The Application Research of Semantic Web Technology and Clickstream Data Mart in Tourism Electronic Commerce Website

Bo Liu

China West Normal University, Institute of foreign languages, Sichuan, Nanchong, China

Liubo123@163.com

Keywords: Semantic Web, Ontology, clickstream information, Multi-agent, Tourism Electronic Commerce.

Abstract. With the wide application of Web, the exploitation of clickstream information resources has a remarkable influence on the aspects of helping the website be adapted to the needs of users and improving the users' satisfaction of Web sites. However, at present the exploitation of clickstream information resources has been limited to the grammatical level, that is to say, it can only identify, reason and judge on the purely formalized level with low accuracy and low satisfaction degree of the users' interests. In view of this situation, this paper makes the ontology-based semantic web technology be introduced into the exploitation of clickstream information resources, and makes Web application system not only utilize the clickstream information on grammatical level but also integrate the clickstream information on semantic level. Thus, these things can improve the satisfaction of the users' interests and provide better services for Web users.

The Thought and System Structure of Semantic Web

According to Berners-Lee's viewpoint, the semantic Web is not entirely new Web, but it is the expansion of the existing Web. It is different from traditional Web because the semantic information can be well defined in the environment of semantic web, and it can make the computer and the human be able to work together better. In other words, the goal of the semantic web is to make the information on the Web can be understood by machines so as to realize the automatic processing of Web information. This way can adapt to the rapid growth of Web resources and provide better services for people.

W3C has described the semantic web like this: Making the data which the machine can understand be published on the Web is becoming a high-priority job of many organizations. Only the Web becomes an automated tool and platform for people to share and manipulate data, it can show its maximum potential. As far as this range of Web is concerned, the programs must be able to share and deal with the data in the future, even if they are individually designed. The semantic web is such an assumption: We make the data on the Web be defined and linked in a way that can be understood by the machine. It not only takes the display as the purpose, but also has the purpose of automated integration and reusing the data on different platforms.

The defined and linked data on the Web can not only display, but also can be processed, integrated and resued automatically. Only when data can be shared and processed automatically not only by human but also by machine, the Web can maximize its potential. However, the "machine-understandable way" does not mean that the machine can understand the language of human. It just shows that according to well-defined data the machine performs well-defined operations to solvewell-defined problems.

On December 18, 2000, Tim berners-lee officially proposed the layered architecture based on semantics at XML2000 Conference, namely the architecture of semantic web. The architecture consists of seven layers, as shown in Fig. 1.

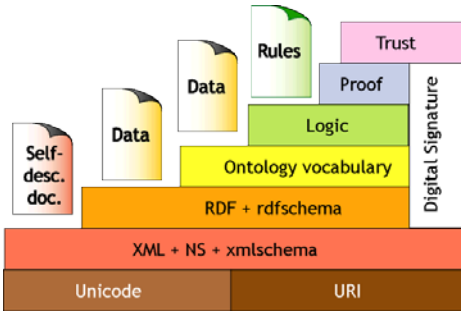

Fig. 1 The layered Architecture of the Semantic Web

The Scheme of Individualized Information Service Based on Clickstream Technology. When the way of individualized information service based on Clickstream technology compares with the present way of individualized information service, the advantages of the former is mainly reflected in the following three aspects:

It tracks the users' interest on information access dynamically so as to provide the users with the required information. Clickstream technology tracks the users' information behaviors. According to the latest information records which the users browse, it reflects the change of users' information needs timely and adjusts the recommended information timely to be more closer to the users' needs. It won't always give spams to the users because the users don't update the contents of custom information timely.

It can provide the users with the required information accurately. According to users' information behavior records, the analysis of the feature value combinations with information interest actually provides he advanced method of information query to provide the users with accurate and comprehensive information. The users may be not familiar with the information classification of website. The customized model of keyword requirements cannot reflect the users' interest completely and accurately. However, the way of individualized information service based on Clickstream technology tracks the user's information behaviors to reflect the users' actual needs with information better.

It doesn't need the users to customize or update the information needs. The traditional way of customized service with information requires the information users to provide the requirements of information change timely so as to send the correct information. The way of individualized information service based on clickstream technology can analyze the changes of the users' information needs automatically and send information.

System Implementation Techniques

The Extraction ofWebpage Keywords. The content of webpage is mainly composed of the text in the webpage. And the text is composed of words. When we parse the webpage in the website, as long as we make the words in the webpage be extracted, the content of webpage can be expressed. Nowadays most of webpages are composed of HTML language. And the text mentioned here refers to the headlines and text of the news and so on. However, the markup in the webpage, pictures, multimedia information,etc don't belong to the content of the text. When we process these things, we should remove them. And the remaining content constitutes a pure text information. However, the quantity of the remaining pure text information is still very enormous. If these information are used indiscriminately, it will increase the data dimension [26]. Therefore, in order to reduce the noise of data, we need to remove the modal particles, interjections and other meaningless words contained in the text information before we calculate weighting. Finally, we make the remaining pure text information be analyzed into the dictionary D which is composed of words.

We use the the vector space model W to represent the dictionary D in the document and use the TFIDF method to determine the weights of words. The TFIDF method is acknowledged as the best method at present which makes use of the words' frequency appeared in the document to determine the weights of the words. We make the words in the dictionary D represent as a n dimensional feature vector.

$$W=\{(d_1 ,w_1) ,(d_2 ,w_2) , \dots , (d_n ,w_n)\} \quad (1)$$

Every word in the dictionary d di has a weight Wi. We use TFIDF formula to represent the weight of words:

$$w_i=[0.5+0.5\frac{tf(i)}{tf_{\max}}][\log\frac{n}{df(i)}] \quad (2)$$

In the formula (2), tf(i) refers to the number of occurrences of the word di in the document(word frequency). df(i) refers to the number of documents containing the word di in the whole document. n refers to the total number of documents. Tf_{max} refers to the maximum frequency appeared in all the words(the maximum word frequency).

When the weight of the words are completed the calculation, we remove the repeated phrases in the result. For example, among {MP3, 0.53} and {MP3 download,0 46}, we only take the maximum word of the average weight, that is to say, we take {MP3 ,053}. Thus, we take out the top five words as the keywords which can best represent the webpage and write them in basic information database of the information processing center.

The Identification of the Users' Current Session. A website access of the users is called a session of the users. That is to say, a series of HTTP transactions which form the user's specific behaviors of the users when the users access a website are called a session of the users. This series of log records about HTTP transactions reveal all the behaviors of the users on a Web server. They include the first webpage which the user visits, namely the login page, and the last webpage, namely the webpage left off and so on which can show the users' interests and preferences and other information.By comparing the timestamps in the subsequent log records, we can calculate the time which the users spend on all the webpages except for the last webpage. Because there is no continuous log records on the webpage which the users left off, we cannot calculate the time which the users spend on the webpage. The time which the users spend on a webpage is called the sojourn time.

When the users first access the Web site, the Web server will generate a unique cookie and set up according to the users are accessing the file. Then when the users access any file on the Web site, the Web server will reset the cookie. Because the cookie is valid in the entire domain. The users' browser will send the cookie when the users request the webpage each time. Suppose that the browser receives cookie while the Web server is recording them, at the time the server can gather each request sent by the users and recreate their session. The following example shows that the server will record the content of the log file if the same user is ready to access 3 different pages on a Web site.

255.108.216.122- -[12/Jan/1997:00:00:00-0800]
"GET/index.html HTTP/1.0" 200 236 "-" "Mozilla/4.75 [en] (Win98;U)"
"-"
255.108.216.122- -[12/Jan/1997:00:00:00-0800]
"GET/page_one.html HTTP/1.0"200 237"-" "Mozilla/4.75[en](Win98;U)"
"827645474623743846="
255.108.216.122- -[12/Jan/1997:00:00:00-0800]
"GET/page_two.html HTTP/1.0"200 246"-" "Mozilla/4.75[en](Win98;U)"
"827645474623743846"

The value of Cookie appears in the last domain of the log records. It is worth noting that the first record of cookie value is "-", because the cookie must be set in the browser before it is sent back to the server. This means that it does not display the cookie value in the first record which the users access to the web sites.

A obvious shortcoming of this method is that it is anonymous. It only explains that someone is using a certain browser to access the website and which web pages are sent. Unless CookieExpires order is used, otherwise when the same user opens the browser and accesses the website again, the user will receive a new cookie and be regarded as another user.

The work which the clickstream data mart is responsible for is divided into two main parts: the first part mainly changes Web data into the form of business data (namely data preprocessing)which

is suitable for the excavation task; the second part mainly makes the mode discovery and the mode analysis.

Data Preprocessing, Data preprocessing includes three aspects: content preprocessing, structure preprocessing, and using information preprocessing. Content preprocessing refers to we change texts, images, sound, and other multi-media files into the form of Web data when we use the excavation; structure preprocessing refers to we link the structure design between web pages and frame image; using information preprocessing gets the service conversation by making data clearance, the users' recognition, the conversation recognition, the route perfection and business recognition with the web server log.

Mode Discovery, After the data is preprocessed, we will get corresponding business database. On the basis, we need to do some work with two aspects: one is we arrange and transform business database to data storage from which is compatible with certain excavation technology; the other is we use data excavation algorithm to excavate information and knowledge(namely, using the documents in grammatical hierarchy) which is effective, new, potential, useful, and understood in the end. The common technologies are special Web route analysis technology using excavation and common relevance rules, sequence mode, classification and clustering technology in data excavation field.

Mode Analysis, For the modes which are excavated using all kinds of technology, the number is very large and the expression is obscure. Appropriate tools and mechanisms are needed to help the analysis personnel to understand. So mode analysis technology and tool is one of the key technologies in the using process of Web excavation. It includes statistics, graphics visualization, workability analysis, intelligent inquiry and so on.

Visualization technology: It refers to using image user interface to help the users to excavate and understand lots of complex data. It offers lots of convenience for the users to manage and understand a large number of modes. Generally, we use graphics and images to denote the intricate relationship in abstract network. We use characteristics description to explain mutual functions among the modes. It will help to better understand the relationship among large number of data in Web, guide and accelerate the searching process.

Conclusion

The development of network technology and the appearance of electronic business have changed people's views and practical behaviors in the aspects, such as what the transaction is and how to operate it. It makes the prospect of economic development show a new complexion with network and informatization. Although there are some limitations, such as tedious information, no recognizing exactly time stages and visitor, lack of authenticating matters etc, in clickstream data which bring lots of difficulty for actual operation. But the corresponding tools which are for data excavation are in the process of the development. At the same time, clickstream data mart is flourishing too. Even though the key technology—extraction, transform, and load mechanism(Extract-Transform- Load，ETL)still does not have the corresponding software which can deal with clickstream data well, currently the main ETL suppliers such as Informatica, Ardent and Sagentm, can offer the primary supports for the clickstream data. In a word, clickstream research is a new research model which is cross-domain and has strong technical relevance. It has broad development prospect.

References

[1] Tim Berners-Lee,James Hendler.Ora Lassila. The Semantic Web. Scientific American, 2001,184(5):35-43.

[2] Dan Brickley. Semantic Web Technologies.

http://www.jisc.ac.uk/uploaded_documents/jisctsw_05_02bpdf.pdf

[3] RDF. http:// www.w3.org/RDF/

[4] Erol Bozsak, Marc Ehrig, Siegfried Handschuh et al. Kaon-towards a large scale semantic web, In Proceedings of the Third International Conference on E-Commerce and Web Technologies(EC-Web 2002), Springer Lecture Notes in Computer Science,2002.

[5] Pérez A G, Benjamins V R. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods. In: Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods. Stockholm, Sweden, 1999.

[6] Michael K.Smith,Chris Welty,Deborah L.McGuinness. OWL Web Ontology Language Guide. http://www.w3.org/TR/2004/REC-owl-guide-20040210/.