Research of Web mining Technology based on XML

Li lan
College of Computer Engineering
Qingdao Technological University
Qingdao, P.R.China
lgdll@qtech.edu.cn

Rong Qiao-mei
College of Computer Engineering
Qingdao Technological University
Qingdao, P.R.China
qdanny@yahoo.cn

Abstract—Web Data Mining is a new important research field in data mining. In this paper, the conception and characteristic of data mining based on Web are expatiated. At present many websites are built with HTML, which is difficult to achieve real effective and accurate web mining. The appearance of XML has brought convenience for it. Based on the research of web mining, XML is used to transform semi-structured data to well structured data, and a model of web mining system which has basic data mining function and faces multi—data on the Web is built. At the same time, the problem in data mining is analyzed and studied. An example is put forward to prove the solution.

Keywords—Web mining;XML; Multi-Agent system

I. INTRODUCTION

With the increasing popularity of the Internet, through Web access to more data and information, work, study and life style of the great changes taking place, much higher efficiency, resources of information are the greatest degree of sharing. However, because of the web page is too complicated and there is no structural, dynamic, leading it difficult to quickly and easily on the Web to find the necessary data and information for this Web mining research in the field of high technology has become the hot spot[1]. In the face of Internet-based “information ocean”, we need to extract useful knowledge can guide the decision-making. XML can be due to different sources of structured data easily be combined so that the search for diversification, incompatible databases possible for Web data mining has brought new opportunities[2]. This paper is application research in Web Data Mining based on XML and Agent, bring forward a kind of XML- based distributed data mining architecture.

II. WEB DATA MINING

A. Web data mining concepts and categories

Web data mining technology[3] and data mining is a combination of Web, is an integrated technology resources extracted from the WWW information (or knowledge) of the course, is the implication of Web resources, interest, unknown, The potential value of the mode of extraction[4]. It repeated use of a variety of data mining algorithms from the observation data to identify patterns or a reasonable model, but also to data mining technology and application of the theory of www resources to carry out excavation of a new research field[5].

According to the Web targeting a different group of mining, Web data mining is divided into: Web content mining, Web mining and the structure of the Web Usage web data mining classification as shown in Figure 1.

![Figure 1. Classification of Web data mining](image)

Web content mining from the document described its contents or takes a course of interesting knowledge, is a web-based content of the elements of the target Web mining[6][7]. These elements have targeted text and hypertext data as well as graphics, images, and other multimedia data; both from the database of structured data, it also uses XML tags of HTML or semi-structured data and unstructured text of the free.

Mining is the structure of the Web page from the hyperlink found in its structure and its relationship with each other. Through to find hidden in a page after the link structure of the model will be able to take advantage of this model on the Web page re-classification, can also be used to find similar sites. Based on the hyperlink topology, Web mining structure can be classified pages, summed up the page and site structure, such as the generation of similarity between the Web site, the relationship between the Web site.

Web Usage Mining is the user "visit marks" to obtain valuable information on the Web log data and data mining. These data include: client, server-side data and data-side proxy. Web Usage Mining can be divided into general and special access to track the path of track. The former is used KDD[2] (Knowledge Discovery in Database, access to knowledge from the database) to visit the general understanding of the technical patterns and trends, such as Web log mining; the latter is an analysis of each and every time the user visits the model, on the basis of these sites will
automatically Mode Built structures, such as adaptive site. Web use records of the excavation is aimed at forecasting the on-line users, compared with the actual site and look forward to the use of the difference, according to the user’s interest to adjust structure of the site[4].

B. XML in Web Data Mining

XML has become a standard, developers can use the XML format for data exchange and tags. XML in the three-tier system for data processing provided a good way. The use of XML, Web designers can create text and graphics, but also to build a document type definition of multi-level, interdependent systems, tree data, metadata, hyperlinks and style sheets.

Web-oriented data mining is a complex technology, which enables the structure of the different sources of data easily combined, making the search for diversification is not compatible database possible, so as to solve problems with Web data mining to hope. XML’s flexibility and scalability is to allow XML to describe different types of applications in the data, which describes the Web page to collect the data record[8]. At the same time, based on the XML data is self-described, the data do not need to be able to describe the internal processing and exchange.

C. Agent in Web Data Mining

Agent technology in artificial intelligence, particularly the Internet network technology development and decision support systems based on technology developed. Agent self-control to the state and behavior, no one can or other procedures involved in the operation and when to run. In data inquiries by the Agent to complete a complex examination of information, analysis and processing, can form intelligent data warehouse[7]. Agent can be carried out in collaboration with the other Agent, interactive, making different locations between data can be easily shared, rules and methods, and other resources. It can be used to support distributed component technology, so as to solve the data warehouse as a result of heterogeneous information sources, the distribution of the resulting information can not be fully interactive, the problem is incomplete, to better support decision-making groups. In addition, the Agent can move independently in the heterogeneous network, in accordance with certain rules of the mobile to find the right computing resources, information resources or software resources, and utilization of these resources in the same network or a host of advantages, processing or use of these Resources on behalf of the user to complete a specific task[9]. Agent mobility with the view to resolve is the issue of safeguarding the provision of a new program. For these reasons, the Agent will be the introduction of data mining techniques, better, faster decision-making, Agent play the characteristics of the source of information to maintain the autonomy, independence.

III. WEB DATA MINING MODEL BASE ON XML

A. Principle

The current Web page is to into XML format, and use the tools to deal with the structure of XML data in order to extract the appropriate data. HTML files can be used to correct common errors in the layout and format to generate the equivalent of a good document, you can use Tidy generate XHTML (XML subset of) the format of the document. By constructing a XMLHelper to complete the Java-type data from XML to HTML conversion, as well as with other XML-related tasks. Data extraction process is shown in Figure 2.

![Figure 2. The process of extracting data in Data Mining](image)

The main steps are as follows:

- Identify the source of data and map it into XHTML. In most cases, the source of information is obvious, but in a dynamic environment to be extracted for use, reliable and stable sources of information more difficult. To determine the source of information, through the structure, called the Java class of XMLHelper to complete the data from XML to HTML conversion..

- To find the data points used. Both the Web page and XHTML source in view of the vast majority of information has nothing to do with the information collected, the next in the XML tree to find a specific region, the need to extract the data. We find the data generally contains the same elements <table> this table will contain general information required for key words, the note observed, the analysis of the page generated XHTML, and the table as Reference points, or anchor.

- Data will be mapped into XML. You can create data taken from the actual codes when you find the anchor, the code will be XSL file the form. XSL document is intended to anchor logo, and specify how to get from the anchor is looking for data, and by that we needed to construct an XML format output files.

- Combined results of the data. If only the implementation of a data extraction, in accordance with the above-mentioned steps have been completed. However, Web data mining is a week of back and forth, a few simple data collected has not yet completed the task of data mining. Web data mining for the special, it is necessary to keep the Internet on the collected data and the results into XML data files.
B. System Model

Web data handling and standardized system for Web mining provides a good source of data, the use of data mining ideas and Web mining technologies has become the focus of the design[10].

Through the above analysis of XML-based Web Data Mining System Model of Web mining is a basic function of mining for a variety of Web data mining system model. The basic framework is shown in Figure 3.

![Diagram](image)

Figure 3. The basic framework on Web data mining

C. Mobile Agent technology into Web Mining System

- Mining algorithm based on the Agent. Each standard for data mining made an Agent. From the user’s excavation mission at the time of distribution can create a mining algorithm based on the Agent, and then moved to the Agent on the implementation of the objectives of the host mining tasks[11][12]. End in one place after the excavation can be moved to another host on the excavation.
- Based on the data source Agent. Each Agent and data source match. For a data source, an Agent can be excavated to express a variety of forms of knowledge, such as clustering, association, the concept of knowledge level.
- Mixed Agent. Take into account the actual needs of mining, excavation and sometimes uses a method can not get a better result, the need for a multi-Agent in the integration algorithm. But the presence of an Agent is not a data source, not just a mining algorithm. Agent can be mixed in the target mobile host asked.
- Multi-Agent system is composed of, on behalf of each of the Agent a special data mining classifier, classifier for data classification. Classification of data mining is a typical mission, known instance of a set of attributes and their values to a new example of some of the attributes to predict when, it is necessary to use classification.

IV. System Framework

A. Agent adapter

Agent adapter used to initialize Agent, to realize the communication between the Agent, as well as data mining and long-range communications systems. It kept Agent is running, running position, the current use of system resources. Agent in each adapter in alias, so that when Agent communication need to know the alias will be able to without having to know the location of the operation. Another adapter is a function of decomposition mining request, and then sent to the appropriate DMAgent, in DMAgent completion of the excavation process, the comprehensive results of the excavation DMAgent longer to process applications.

B. Agent search engines and Web data set

Agent uses multi-search engine Agent architecture of the Web to achieve the parallel, distributed process-ing, in order to solve large-scale Internet to gather information and improve the accuracy of information retrieval[6]. Agent between the knowledge-based re-asoning model of organizational forms, with different members of the machine learning methods to achieve the basic functions and high-level collaboration, to design the structure of the members of the Agent prior knowledge come from search-based keyword matching, Catalog-based classification and retrieval based on the hyperlink, such as retrieval of the search algorithm. Web Database Management Agent Search Agent will have access to the results of the format, and regularly update the Web to generate data sets.

C. Agent Server

It has preserved the local source database data in the form of local data, table structure, such as the scope of the data. Each Agent Server adapter to register through the registration, each adapter Agent Server knows the source of data, allocation of tasks in order to produce its desired effects.

In the system design process we take into account the operation of distributed will be digging into mining request queue for each of the agents to enter data mining incident distribution of a time stamp, to achieve simultaneous events, with this balanced approach to migration in all[7]. Deputy mining system load and at the same time reducing the costs of communication between the incident and improve availability.

V. Conclusion

With the computer technology and the rapid development of Internet technology, data mining has put forward higher requirements. Data mining technology is facing more problems and challenges: how to improve the efficiency of data mining methods; the development of adapt more types of data to allow for the noise of the excavation methods; dynamic data and knowledge of the data mining; network and distributed environment, such as data mining. In addition, in recent years database rapid development of multimedia oriented database mining and the next thing will
be the focus of research and development. Therefore, as a network of data mining Main way the web of data mining technology to further.

In this paper, object-oriented database to create a type of data dictionary approach, the establishment of such a dictionary definition of the structure of the method to mapping pointer corresponds to the type of content will be targeted at different management record, so that improved the efficiency of storage, faster and at the same time Considerable savings in storage space.

REFERENCES