I.J. Engineering and Manufacturing **2011, 5, 59-65**
Published Online October 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2011.05.08

Available online at http://www.mecs-press.net/ijem

Research of Distributed Data Mining System Based on Web Services

LIU Weiming[a], YANG Jian[a], MAO Yimin[a]

[a] *Jiangxi University of Science and Technology, Ganzhou, Jiangxi, China*

Abstract

For the following reasons, the original centralized data mining became more and more out of date. First, the data source need to be processed is distributed on different computers in the networks. Second, because of the constrain of networks band, the privacy and safety of data, the incompatibility of systems and so on, it is not realistic to put all data source in a place for centralized data mining. Third, more and more demands have addressed on the openness and easy accessibility. The distributed data mining technology was presented for the problems mentioned above. This article introduced the latest technology for distributed component technology—web services technology into distributed data mining field, and took some tentative efforts in solving the problem of designing for suitable architecture of distributed data mining systems and corresponding distributed mining algorithms.

Index Terms: Data Mining; Distributed Computing; Component Technology; Web Service

© 2011 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science.

1. Introduction

With the rapid development of information technology, we can easily access and store various data. However, we are now faced with issues that a wealth of information resources but lack of knowledge. People will not be satisfied with surface treatment of data, such as statistics and inquiry, so dig hidden information and intrinsic relationship between the data, naturally became an important task.As the largest data source Internet is, there are a lot of different types of information resources, and contains the knowledge of great potential value. Equally, there are a lot of data sources of rich information, people are eager to obtain valuable resources and knowledge from these data sources. Data Mining is a technology which intelligently and automatically converts data into useful information and knowledge, it become the focus of information technology. This is a new cross-disciplinary which based on statistics, pattern recognition, artificial intelligence, machine learning, database technology and high-performance parallel computing and other fields, it has been successfully applied in the economic, financial, astronomy and other industries field. The original data set can be structured, semi-structured, even heterogeneous data which distribute on the network. Mining knowledge can adopt mathematics, non-mathematical, deductive and inductive methods.

2. Overview of Distributed Data Mining

2.1. What is distributed data mining

Distributed data mining has two meanings: first, by using distributed algorithm, it is a procedure that discovers knowledge from logical or physical distributed data sources[1]. Distribution is emphasized here. Second, users, data, mining software and other software components which relate to a certain data mining task are geographically dispersed. It is emphasized on dispersion of the soft component.

2.2. The problem to be solved by distributed data mining

1) Global centralized control: in order to facilitate the realization of distributed data mining, a site which is used for centralized controlling is necessary, in the case that there is no global control site, the communication overhead the whole system is very large. In order to get the global knowledge, all the sites will be a lot of radio, comparing to global control site, there is no doubt that cost and difficulty is much bigger.

2) Parallel and distributed data mining algorithms: this is actually past for performance issues, running data mining algorithms on the large-capacity data set will take a long time, because the time complexity of data mining algorithm is high, a better approach is to use parallel data mining algorithms, partition data set into several subsets, and combine the results of each subset of mining after parallel processing [2].

3) Knowledge sharing: when making distributed mining between each site, it is necessary to select a understood knowledge form. One is that distributed data mining generally consists of the excavation for the knowledge, so we must take a unified understood knowledge representation in order to achieve synergy mining between each site. The other is that users may need access to knowledge on other sites, it also requires a general knowledge representation.

4) Distributed software design: soft component is a distributed object which does not bind with other specific program or computer language. It can across heterogeneous platforms, with encapsulation. It interacts with the outside world by re-defined application program interface. Its biggest advantage is supporting for software reuse, system designers can use existing software components, this will optimize the division of labor, greatly reducing the coding workload.

3. Web Services Technology

3.1. Web services architecture

A typical web service architecture is shown in Figure 1.

There are three roles in the web services architecture: services provider, service registry and service requestor. Service provider is supplier who provides the final web services, it implements a application which is written for a particular demand and placed in the online server for others to use. From a business perspective, service provider is the owner of web services and in charge of publishing, updating and recycling of the services, service requestor is the consumer of specific services [3]. From the perspective of web services architecture, service provider is the platform that achieve the web services, services requester is the user of the services and application of finding and invoking a particular service. Service registry is a web services registration, collects many online web services. Generally, when web service provider installs the web services onto the online server, it will post the web services to service registry. For the service requestor who wants to use web services, first, he have to search the service registry. When found the right web service, he will get the technical information reference from the service registry, then find web services and related technical information by these references to complete the binding between the service requestor and service provider.

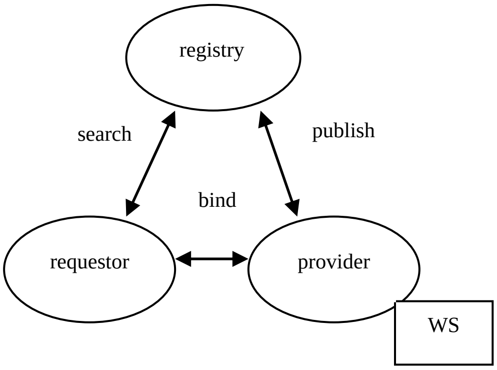

Fig. 1. Web services architecture

3.2. Web Services protocol stack

Implementing a complete Web service system needs to be supported by a series of protocol specification. Figure 2 shows the system of the web services technology—web services protocol stack.

? ? ?	? ? ?	Management	Quality of Service	Security
Routing, Reliability and Transaction	? ? ?			
Workflow	WSFL			
Service Discovery, Integration	UDDI			
Service Description	WSDL			
Messaging	SOAP			
Transport	HTTP,FTP,SMTP			
Internet	Ipv4,Ipv6			

Fig. 2. Web services protocol stack

Web services must be accessible by network, so that it can be called by requestor. Web services use a common Internet protocols (HTTP, FTP, SMTP, Ipv4, Ipv6). The four layers that Workflow, Service Discovery, Service Description, Messaging are messaging and the service description layer of XML, which are the standard protocols for web services, the upper two layers are protocols about routing, reliability, and transaction and other aspects. The right part that Security, Quality of Service and Mangement is the supporting infrastructure of protocol layer. In order to make the application of web services meet requirements of e-commerce, e-government applications, we have to provide enterprise level infrastructure, even city level, including security, association, service quality and so on. The infrastructure is to be involved in each layer, and corresponding solutions of each layer can be unrelated to each other. In the web services protocol stack, the more the underlying technology, the more mature and standardized. The high-level content and supporting infrastructure needs further development and standardization.

4. Architecture for Distributed Data Mining Based on Web Services

The point of integration of distributed data mining and web services technology [4].

1) Consistent storage and representation mechanism of the data is one of the basic problems that distributed data mining system should solve well. But the actual situation is often, different branches of the same institution use incompatible computer systems (such as different operating systems and database systems). In order to facilitate the exchange of data, we need a cross-platform means of data coding and organization. XML technology provides a vendor-independent data representation mechanism, so that the problem that data private and not compatible no longer be barriers to different application data exchange.

2) After the Web services technology, distributed data mining system will greatly enhance the ease of use. Web services technology is based on the Internet, once the service is deployed on the network, They can be used anywhere by SOAP and WSDL. Because of the platform independence of interfaces, the system has an architecture from application to application, rather than the traditional structure of the user to the application. Many data mining systems, such as IBM 的 Intelligent Miner、XELOPES and PolyAnalyst provide a platform independent interface, can be easily integrated into their data mining systems.

3) Machine learning and data mining research in the field with each passing day, and the needs of users is not static. A good distributed data mining system must have a dynamic, scalable architecture. After the introduction of Web services technology, the different components published in the form of web services, which were provided by different suppliers, can be dynamically invoked at run time. Data mining system could keep up the pace of Development of new technologies and user needs change.

It can be seen that, combination of web services technology and the distributed data mining, will have far reaching implications on distributed data mining.

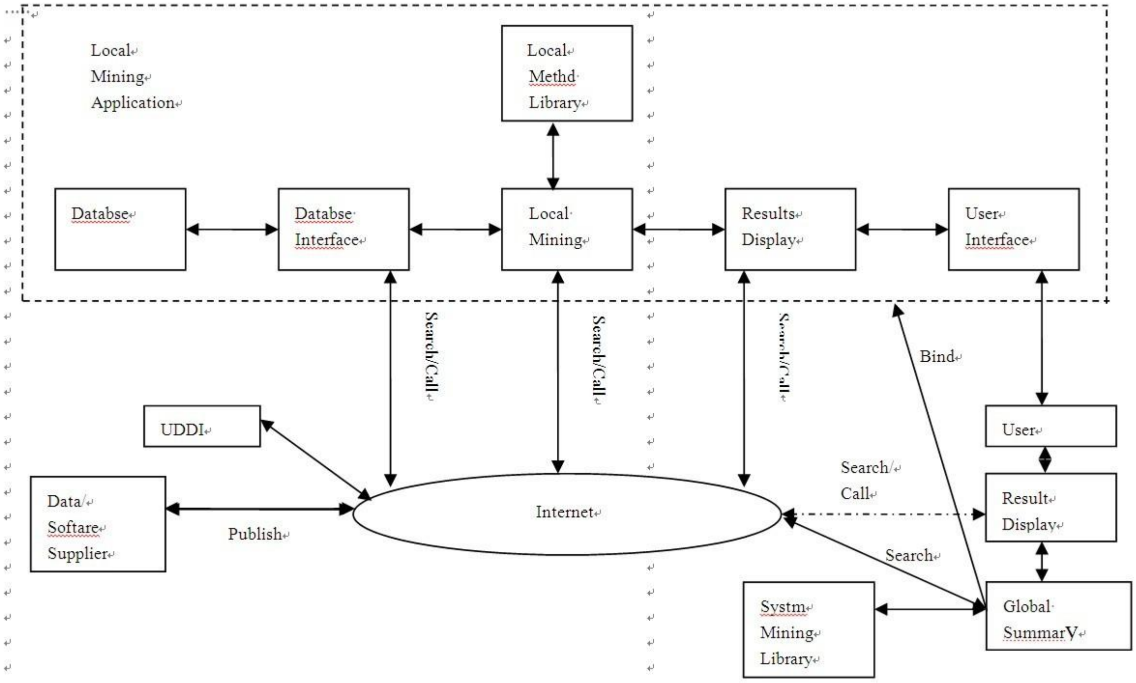

Fig. 3 distributed data mining system Architecture based on web services

Fig. 3. Distributed data mining system Architecture based on web services

Architecture of distributed data mining system based on web services

The system is based on web services, distributed computing system, Each local data mining module should be registered and published as independent web service. It has good portability, so that can cross platforms and heterogeneous data structures, and communicate by acrossing firewalls and proxy servers. It also has a good user transparency and scalability. The architecture includes the following modules: User interface, global summary of data mining, registration centers, local application of data mining and so on [5]. Structure of the system shown in Figure 3.Functions of main modules:

1) User Interface: user interface used to interact with the user and the computer, accept the request of the user's mining, show to user with mining results in the form of intuitive and easy to understand. In addition, user interface can also include user information databases. According to the user information, user interface analyze the user's identities and requirements, and the results presented to the user information databases. User information databases contain the two types of information: one is the user management information,which is used to register, maintain and manage the user information. The other is the user's interests and hobbies and other information, which is used as the inference rules of user interface interacts with the user. In order to facilitate user access, independent of the platform, reduce establishment and maintenance costs, user interface selects web form.

2) Global data mining summary [6]: it determines algorithm and data sets combining with global knowledge library according to submitted mining request. Then query the registry center, find relevant web services, implement the dynamic binding of local data mining, analyze the results of each local mining together and the final results presented to the user interface. You can also store the final ming result in the lobal knowledge library, for later analysis and comparison. Equally, global data mining application can also be packaged as Web services to register and publish, called by other applications.

3) Systematic mining knowledge library: the knowledge library store the knowledge in an appropriate manner which is mined by different user, in order to provide the basis for the user's decision-making [7].

4) Registration center: registry is a web service registration, which brings together many online web services. The web services application of the various local data mining

register here, in order to be searched by caller. In addition, mining components and data provided in the form of web service by other supplier can also register here, in order to be called by other application.

5) Local mining applications: the module is used to analysis the data on local database. According to the mining request of local user, select the appropriate mining algorithm and mining local data., and then submit mining result to display module, display module submit the ming result to user in the form of intuitive. Through the web services encapsulation layer, it can also accept the dynamic calling which come from the global mining application, analyze the calling, and select the appropriate data mining algorithm for local mining. Then pass the mining result to the global application. Because web services platform is independent with language, the module can be encoded using any programming language.

6) Data Mining Library: this module stores local various data mining methods, which is used for calling according to the need of local mining module.

7) The database engine and databases: present all major database vendors have launched their own database system, such as the famous ORACLE/DB2/SQLSERVER2000 etc. The database engine such as ODBC / JDBC and so on, is the role of a unified user interface to access these different databases.

When you want to be involved in the global data mining with several local data mining module [8]. The basic working principle of the system are as follows:

1) The user issues mining requests.

2) User interface accept the mining request, and transmit mining request to global aggregate modules in the predefined format.

3) Global aggregate modules analyze the mining request, determine the involved local mining applications.Then look for Registration Center, binding with the corresponding web service.

4) Global summary module passed mining requests to the local mining applications, local mining applications make local data mining according to mining request, and pass the result to global summary module.

5) Global summary module makes a comprehensive analysis with the result that were submitted by local applications, and comes the final result.

6) Submit the final results to results display module, display module will show them to users in form of intuitive.

To achieve the system, we choose multithread technology.The example code as follow:

```
／／Generate the local proxy object
WS=new DCD.1ocalhost.DCDw s();
Wsl=new DCD.1ocalhost1.DCDw s();
／／Send data using multiple threads
AsyncCaIlback1 一new AsyncCaIlback(CallBack);
AsyncCaIlback2 一new AsyncCaIlback(CallBaek1);
WS.Begin GetDb(dbl,AsyncCaUback1,nul1);
ws1.Ba ginGetDb(db2,AsyncCaIlback2,nul1);
privatevoid CallBa ck(IAsyncResult assignHandle)
／／Callback function1
{startlndex+一number;
Array.Copy(db,startlndex,dbl,0,number);
WS.BeginGetDb(dbl,AsyncCaIlback1,nul1);}
privatevoid CallaB ekl(IAsyncResult assignHandle)
／／Callback function2
{startlndex+一number;
Array.Copy(db,startlndex,db2,0,num ber);
WS.aBginGetDb(db2,AsyncCaIlback2,nul1);}
```

The example code implements the function that data mining client send data to two data mining web service at the same time.

5. Summary

With the development of distributed database system and Internet , because of the following reasons, the previous centralized data mining can not meet the needs of a distributed transaction:

1) In the real world, most of the large databases are in the form of distribution, therefore, it is necessary to propose a new distributed data mining system architecture. Meanwhile, we should design new distributed mining algorithms according with the feature of distributed data mining.

2) Because of limitations of network bandwidth, privacy and security of data, incompatibility of system and so on. It is difficult to concentrate all the data source in one place for centralized data mining.

3) System scalability is rather poor. When new data or mining component appears, it is difficult to integrate them into the system.

To solve the above problem, we propose a distributed data mining system architecture based on web services. The architecture builds on web services distributed computing architecture. Various local data mining modules are registered and published in the form of web services independently. It has good portability and heterogeneous of data structures.

References

[1] Aronis, J.M.Kolluri, V.Provost, F.J et al. The World: Knowledge discovery from multiple distributed database.Technical Report ISL-96-6.Department of Computer Science, University of Pittsburgh,1996.

[2] H.Kargupta,B.Park,E.Johnsom,et al.Collective data mining from distributedvertically partitioned feature space. Workshop on distributed data mining. Conference on Knowledge Discovery and Data Mining. New York,USA,1998.

[3] A.L.Prodromidis,P.K.Chan,S.J.Stolfo.Meta-learning in distributed data mining systems: issues and approaches. Advances in distributed data mining. AAAIPress, 1999.

[4] H.Kargupta,I.hamzaoglu,B.Stafford.Scalable distributed data mining using an agent based architecture.Proc of KDD97,Menlo Park,CA,1997.

[5] U.M.Fayyad,G.Piatetsky-Shapiro,P.Smyth et al.Advances in knowledge discovery and data mining.AAAI Press,1996.

[6] Liu B C,Ma X X,Ge S eta1． Research and Implementation on Web Services-based Service-oriented Software architecture[J]. Journal of Beijing Universityof Aeronautics and A stronautis,2004.

[7] R.Heckel.Towards Modeling Service-Oriented Architectures: Method and Semantics[EB/OL].http://wwwcs.upb.de/cs/reiko.html,2008.

[8] LIU Si-pei,LIU Da-you,QI Hong,et al.Service Community Chain Based Approach for Web Service Composition ［J］.Journal of Jilin University:Engineering and Technology Edition,2010.