Application of Web Service in Web Mining

Beibei Li[1] and Jiajin Le[2]

Computer Science Department，Donghua University，
200051 Shanghai，China
[1]bb@mail.dhu.edu.cn
[2]lejiajin@dhu.edu.cn

Abstract. To solve the problems we now encounter in web mining, We first propose a new distributed computing strategy——web service. It suggests building a web mining system based on web service, which can share and manage semi-structured data from heterogeneous platforms. Moreover, the system can integrate the mining services and algorithms, improve the efficiency of web mining, and make the mining results easier to access. We also conduct an experiment on selecting useful words to simulate the realization of the web mining system on the Microsoft.NET platform, which demonstrates the importance of Web service in Web mining.

1 Introduction

With the rapid development of online application, the quantities of data stored on the web have increased exponentially since 1990s. As a result of this increase in data quantity, how to discover and obtain useful information faster and more accurate from this vast data resource has become a focus of researchers' interest. Web mining was studied just under this situation, which refers to the procedure of extracting interesting patterns and knowledge from web resources, such as web contents, web structure and web access. It contains three different data mining tasks: Web Contents Mining, Web Structure Mining, and Web Usage Mining [1].

Now the prospects of Web mining technology have attracted the attention of researchers and commercial organizations. They have developed many mining systems and proposed lots of mining algorithms. However, Web mining is a complex procedure and it still faces many problems.

First, as to the Web Contents and Web Structure Mining, most conventional mining algorithms could not get used to the hetero-structured web resources. Because the web resources contain not only structured data in conventional databases, but also non-structured data such as multimedia, sounds, and images. Web mining has extended the objects of data mining from simple and structured data to complicated and semi-structured web data, which hinders those conventional mining algorithms from working effectively.

Secondly, as to the Web Usage Mining, usually the usage records of users are not static. Since some clients use local cache and proxies, the web server could only take

the information of those proxies, while missing the users' real information behind them, which prevents catching the behavior of single web user.

Thirdly, Web mining is mostly used by web servers only. So no matter which kind of mining, Web Contents Mining, Web Structure Mining or Web Usage Mining, they all have to be performed by the servers. Therefore, the lack of clear and better design for clients has caused low mining efficiency.

Fourthly, now there is contradiction between the large variety of Web mining algorithms and the specialty of their application fields [2]. The users could use those algorithms provided by certain system only, and never could add, delete or organize any of them, which makes the mining systems quite closed and inflexible.

Lastly, it is difficult for other systems to process the mining results further, because the formats of most mining results are quite closed. And this closure on the output formats is really bad for the improvements and use of those mining results.

So we can think of a new technology nowadays——Web Service. It is a new distributed computing technology and helps solve the problems above.

2 Advantages of Web Service in Web mining

2.1 The characters of Web Service

Web service is a new distributed computing model on the Internet [3]. It can be used for coupled network environments and provides an interface, which can be accessed by XML message through the network. This interface defines a group of accessible operations, in which functions can be easily realized just by using this standard interface. Figure 1 shows the related standards and architecture [4].
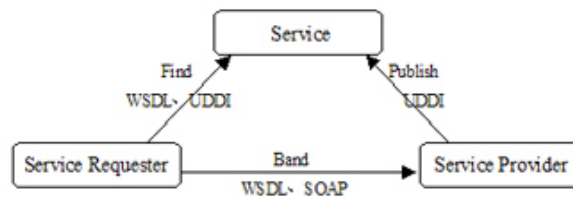

Fig. 1. Web Service Architecture and related standards

Therefore, the key point of Web Service lies in making remote process calling easier and faster by providing a set of Internet standard protocols, such as WSDL, UDDI, SOAP. Meanwhile, it has the compatibility with hetero-structured system platforms, which can eliminate the isolation of information, thus sharing and managing resources over different platforms.

2.2 How to solve the problems in Web mining

From the above discussions and explanation, we can design a new Web mining system. This new system follows all the standard protocols already defined, and integrates various Web mining services based on different kinds of algorithms, operating platforms, and data resources. Figure 2 shows its architecture.

The entire system consists of three layers: the Clients Layer, the Server Layer and the Web Mining Service Providing Layer. The Server Layer is its kernel, which is deployed on the network as the form of web service, finding useful mining services from the Web mining Service Providing Layer according to the clients' particular requests, integrating them with local service, and then providing it to the clients as a whole web service. Obviously, using the web service technology, this new system can easily overcome those problems we encountered in the conventional Web mining process.
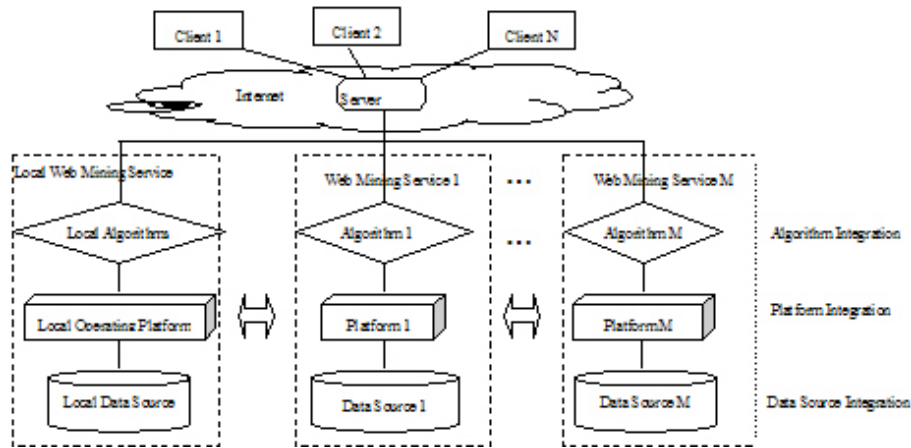

Fig. 2. Architecture of a new Web mining System Based on Web Service

1. Web service technology can solve the problem about sharing and managing the hetero-structured data in Web Contents Mining. Because web service is based on the standard XML language, it can integrate hetero-structured data sources and provide structured descriptions to them, thus making those data from different sources easily bonded together. Therefore, the service requesters could obtain the data directly from the web with the original data formats, rather than allowing the original data formats to get lost in the semi-structured description information with various formats.
2. Web service technology can help clients improve efficiency in the Web Usage Mining. In the proposed new system, the clients can call the mining service interface provided by the server conveniently and directly just like calling its local application procedures, which makes the clients not subordinate and passive any more, but as active as the server. So the Web mining is not confined to the server, and changes from the conventional "One-to-Multeity Mining" to the parallel "Multeity-to- Multeity Mining", which improves the mining efficiency. Meanwhile,

by calling the Web mining services, the clients could directly give the feedback of the local usage pattern analyses to the server, rather than being screened by those proxies or local cache, which makes tracking the behavior of individual users more easily and plays an important role in building personalized recommendation systems.

3. Web Service technology can help strengthen the specialty of the Web mining algorithms. For example, in the Web Structure Mining, the new system can integrate all those Web Structure Mining algorithms that exist in the form of web services, such as HIT, PageRank, improved HIT and etc, and make them form a unified mining service interface. Then the users could select, add, delete or optimize the algorithms according to their own mining tasks, which forms a custom-made mining developing platform and strengthen the specialty of Web mining.

4. Web Service technology can also deal with the results of Web mining effectively. Since the mining results are described and output by the XML language at last, the contents and formats of them are totally separated. The users can figure out the results easily and correctly, use various analyzing tools to visualize them, or make further use of them. Thus this system realizes the sharing of Web mining results over different kinds of systems and platforms.

3　Experiments

3.1　The implementation of a simulation system

Web service develops very fast nowadays, and many companies have offered various sorts of platforms to design, develop and deploy this new distributed web application. Such as Sun One, IBM's WebSphere, and Microsoft's .NET. Now we select the .Net platform from Microsoft to develop the Web mining simulation system.

First, we deploy a local web service named "StartWithA" on the local server, which is used to choose the words starting with "A" in certain text. At the same time, we deploy a remote web service named "EndWithD" on another server, which is used to choose the words ending with "D" in certain text. Finally we add its web reference to the local web service.

Then we can use those two integrated services through the local server at any client machine, only by adding the web reference of the local web service. The working interface of this simulation system is shown in Figure 3.

We input all the text into the textbox and click the "Select" button. Then the system can automatically find those two web services we already deployed and call them rapidly. Finally it returns the result we want——all the three words, "and", "aloud" and "around", in the text, which both start with "A" and end with "D". Figure 4 shows the result page.

Fig. 3. The Interface of the simulation system

Fig. 4. The result of selection

3.2 The results

Through the process of developing the simulation system we could easily find that when dealing with the real Web mining work, the Web Service technology can not only help the clients by local service, but also integrate a large variety of remote mining services by local server just as fast and convenient as calling local procedure. Meanwhile, by using Web Service as the interface between server and clients, the users could easily write the clients software by themselves, or even integrate all the function provided by this entire system into another system as its subsystem, as long as they conform to the rules when calling the Web Service of this system.

Besides, thanks to the B/S (Browser/Server) structure based on the Web Service, there is no need for the clients to worry about the realization of the mining algorithms, the working platforms, the program languages, or the formats of data description. All the service details are transparent to them and they only need to know the address of the local server. The clients get the requests of the users, submit them to the server,

and receive the resulting documents returned from the server. Then the users can choose to see the results in a visualized form.

4 Related Work

Now there are already some Web mining systems based on Web Service. Such as the Web Service systems based on database——Amazon and Google. Both of them are better in the Recall and Precision than some other systems at the same time like Yahoo. Besides, some other network companies have also begun to create certain new services for developers, which can allow the developers access the databases of their companies directly without going through certain application procedures. However, all these services are just based on the elementary data sharing and have not gone deep enough into the data processing and data mining.

The TETRIS system, developed by the Fudan University [5], has made great efforts in the open character based on the service integration, from the whole architecture to mining language. It has provided a flexible platform for developing and applying Web mining. But at present this system is not mature enough in practice, and still needs some improvements.

5 Conclusions

Web data is so complex and enormous, there are many problems in the Web mining process. We propose the Web Service technology and suggest building a new Web mining system based on it, which can share and manage those semi-structured data from different kinds of platforms, strengthen the specialty of mining algorithms, improve the mining efficiency, and eliminate the closed character of mining results.

With the optimization and maturity of Web Service technology, it will surely be used in the Web mining much better. And the various web resources will finally become our treasure, thus making our lives fresher and more vigorous.

6 References

1. Yun, S., Han, L., Dong, J., Ceh, D.: KDW Survey: Web-based Data Mining. Computer Engineering. 29 (2003) 284-286
2. Zhang, S., Wang, C., Zhou, M., Zou, Y., Wang, W., Shi, B.: TETRIS: An Open Data Mining Application Development Platform Based on Service Integration. Computer Science. 30 (Spp.) (2003) 264-266
3. Zhang, Y., Tang, Y.: The Analysis of Distributed Computing Techniques and Web Service. Modern Computer, 1 (2004) 42-45
4. Kreger, H.: Web Services Conceptual Architecture. http://www-900.ibm.com/developer Works/cn/webservices/ws-wsca/index.shtml，May (2001)