, they can acquire the resource services of the platform through the " service platform "; for business people, they control and manage the complete process of collecting and saving web archiving through "administration platform"."service platform" and "administration platform" data and requirements through standard interface and they are transparent to each other. The " administration platform" manages the complete process of online resources collection and storage for all organizations and it can realize hierarchical resource sharing and integrated management in the light of the step-by-step cloud storage infrastructure.

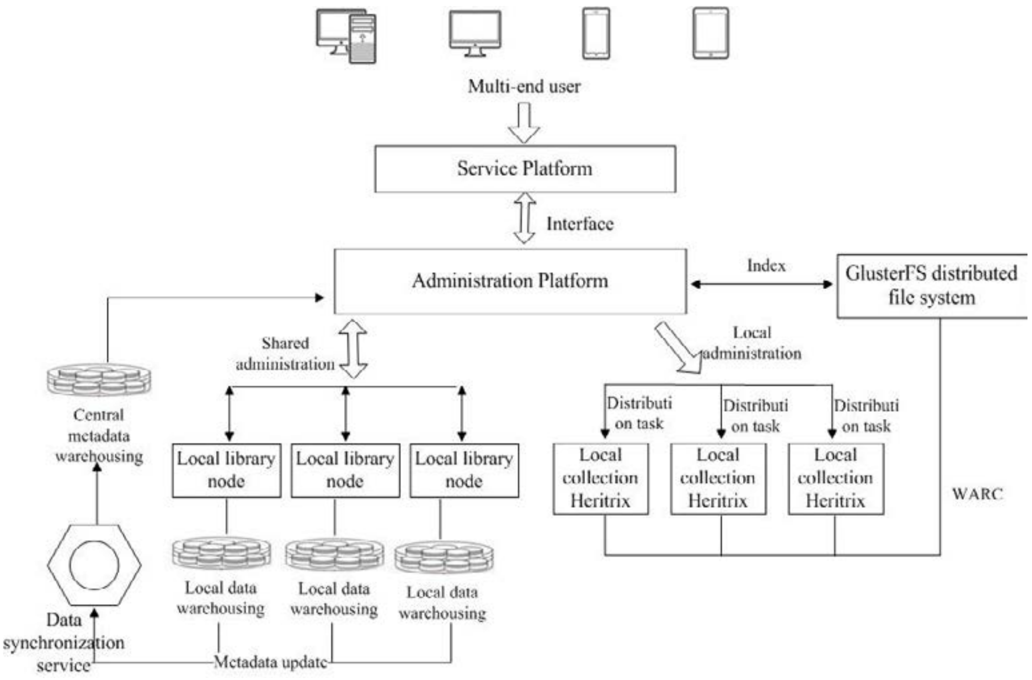

Figure 2: The Architecture of the web information Preservation and Service platform

1. *The Technical Implementation of Distributed Acquisition and Storage*

(1) Implementation of Unified scheduling

The “ administration platform “ is deployed on a master server and the acquisition nodes are deployed on a plurality of servers. Through the communication feedback established with the acquisition server nodes, the administration platform uniformly configures and schedules the acquisition tasks among the acquisition nodes and multiple acquisition threads on the nodes. It is responsible for the scheduling and management of the task queue composed of all the completed tasks to be collected, allocating the free resources of the acquisition node and dispatching the task to be collected for crawling.

(2) Implementation of Distributed Acquisition

With distributed storage architecture, the platform can parallel expansion increase physical servers. Each physical server extends the heritrix collection node with Virtualization Technology Taking advantage of the characteristics of heritrix multithreading, one or more URLs parallelly runs in multiple threads, it makes full use of the performance of the acquisition server and improves the acquisition efficiency. In order to ensure the high efficiency and unity of distributed acquisition, the platform should monitor the execution of all acquisition tasks and balance the network and hardware resource efficiency of each acquisition node for scheduling heritrix to carry out the task capture.

(3) Implementation of distributed Storage and Servic

In order to realize the unified archiving and effective index of collected data from plurality nodes, GlusterFS DFS[10]which has high availability, high performance, strong expansibility and low hardware performance is used to realize the storage management of data.In the application of GlusterFS distributed technology, the platform abandons the metadata server to eliminate the single point failure of the whole system and improve the parallel access speed of the data. It stores the metadata using mysql as the database service and accelerates the convergence rate of the algorithm by the OMP under the same accuracy requirement. According to elastic HASH algorithm, the archived data is stored in each file. Each storage node has the same directory structure and the HASHvalue of data in each node is 66635 / node. The number of the first node is 0. 65535 / node. The second node is 65535 / node-2 × 65535 / node, and so on, it solves the problem of single point failure.

The acquisition node of the platform communicates and transmits data with the storage server which deploys GlusterFS DFS through InfiniBand. The captured data is stored in warc format. The acquisition node transfers the warc format data to the GlusterFS DFS at the end of the acquisition task. The data is unified into  the GlusterFS cluster virtual storage pool for unified organization and storage by using its global unified namespace management [11]. GlusterFS DFS uses multiple Brick to form the storage volume and mounteto the GlusterFS client. All the collected data is shared through the mount point and displayed the storage path in the front end of the platform. The platform performs URL and full-text indexing of the archived data by deploying Wayback and NutchWAX components.

In the initial stage of the platform, three servers were used as the acquisition server. When the platform was running for a period of time, the acquisition server was expanded from 3 to 7, and each server ran the following mount commands.

a. yum install centos-release-gluster
b. install -y glusterfs glusterfs-server glusterfs-fuse glusterfs-rdma
c. service glusterd start
d. gluster peer probe 192.168.xxx.xxx (xxx)
e. mkdir -p /mnt/glusterfs/
f. gluster volume create gv0192.168.xxx.xxx:/home/GlusterFS 192.168.xxx.xxx:/home/gv0
g. mount -t glusterfs 192.168.1.168:/gv0 /mnt/glusterfs/

### *2. The Technical implementation of hierarchical Resource sharing and Integrated Management*

The NLC is the center node as the manager to unified manage the web archiving tasks of all institutions connected to the platform. The center node is the information aggregation point for all nodes and real-time monitors the progress of each node's acquisition task for managing and adjusting the related business. The NLC deploys the complete hardware architecture and software system of the platform, and other libraries can deploy their own business modules in a modular way according to their business needs. The hardware resources of all libraries (institutions) together form the hardware environment of the Web Archiving and Service platform under the unified management of the cloud architecture. Each library can use its own local server, network and storage space to complete the web archiving but will not be interfered with each other. The metadata of all the libraryis synchronized and centralized into the database of the central node for realizing the management mode and service mode of the meta-set.

## III. The Operation Effect of the  Platform

The Web Archiving and Service platform has released in the NLC and 3 provincial libraries for months. By May 2019, the archived resources are up to 7.8TB (compression) and  23720 URLs, the average acquisition speed up to 23.75KB/ s, maximum acquisition speed up to 403KB/s.

The staff of the center node and the institutional node complete workflow by login browser client. At present, the platform works well and greatly improves the degree of work automation and saves manpower cost in the aspects of tasks' automatic distribution, bulk deployment and the task scheduling. The central node of the platform has been expanded from 3 to 11 physical servers and more than 200 collection nodes of 33 virtual machines realized the distributed collection and ensure the stable and efficient operation of the platform. The extendibility of the system is fully verified.

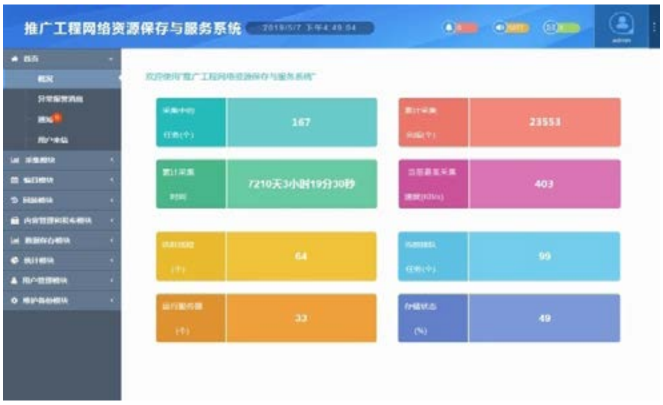

Figure 3 Platform front page diagram

## IV. Conclusion

The NLC has been devoted to the research of the online archiving related fields for many years and actively explored the relevant technologies, policies and development trends. The Web Archiving and Service platform independently developed has laid a foundation for the national, multi-agency, hierarchical web archiving. In the future, the NLC will continue to expand the scale of online archiving, strengthen technological research and innovation, and explore new models for data preservation and analysis in order to meet the changing needs of business.

## References

[1] L. Chen, S.-Z. Hao, Z.-G. Wang.Introduction of Acquisition and Preservation of Web Information to WICP and ODBN Project of National Library[J]. Journal of The National Library of China,2004(01).

[2] L.-Q. Zhao. Review of the Research on Web Information Preservation in China[J]. Reserch on Library Science,2011(2):5-7.

[3] Q. Sun, W. Zhang. Research on collecting and selecting Strategy of Web Information Preservation in Chinese Libraries[J]. Reserch on Library Science,2016(17):28-32

[4] Heritrix [EB/OL]. [2014-08-05]. https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[5] J. E. Halse, G. Mohr, K. Sigur\djsson, M. Stack, and P. Jack, Heritrix developer documentation, Internet Archive, 2004.

[6] ISO 28500:2009 Information and Documentation -- WARC File Format [EB/OL]. [2014-08-05]. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=44717.

[7] IIPC [EB/OL]. [2014-08-05]. http://netpreserve.org/.

[8] WebArchiveAccess[EB/OL].[2014-08-05]. http://sourceforge.net/projects/archive-access/files/wayback/.

[9] NutchWAX [EB/OL]. [2014-08-05]. http://archive-access.sourceforge.net/projects/nutch/.

[10] Davies A Orsaria A. Scale out with GlusterFS [J]. Linux Journal. 2013, 2013(235):1.

[11] L.-H. Wang, X.-Y. Li, Y.-B. Zhang. Mass Small-File Storage File System Research Overview[J].Computer application and software,2012,29(8):107.