BibServSys:面向科技文献分析的服务系统[①]

龙海燕 吴 斌 杨胜琦 毕 然

(北京邮电大学 智能通信软件与多媒体北京重点实验室 北京 100876)

BibServSys: Towards a Bibliographic Service System

Haiyan Long, Bin Wu, Shengqi Yang, Ran Bi

(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract: With the growing amount of bibliographies and expanding of scientific research, more in-depth research about bibliographic data is needed. The emerging of Complex Network makes a brand-new insight into the demanding. To provide a comprehensive bibliographic service, in this paper, we develop a service system, called BibServSys (Bibliographic Service System). We propose a methodology to construct a service system which makes its components and even itself reusable. The architecture of the BibServSys is elaborated with components, connectors in it. And at last we also take examples to show the comprehensive bibliographic service the system provided. Most parts of the paper are concentrated on the design and implementation of the system as well as the features of the system.

Key words: bibliographic service system; constructing methodology; SOA; complex network; scientific research network

1 Introduction

With the coming era of mass data storage, people become more and more unsatisfied with merely finding simple plane information of different types about bibliographies through indexing, such as titles, authors and publishing date. After getting the results indexed, people may be much interested in multi-dimensional information about bibliographies, such as ranking of results indexed in different criterions, relationship among selected results or authors of these bibliographies and even more valuable and in-depth information about scientific research. Bibliographic Service System (BibServSys) mainly aims at providing comprehensive service which integrates retrieval and statistic analysis of bibliographic information with mining and visualization of scientific research network.

Bibliometrics has a set of methodologies to analyze bibliographies by using statistical approaches[1]. And as a member of social network analysis, bibliography related scientific network analysis usually use both statistic and visualization methods which generally be employed to conduct SNA (Social Network Analysis). But it is characterized by its property of Bibliography that can hardly be analyzed by SNA thoroughly. Recently, many methodologies and technologies of data mining, statistic indicators computing and relation visualization have been devised. In order to make these constructed components and systems can easily be reused

[①] Supported by the National Natural Science Foundation of China under Grant No.60402011; National Key Technology R&D Program of China under Grant No.2006BAH03B05; Specialized Research Fund for the Joint Laboratory between Beijing University of Post and Comminications and IBM China Research Laboratory under Grant Nos.JTP20071002-4, JTP200806014-3

in other systems, the issues come to: 1) How to construct independent components for those parts above as well as how to connect them in an appropriate approach which allows these components interact smoothly and efficiently. 2) How to provide the system as a service as well as how to deploy it.

To solve these problems, idea of Service Science is adopted. Due to the multiformity and complexity of the current situation of Service Science, constructing an appropriate service system for bibliography analysis should make many decisions among existing ideas and technologies. This paper presents such an avenue that is about how to construct the BibServSys and how to provide the system (without representation layer) as a service to a host system (such as a comprehensive SNA related system) which can employ BibServSys to provide particularly insights into the science of Bibliography. The methodology is concluded from actual developing experience of BibServSys to provide a viable methodology solution on constructing of other similar systems.

Here we present a list of main tasks of our system which can be summarized as the following 4 contributions:

(1) We illustrate the method to construct a service system using a self-developed system called BibServSys.

(2) We propose a novel method to make reuse of existed algorithms and systems and also make new components and even the integrated system reusable.

(3) We illustrate the approach to provide comprehensive bibliographic service by integrating statistic and network analysis.

(4) Incorporating the social network analysis, we resolve to provide a visual perspective and a predominant solution to bibliographers.

This paper is organized as follows. Section 2 reviews the related work. Then Section 3 elaborates the constructing of service. Section 4 introduces the design and implementation of BibServSys. At last, Section 5 examples a case to illustrate how to conduct research on BibServSys.

2 Related Work

There are a number of software tools designed to help analysts to understand social networks. Tools such as UCINET, Pajek[2], KrackPlot[3] focus on statistical analysis and feature limited interaction in their visualizations. Others (such as NetDraw and Tom Sewyer) focus on visualization, but lack many statistical analysis. These tools used to analyze social network can also used to conduct bibliography related mining, but not sufficient yet.

There are also many existed systems or tools providing bibliography related service (such as Scopus, CiteSeer [4] and RefVis). But most of them tend to focus their efforts to analyze some aspects of bibliographic service. And these systems mentioned above only conduct analysis around data and ignore relations among data of different types.

When considering of relations of data, in order to analyze all kinds of relation well, analyzing methods in complex network are needed. Many systems oriented to scientific research network mining are already existed, such as ArnetMiner[5], CiteSpace[6] and NWB. But they are not comprehensive enough yet.

Though many related tools and systems are developed and many algorithms, process are the same in these systems, but few have been benefit of practicing the concept of SOA and making the algorithms and processes reusable. Therefore, we propose a method to design and implement a service system which makes itself reusable as a service as well as services in it.

3 The Constructing Methodology

As Fig.1 presents, in our methodology, we use understanding of service and service lifecycle as a guideline for the system architecture. Then service scenario analysis is conducted in order to extract services in the system. According to the architecture and services which act as components in the service system, we depict the processes which act as connectors among components.

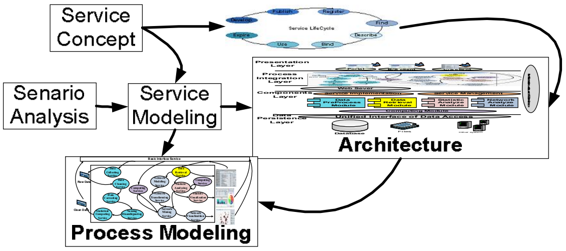


Fig.1 Constructing methodology

3.1 Service concept

In Service System, Service can be defined as the application of competences for the benefit of another [7]. It means that service is a kind of action, performance, or promise that' s exchanged for value between provider and client[8]. Generally speaking, there are three distinct understanding of service concept: Fine-gained service, Coarse-grained service and Mixed service [9].

In bibliographic service system, the service provided to clients are information retrieval, scientific assessment, study of scientific research network and hunting for commercial opportunities derived from the research area. These services demanded by various clients are different and complex. In such a situation, composition of services is necessary, which means that the service emphasize particularly on application-level operation. So the fine-grained service concept is employed to help understanding.

Under the understanding of service, in order to find services needed in service system, it is essential to depict the scenarios

3.2 Application scenarios analysis

Scenario analysis is a process of analyzing possible future events by considering alterative possible outcomes (scenarios). The analysis is designed to improve decision-making by the consideration of outcomes and their implications. One usage of scenario analysis is in economics and financial to forecast several possible scenarios for the economy (e.g. rapid growth, moderate growth, slow growth) and financial market returns (for bonds, stocks and cash). And another usage of it is in software architecture evaluation as in Ref.[10].

In this paper, we get rid of the inherent complexity of scenario analysis and simplify it to application scenario analysis which concentrates our efforts on which services needed to provide in each application scenario.

There are three kinds of users may be benefited by emphasizing particularly on different aspects of the services the system provided:

(1) Common scientific researcher: To find bibliog.-raphies by specific key words research communities in specific area and potential cross-disciplines.

(2) Researcher of scientific assessment: To Analyze and assess the direction and influence of research departments. To do research on the relations among various academic communities. To assess personal achievements and influence of researchers.

(3) Cooperating opportunities hunter: To find outstanding laboratories on specific areas to cooperate.

From these scenarios, we extract five main services that BibServSys needed to provide: data preprocessing service, data retrieval service, statistic analyzing service, network analyzing service and computing service. And since each service could be widely used in other systems (e.g. network miming service and visualization service are necessary components in any system which conducts research using complex network), we also provide them as services while acting as components in the BibServSys.

3.3 Service lifecycle

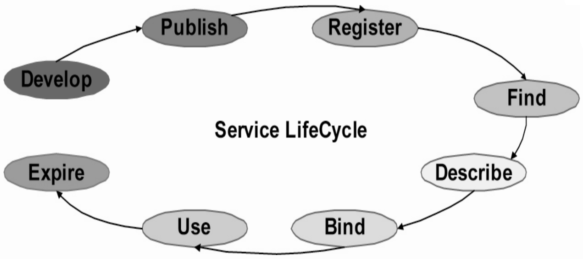


Fig.2 Service lifecycle

Service lifecycle could be considered as a set of ordered phases that services may experience in the service system[7]. Based on the typical service lifecycle, we propose another lifecycle according to the under- standing of service. Services in BibServSys and even BibServSys itself will experience such a lifecycle. As it is depicted in Fig.2, there are eight phases in the lifecycle: develop, publish, register, find, describe, bind, use and expire.

4 The Design and Implementation of BibServSys

Under the consideration of service concept and service lifecycle, in this sector, we elaborate the design and implementation of BibServSys including architecture, components and connectors.

4.1 Architecture of BibServSys

As shown in Fig3, the system could be considered as a five layer structure.

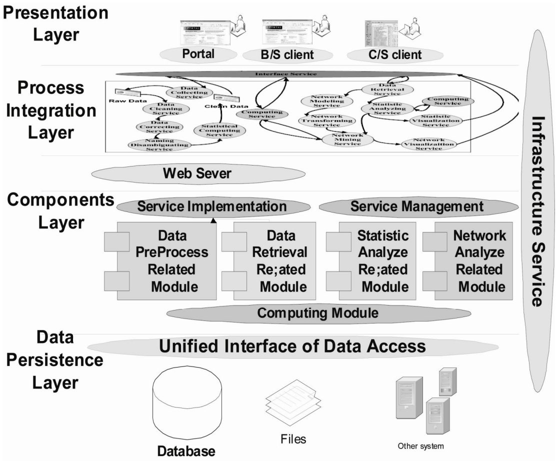


Fig.3 Architecture of BibServSys

Data Persistence Layer: Mainly handle various format data, including database, files and even other existed systems. The layer transforms various data sources into a unified data model, which could be used by components in the higher layer. Besides, a unified interface is provided to data access.

Components Layer: Encapsulate the existed systems in Data Persistence Layer to different components and use the unified interface to obtain the data from underlying layer. The Service Implementation module is responsible for constructing services based on components in these layers while the Service Management module is mainly responsible for managing services just as its name implies.

Web Server Layer: Publishing and registering services constructed in the underlying layer and make the legal services available for the Process Integration Layer.

Process Integration Layer: To construct processes using the encapsulated service in the underlying layer. Through compositing, arranging and coordinating various services to an integrated service (as Fig.4), it can provide an efficient organizing process to aid relevant users to conduct analysis.

Presentation Layer: To provide interface to user in an appropriate way. For the system is constructed based on SOA, multiple user clients are allowed, user can access the system through portal, simple B/S client and also C/S client.

4.2 Components in BibServSys

For the complexity of the application, the composition form of services is widely applied to construct system. In our work, a process issued to integrate several services to a unified one as Fig.4. The following subsections introduce each unite in detail.

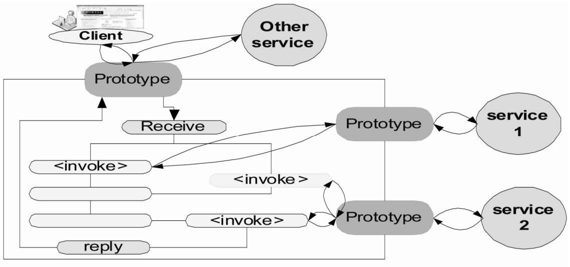


Fig.4 Service cmposition

4.2.1 Data preprocessing service

Data preprocessing is the basic service that is executed before or during any other services. For the incomplete and ambiguous nature of data source (e.g. duplicated names among different authors), data preprocessing service concludes abundant trivial works, such as data collecting, data cleaning, data correcting, name disambiguation and some basic statistical computing tasks. These works will be executed sequentially. To Construct the Service, Service Pipeline (SP) architecture [11] is adopted.

4.2.2 Data retrieval service

Data retrieval service mainly aims at providing user with the results indexed as quickly as possible. For example, when using "DNA" as a topic, all articles about DNA can be displayed to user. During the process of retrieval, many services also could be used, such as data preprocessing, data querying and results displaying. To construct the Service, Call-Return (CR) architecture is adopted.

4.2.3 Statistic analyzing service

Statistic analyzing Service mainly aims at providing user with comprehensive and explicit statistic information of assessing indicators about bibliographies. In addition, it instructs user to conduct network analyzing more efficiently, for it can guide the user to persons or departments that they are actually interested in. Three main services are constructed in the statistic analyzing service (Call-Return architecture):

(1) Statistic Indicator Computing: To compute various assessment indicators and rank the objects

according to computational results. The indicators include: amount of published papers, annual amount of cited papers, H factor, Price Factor, Impact Factor, Immediacy Index and so on.

(2) Statistic Results Displaying: To display statistic results in a user-friendly way. Usually, visual tables and lists are employed.

(3) Statistic Visualization: To display statistic results in a comprehensible visual form (charts, tables).

Since abundant computations are needed in the statistic analyzing service, there will be many interactions with the computing service during its execution.

4.2.4 Network analyzing service

Network analyzing service plays an important role in the BibServSys. Many concepts in Complex Network are used in the process of analyzing. And in this part, we propose a novel network modeling method which constructs the network with a heterogeneous structure. This means the nodes in the network are possibly of different types. And the network can be transformed from heterogeneous structure to homogeneous one. The merit of this model is to provide multi-dimensional views about the network. Five main services are included in the network analyzing service, we elaborate them as:

(1) Network Modeling: To map the data into a unified network which can reveal the relations among data, a novel heterogeneous modeling method is proposed. In which, the type of nodes may be author, paper, keyword and even publisher, and the edge may be the relation of publishing and occurring. Heterogeneous network provides user with multi-dimensional insights into the data.

(2) Network Transforming: To convert the network from the heterogeneous structure to a homogeneous one (e.g. co-citation network, collaboration network and so on), on which we can conduct lots of analysis about the network and find in-depth insights. Based on this transformation, it is possible to offer comprehensive analysis on network.

(3) Static Characteristics Analyzing: To analyzing the network by degree distribution, Cluster Coefficient, expanding rate, betweeness, etc. These statistical characteristics can well reflect the features of the network. For example, the author with low degree and high betweenness in co-author network may indicate that the author connecting two research areas. Attentions on these nodes may reflect meaningful information of the network.

(4) Cluster Analyzing: To divide data into useful groups (Nodes in a group has similar attributes.). For example, when the dividing criterion is the degree of the nodes, we can find various group, authors in a same group has similar degree. Accordingly, we can rank author groups by degree level to reveal authority of groups.

(5) Community Detection: To find communities in which nodes has denser connectivity than periphery nodes. As in collaboration network, community detection could help researchers to find community with the same research interest. And in keywords network, sets of keywords that can describe a specific topic. In our system, we integrated lots of existed algorithm, such as GN, FAST, CPM, etc.

Network Visualization: After various analyzing services' processing, an intuitive way to display the analyzing results to guide the user to more in-depth and useful ideas and discoveries. Displaying information in terms of graphs and pictures is more understandable than in terms of tables. What's more, the former makes visible what would have otherwise remained hidden in the numbers[12].

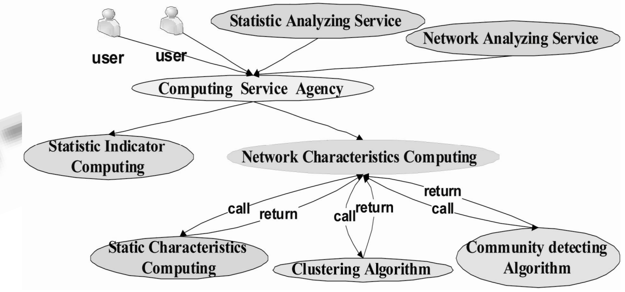

user
user
Statistic Analyzing Service
Network Analyzing Service
Computing Service Agency
Statistic Indicator Computing
Network Characteristics Computing
call
return
call
return
return
call
Static Characteristics Computing
Clustering Algorithm
Community detecting Algorithm

Fig.5 Computing service architecture based on Agency and Call-Return (CR)

4.2.5 Computing service

Massive calculations are ubiquitous in both statistic analyzing service and network analyzing service. Thus a computing service is necessary. The service could be executed during the statistic analyzing process or network analyzing process or be used directly by user. So agent architecture is adopted to construct the service which can allocate appropriate computing components and algorithms to satisfy demands of users and other services. Fig.5 presents the computing service architecture.

4.3 Deploy services

For the aim of reuse, we deploy the services introduced in Sector4.2 to the web server, and register them to the registry so that user can find and use them.

4.4 Connectors among components: process

Services are illustrated discretely, then how to integrate them together and provide a systematic analyzing platform to user. The issue comes to what are connectors among various services As the architecture of BibServSys shown in Fig.3, process plays the role of the connector. Fig.4 gives an illustration about how the process connects various services. Services in a process are so loosely coupled that we can only make little change about the definition of process (BPEL would be used to define the process) when it’s necessary.

Besides the preordered services, some processes in BibServSys are user-participated. Fig.6 presents processes in the system. The interactions occur during the execution of processes. For example, during the execution of network analyzing service, computing service is needed, so it is integrated into the network analyzing process.

4.5 BibServSys as a service

Bibliographic Service System can also be implemented as a service and to be published, so that other systems with the same architecture (as Fig.3) can find it and use it separately or integrate it into some more complex services. The Bibliographic Service itself is a composition of Data Preprocessing service, Data Retrieving service, Statistic Analyzing Service, Network Analyzing Service and Computing Service.

5 Case Study

As large amount of analysis are supported by BibServSys, we just take a simple case to show how the system works.

Data Preprocessing: Millions of records of articles about life science spanning from 1997 to 2007 are exploited by our system. Since the data are incomplete, noisy, ambiguous and inconsistent. A great deal of work has gone into data cleaning. After the processing, the number of authors descent from 2,653,486 to 2,367,632.

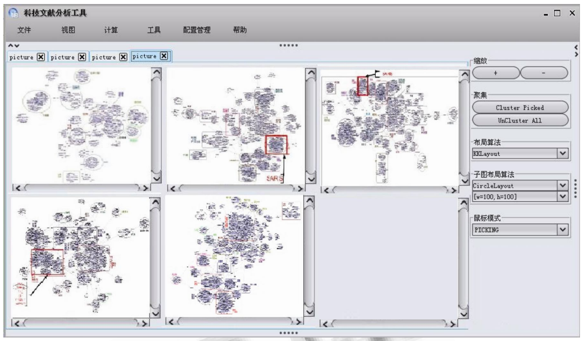
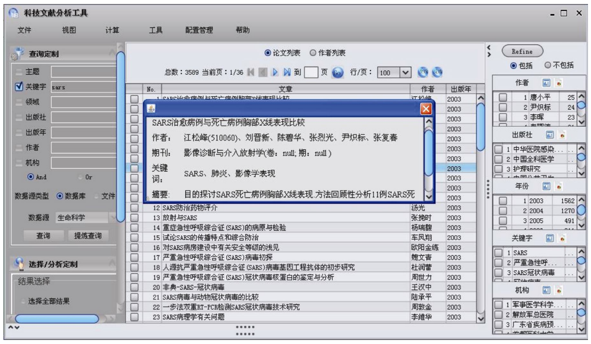


Fig.6　Evolution of bibliographies in life science

To get an overall idea about the records year by year, we need records in 2002, 2003, 2004, 2005 and 2006 for research separately, Data Retrieval Service is adopted. And we conduct community detection and its evolvement on these records. (Network Analyzing Service) Fig.6 presents the community evolution of Bibliographies in Life Science from 2002 to 2006.

What insights can we get from observing these graphs Community detection divide bibliographies into various communities which stand for interesting research topics, such as hospital, doctors, nurse, patients, insulin, diabetes and hypertension. From the evolution visualizetion, we can learn that as follows: Relations between different disciplines are closer. Cell Science, Gen Engineering and RNA are jumped-up research area in recent years. What’s more, some area shows obvious timing property. For example, research about SARS was rising in 2003(Because it bursted out in the beginning of 2003) and was not so hot since 2006(For we can’t find any obvious community that about SARS, it faded away rapidly).



□ 主题
☑ 关键字 sars
□ 领域
□ 出版社
□ 出版年
□ 作者
□ 机构
● And　○ Or
数据源类型 ● 数据库　○ 文件
数据源 生命科学
查询　提炼查询

● 论文列表　○ 作者列表

总数：3589 当前页：1/36　到 　页　Go　行/页：100

	No.	文章	作者	出版年
□	1	SARS治愈病例与死亡病例胸部X线表现比较	江松峰	2003
□				2003
□				2003
□				2003
□				2003
□				2003
□				2003
□				2003
□				2003
□				2003
□				2003
□	12	SARS防治药物评介	汤光	2003
□	13	放射与SARS	张挽时	2003
□	14	重症急性呼吸综合征（SARS）的病原与检验	杨瑞馥	2003
□	15	试论SARS的传播特点和综合防治	车凤翔	2003
□	16	对SARS病房建设中有关安全等级的浅见	欧阳金练	2003
□	17	严重急性呼吸综合征（SARS）病毒初探	魏文青	2003
□	18	人源抗严重急性呼吸综合征（SARS）病毒基因工程抗体的初步研究	杜润蕾	2003
□	19	严重急性呼吸综合征（SARS）冠状病毒核蛋白的鉴定与分析	周世力	2003
□	20	非典-SARS-冠状病毒	王汉中	2003
□	21	SARS病毒与动物冠状病毒的比较	陆承平	2003
□	22	一步法双重RT-PCR检测SARS冠状病毒技术研究	周敦金	2003
□	23	SARS病理学有关问题	李維华	2003

SARS治愈病例与死亡病例胸部X线表现比较
作者：　江松峰(510060)、刘晋新、陈碧华、张烈光、尹炽标、张复春
期刊：　影像诊断与介入放射学(卷：null; 期：null)
关键词：　SARS、肺炎、影像学表现
摘要：　目的探讨SARS死亡病例胸部X线表现.方法回顾性分析11例SARS死

● 包括　○ 不包括

作者

□	1	唐小平	25
□	2	尹炽标	24
□	3	李晖	23

出版社

□	1	中华医院感染...	..
□	2	中国全科医学	..
□	3	护理研究	..

年份

□	1	2003	1562
□	2	2004	1270
□	3	2005	491

关键字

□	1	SARS	...
□	2	严重急性呼...	...
□	3	SARS冠状病毒	...

机构

□	1	军事医学科学...	..
□	2	解放军总医院	..
□	3	广东省疾病预...	..

Fig.7　Data retrieval results with the keyword “SARS”

In allusion to SARS, we can find that the statistic properties (Data Retrieval Service (retrieval results in Fig.7) and Statistic Analyzing Service is employed.) about SARS is corresponding to the community evolution (histogram in Fig.8). Suppose we are interested in bibliographic information about SARS, we can get all records about SARS through data indexing , and from the statistic analyzing, we can get the experts (Who is more productive and Who has a higher impact)

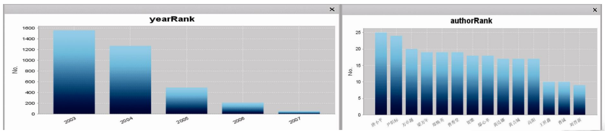


Fig.8 Rank of year and author on SARS related bibliographies

As the research that could conduct using BibServSys is so many that we don't give unnecessary details. And the aim of this case study is to illustrate how the services introduces above are used in BibServSys and how a comprehensive bibliographic service the BibServ Sys can provide.

6 Conclusions and Future Work

This paper presents the method of constructing a service system and illustrates advantages of using SOA architecture. In addition, a novel method which integrates statistic analysis with network analysis is proposed. And we emphasize particularly on the design and implementation of BibServSys which is a comprehensive bibliographic service system developed by ourselves and has many outstanding and novel features. At last we give a scenario to demonstrate the system, the outcome is quite good.

When the amount of data comes to a large scale, the speed of data retrieval and network visualization may slow down. Future study is to embed the service system to a host system and improve the performance of the system when conducting massive computing.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant No.60402011, and the National Key Technology R&D Program of China under Grant No.2006BA H03B05. It is also supported by IBM China Research Laboratory, the Specialized Research Fund for the Joint laboratory between Beijing University of Posts and Communications and IBM China Research Laboratory (ProjectNo.JTP20071002-4,JTP200806014-3). We would like to acknowledge the implementation works by Lijun Suo and Deyong Hu.

References

1 Catharina R, UlfKronman K. Bibliometric handbook for Karolinska Institutet. Karolinska Institutet. Karolinska Institutet University Library publications, 2006－11.

2 Nooy D, Mrvar WA, Batagelj V, et al. Exploratory Social Network Analysis with Pajek. NY, USA, Cambridge University Press, 2005.

3 David K, Jim B, Cathleen M. KrackPlot 3.0: An Improved Network Drawing Program. Connections, 1994,17(2):53－55.

4 Giles CL, Bollacker KD, Lawrence S. CiteSeer: An Automatic Citation Indexing System. ACM Press, 1998:89－98.

5 Jie T, Jing Z, Limin Y. ArnetMiner: Extraction and Mining of Academic Social Networks. KDD'08, 2008:990－998.

6 Chen CM. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. Journal of the American Society for Information Science and Technology, 2006.

7 Vush RE, Vargo SL. The Service-Dominant Logic of Marketing: Dialog, Debate, and Directions. Armonk, NY. M.E. Sharpe, 2006.

8 Spohrer J, Maglio PP. Bailey J, Grahl D. Steps Toward a Science of Service Systems. IEEE Computer, 2007, 40(1):71－77.

9 Hou J, Liu SJ, Meng XX, Li H. Research on the Constructiong Methodology of Service Ecosystem. Journal of Harbin Institute of Technology, 2008,1(15).

10 Mugurel TI, Dieter K, Hammer, Henk O. Scenario-Based Software Architecture Evaluation Method:An Overview.

11 Wang ZJ, Xu XF, Mo T. Service Architecture: High Level Descriptions of Service System. ICSS08, 2007.

12 Gershon ND, Eick SG. Informattion Visualization. IEEE Computer Graphics and Applications, 1997,7－8:28－31.