Acta Technica 62 (2017), No. 4A, 283–288

Research on the present situation and development of web data mining

Zeng Jun[2]

Abstract. With the database, data warehouse, data warehouse technology based information system application in all walks of life, making huge amounts of data being produced. There are so many data makes it difficult to digest, they cannot see the useful information contained on the surface, not to mention the effective guidance for further work. To find useful information from a large amount of data has become the focus of attention, data mining technology is accompanied by a demand from research to application.

Key words. Web, data mining, present situation development.

1. Introduction

In recent years, with the rapid popularization and development of Internet/Web technology, make all kinds of information can be available at very low cost in the network, because the Internet/WWW in the global interconnection, the amount of data can be obtained from it is difficult to calculate, and the development trend of Internet/WWW continued bullish, especially the rapid development of e-commerce provides a strong support for network applications, how to WWW the world's largest collection of data found useful information will undoubtedly become the focus of research on data mining.

2. Methods and Materials

Web mining refers to the use of data mining technology in the WWW data. The potential and useful patterns or information.Web mining research covers a number of research areas, including database technology, information acquisition technology,

[1] Acknowledgment - Fund project of Chongqing Municipal Education Commission, parallel mining research, KJ15012021 big data under the Hadoop framework; the Ministry of education "spring" project, preliminary application of big data intelligent agriculture platform networking, S2016038.

[2] Workshop 1 - College of Computer Engineering of YANGTZE Normal University,ChongQing,China

statistics, machine learning and neural networks in artificial intelligence.As shown in Figure 1.

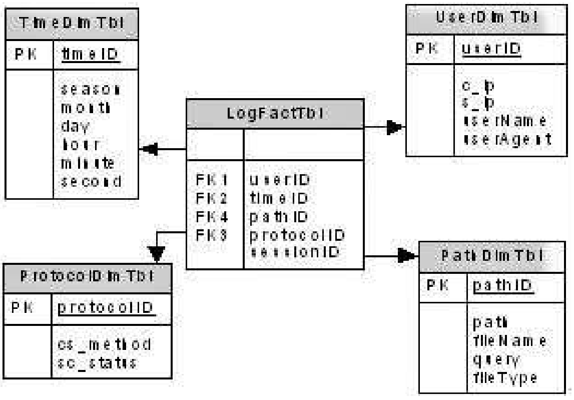

Fig. 1. For mining Web log data warehouse system

3. Analysis and discussion

3.1. Web mining process

Compared with the traditional data warehouse, Web information is unstructured or semi-structured, dynamic, and confused, so it is difficult to directly to Web data on a web page data mining, and through necessary data processing. The typical process of Web mining are as follows[1]

Find resources: the task is to get data from the Web document, it is worth noting that sometimes information resources are not limited to online Web documents, including electronic documents, e-mail, newsgroups, or even web log data is formed by Web in transaction database data.

Information selection and preprocessing task is obtained from the Web resource information and eliminate useless information to make the necessary arrangement. For example from the Web document automatically remove ads connection, remove redundant format tag, [2]automatic identification and data fields or passages are organized into logical form regular even form.

Pattern discovery: automatic pattern discovery. Can be carried out within the same site or at multiple sites.

pattern analysis: verification, a step on the interpretation of patterns can be done automatically. The machine, can also interact with the staff to complete the analysis. Web mining as a complete system, in the mining information before IR (Information Retrieval) and information extraction (Information IE Extraction) is very important. The information gain (IR) aims to find relevant Web documents, it is only the data in the document as a set phrase without sorting, and information extraction (IE) aims to To find the data needs of the project from the document to the document, it structures the meaning of the expression of interest, it is an important task is to organize the data and information obtained. [3]Appropriate index (IR) and information extraction (IE) technique has been for a long time, with the development of Web Technology Web technology, based on IR, IE gets more and more attention. Because the Web data is very large, and dynamic change, using the original manual way of collecting information already inadequate, the direction of the research is to use automated, semi automated method in Web on IR and IE. in W EB environment to handle unstructured and semi-structured document, data processing, in recent years there are two corresponding research results and specific applications, especially has a very good application in the major search engines.as in Figure 2.

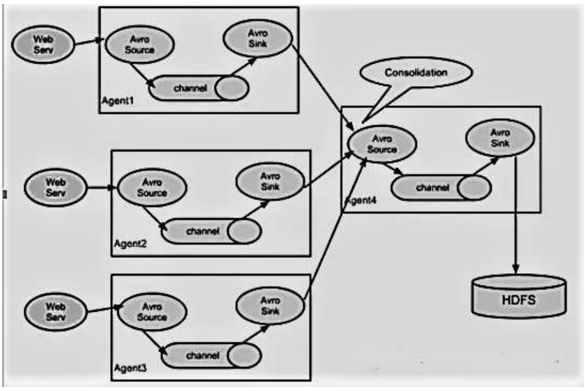

Fig. 2. How do you analyze web sites by using data mining techniques

3.2. Classification and research status and development of Web mining

According to the different degree of interest in the Web data, Web mining can generally be divided into three categories: Web content mining (Web Content mining), Web structure mining (Web structure mining), Web usage mining (Web usage mining)

Web content mining refers to the content from the Web / data / document found useful information, all kinds of information on the Web, the traditional Internet from various types of services and data sources, including WWW, FTP, Telnet and so on, there are now more data and port can be used, such as the government information service, digital library electronic commerce, data objects, and various other Web can access the database through.Web content mining including text, image, audio, video, multimedia and other types of data. [4]The unstructured text Web Mining is classified into text based on Knowledge Discovery (KDT), also known as text data mining or text mining is an important technique in the field of Web mining. It has attracted the attention of many researchers. In a recent Web multimedia data mining has become another hot point. Web content mining in general from two a different point of view to study resource search (IR). From the point of view, Web content mining task is from the user's point of view, how to improve the quality of information and help users to filter the information. From the viewpoint of DB Web content mining task is mainly to the number of Web According to the integration, modeling to support complex queries on Web data.

Web usage mining (Web usage Mining): Web usage mining, has important significance in the emerging field of e-commerce, by mining Web log records related to the discovery of user access patterns of Web pages, by analyzing the log records law, can identify the user's preferences, loyalty, satisfaction, can find potential customers, enhance service competitiveness.Web site using data in addition to the server log records include proxy server logs, browser logs, registration information, user session information, transaction information, Cookie The information in the user query and the possible interaction between recording mouse click stream all users and site. The amount of data visible Web usage is very huge, and the data types are abundant. According to different treatment methods on the data source, Web usage mining can be divided into two categories, one is the use of Web the recorded data conversion and transfer relations in the traditional, then use data mining algorithms for conventional mining on the relationship between the data in the table; the other is the direct use of Web data processing pre recorded and then tap an interesting problem in.Web usage mining in multiple users using the same How to identify a user proxy server environment, how to identify the belonging to the user session and use records, this problem seems small, but has a great influence on the quality of mining, so it is specially studied in this area. Generally speaking, the classical data mining algorithms can be directly used in Web usage mining, but in order to improve the quality of mining, the researchers conducted efforts in the extended algorithm, including complex association rule algorithm, the improved algorithm sequence was found. According to the data sources, data types, the number of users in the data set, the data set of the number of servers will be Web Usage mining is divided into five categories: the personality Mining: for individual users to use records to model the user, combined with the analysis of the basic information of the user and his habits, personal preferences, the purpose is to provide personalized service for the user out of the ordinary in the electronic commerce environment. The improved system: Web (network service database, etc.) the performance and other service quality is a key indicator of user satisfaction, Web usage mining can record the user's congestion to find the performance bottleneck of the site, the site administrator to prompt the improvement of Web cache strategy, network transmission strategy, distribution of traffic load balancing mechanism and data Strategy. Moreover, through illegal intrusion data analysis network system to find the weaknesses, improve the site safety, which is particularly important in the electronic commerce environment. - site modification: structure and content of the site is the key to attract users of.Web usage mining by mining user behavior records and feedback for site designers to provide improved in. For example, how to organize the page links, those pages should be able to directly access. - Business Intelligence: the user how to use the Web site information is undoubtedly the focus of e-commerce vendors, users visit cycle can be divided into attracted, resident, buy and leave four Step, Web usage mining can be analyzed by the user click stream Web log mining user behavior, to help dealers to arrange the sale strategy. The description Web features: this kind of study with such attention through the interaction of the user site visit statistics of each user on the page, the user access feature description the situation.

4. Conclusion

Although the form and research direction in Web mining emerge in an endless stream, but I believe that with the rise and the rapid development of electronic commerce, an important application of Web mining will be the future direction of the electronic commerce system and electronic commerce. And most closely related to the usage mining (Usage Mining), that is to say will be at this the field has been paid more and more attention. In addition, in the study of search engine research, structure mining has been relatively mature, the contents of the text mining based on also there have been many studies, the next step will have more researchers put the research direction for the multi media mining.

References

[1] J. W. Han: *Concept Techniques.* Morgan Kaufmann Publishers (2001).

[2] : *http://www.billinmon.com.*

[3] R. H. Blockeel: *Web research a survey.*SIG KDD Explorations

[4] J. Srivastava, R. Cooley: *Web usage and of usage from web data.*SIGKDDExplorations *1* (2000), No. 2, 12–23.

Received November 16, 2016