Web data mining Using XML and Agent Framework

S.Mukthyar azam[1]

S.C.E.T., Hyderabad, India.

M.Kiran Kumar[2]

RVR&JC College of Engg. Guntur, India

Shaik Rasool [3]

S.C.E.T., Hyderabad, India.

S.Jakir Ajam[4]

Hyderabad, India.

Summary

The World Wide Web is one of the fastest growing areas of intelligence gathering. Currently many websites are built with HTML tags, one problem associated with retrieval of data from web documents of HTML is that they are not structured in traditional databases because the Web pages created using HTML are semi structured thus making querying more difficult than with well-formed database containing schemas and attributes with defined domains. The concepts of XML have brought convenience for it. Based on the research of web mining, XML is used to convert semi-structured data to well structured data, and a model of web mining system which has basic data mining task and faces multi data on the Web is constructed. This paper"Web Data Mining using XML and Agent Framework" bring forward a kind of XML-based distributed data mining architecture. At the end of the discussion problems in data mining is analyzed. An example is put forward to prove the result.

Keywords

XML, Web mining, web structure mining, Multi-Agent system.

1. Introduction

With the rapid development of computer network and multimedia technology and the people 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 problematical and there is no Structural, dynamic, leading it difficult to fast 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 Web-based Data package, we need to extract useful knowledge can guide the decision-making. XML has ability that different sources of structured data easily be combined so that the search for diversification, incompatible databases possible for Web data mining has brought new chance [2].

2. Background of Web Mining

Web mining is a inclusive technology, related to web, data mining, information standard and other fields of science. It can be defined as the analysis of the relation among the content of document, the use of available resources, to find the knowledge which is effective, potentially valuable, and eventually understandable, including the non-trivial process of patterns, rules, regularities, constraints and visualizations and etc.[9]

2.1. The Categorization and realization of Web Mining [10]

The methods of data mining can be divided into two kinds:

1) Based on statistical models and the technology used includes decision tree, classification, clustering, association rules, etc.

2) Establish an artificial intelligence schema mainly based on machine learning, the methods used include neural network, natural law calculation method, etc.

There are three popular fields of Web Mining: web usage mining, web content mining, web structure mining. Its detailed structure is illustrated as Fig.1.

(i)Web Content Mining [11]: web content mining mainly bases on the text information mining, its method and function is usually very similar with plane text mining. Using parts of tags of web text, such as title, head etc. which contains additional information can improve the performance of web text mining.

Web content mining refers to the process of mining from the content of web pages or its report, and extracting the knowledge. There are two kinds of web content mining according to the objects of mining:

Web content mining refers to the process of mining from the content of web pages or its report, and extracting the knowledge. There are two kinds of web content mining according to the objects of mining:

Manuscript received May 5, 2010
Manuscript revised May 20, 2010

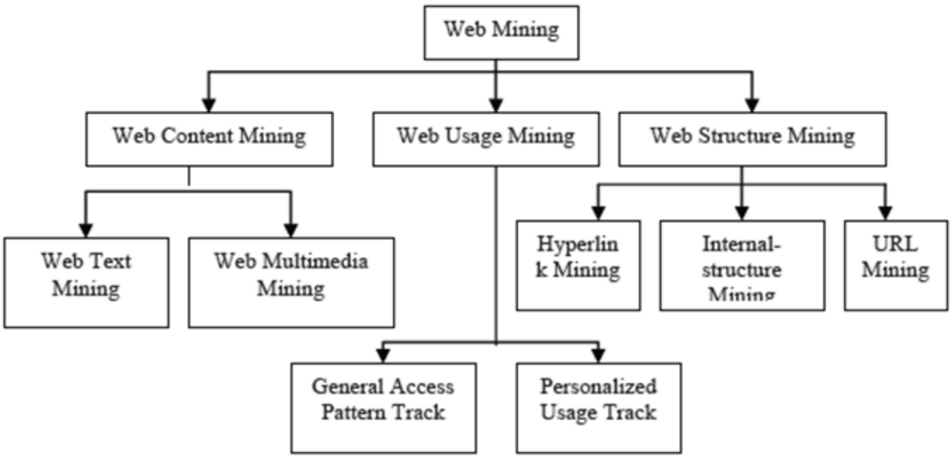

Fig. 1. Classification of Web Mining

- Text documents mining, including the text format, HTML tag, o uses XML tags of HTML or Semi-structured data and unstructured text of the free format and so on.
- Multimedia documents mining, including Image (*.jpg,*.gif, jpeg etc), audio (*.mp3,*.mp4,*.wav so on), video (*.mpeg,*.vlc so on) and other media types.

In web content mining refers to the process of mining from the content of web pages from the hyperlink found in its structure and its relationship with each other.

Text conclusion can extract key information from documents, and summarize and explain the content of the documents with a concise form, so that users do not need to browse the full text. The purpose of text conclusion is to concentrate the text information, and give out a compact description

1) Using part of speech tagging to analyze the segmentation.

2) Using statistical method to extract the high-frequency words and determine the Summary.

Text clustering refers to the combination of a group of objects, and the group of objects can be divided into several categories according to similarity. Its purpose is to divide the document collection into several clusters, and its request is that the similarity of document content in the same cluster should be as much as possible, while the similarity between different clusters should be as small as possible. We can use text clustering to provide the summary of the large scale document content; identify the similarity between hidden documents; reduce the process of browsing related or similar information.

Text classification is the core of text mining .Automatic text classification refers to use a large number of texts with class signs to train classification rules or model parameters, then use the training result to identify the text of which type is unknown. It not only allows users to easily browse documents, but also makes the search of documents more convenient by limiting the search scope.

Discovering the algorithm of associate rules always has to go through the following three steps: data connecting, for the preparation of data; give minimum support and minimum reliability, discover associate rules through the algorithm provided by data mining tools; visualized display ,understand and Assessment associate rules.

(ii)Web usage mining: It refers to mine information from the access logs left on the servers when users visit the web. That means carry out mining from the access methods of visited web sites in order to find out the browse patterns when users visit web sites and the frequency of visiting the pages. There are two kinds. Tracks in the analyzing of users' browsing patterns, the first one is the general access pattern track for user groups, and the second is the personalize use record track for single user. The mining objects are in the serves including the logs such as Server Log Data.

Theses information includes a client-IP, server-side data, authoritative pages and data-side proxy. Web Usage Mining can be divided into general and special access to track the path of track. The past is used KDD[2] (Knowledge Discovery in Database, access to knowledge from the database) to visit the common understanding of 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.

Web Usage Mining performs mining on web usage data, or web logs. A web log is a listing of page reference data. The data for describing users' access in web usage mining includes IP address, reference pages, access date and time, web sites and their configuration information.

There are two kinds method for discovering usage information: One kind is that analyze through log files, including two manners:

1) Pretreatment that is the log data will be mapped into relationship list and use the corresponding data mining technology to access log data.

2) access log data directly to obtain the users' navigation information. The other kind is that the users' navigation behavior can be discovered through the collection and analysis of users' click events.

(iii)Web structure mining: It refers to derive knowledge from the organizational structure of World Wide Web and the relationship of links. As a result of the interconnection of the documents, World Wide Web can provide the useful information besides the content of documents. Making use of this information, we can sort the pages and find the most important pages among them. On behalf of work in this area have Page Rank and Clever. Web structure mining not only includes hyperlink structure between documents, but also includes the internal structure of the documents, the directory path structure in URL.

2.2. XML Usage in Web Data Mining

XML is a way of marking up data, adding metadata, and separating structure from formatting and style. Developers use XML format for data exchange and tags. Because XML is simple .it contains only data and markup. Look comes from a separate style sheet and links are separated not buried in the document .Each can be maintained separately for data processing provided a good way.

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 records [5]. 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

2.3. Agent Framework in Web Data Mining

Agent Framework in artificial intelligence and neural network, particularly the Internet network Technology development and decision support systems based on technology developed. Agent self control to the mode and state, 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 [4]. Agent and can be carried out in collaboration with the other Agent, interactive, making different locations between data can be easily shared, functions and methods, and other resources.

It can be used to support distributed environment, so as to solve the data warehouse as a result of heterogeneous information sources, the distribution of the resulting information cannot 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 [6]. 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.

3. Web Data Mining Process Based on XML Technology

3.1. Basic Principle

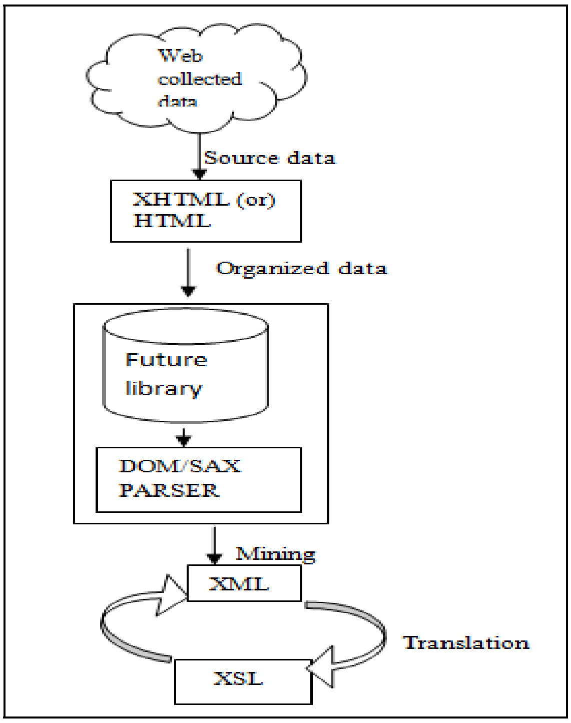

Fig.2 The process of extracting web source data in Data Mining

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 error 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 XML Helper 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.

The main steps are as follows:

A. Recognize the source of data and map it into XHTML (Or) HTML. 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 XML Helper to complete the data from XML to HTML conversion.

B. 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.

C. Data will be mapped into XML. You can create data taken from the actual codes when you find the anchor, the code will be Extensible Style sheet Language 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.

D. Joint 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.

3.2. Web mining Model

Compared with data mining based on relation database or data warehouse web data mining is much difficult. If think the web as a huge and distributed database, each site is an independent data source, and their data organization forms and structures are not the same. Therefore, the information on the web can be regarded as a heterogeneous database environment. In addition, apart from the heterogeneous of different sites, the large number of data on web pages is always some texts and multimedia information which is semi structured or unstructured, because of this, it is necessary to do some data processing instead of mining date on web pages directly.[7][12] For example, in the online education platform we can regard the resources of each course as an independent web site, and integrate the independent course into a full resource through web mining.

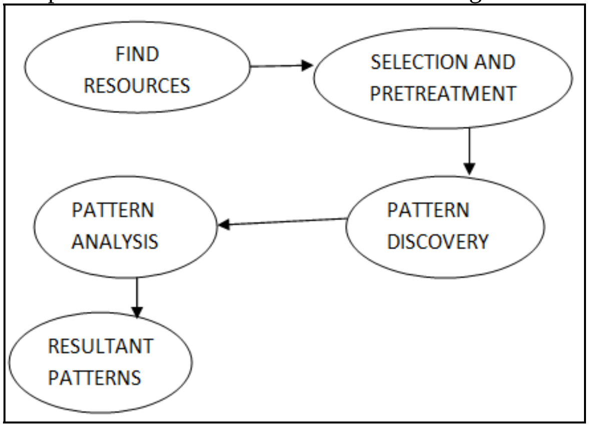

Fig.3 The basic framework on Web data mining

Typical web mining process is as follows:

A. Find resources: its task is to obtain data from goal web documents; it is worth noting that in some case information resources are not limited to online web documents, but also include email, electronic documents, news groups or web site log data.

B. Selection and pretreatment of information: its task is to remove the useless information from the obtained web resources, and to take some necessary editing for the information. For example, automatically remove the ads, redundant tag format, automatic identify paragraphs or field from web documents, and organize data into a logical form even relationship tables.

C. Pattern discovery: automatically discover the patterns; it can be achieved within the same site or between multiple sites.

D. Pattern analysis: verify and explain the pattern generated by the previous step. It can be finished automatically by machines, as well as by the interaction with analysts.

3.3. Mobile Agent program into Web Mining System

In the context of computer science, an agent is an independent software program that runs on behalf of a network user [14]

- 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 [8][13]. End in one place after the excavation can be moved to another host on the excavation.

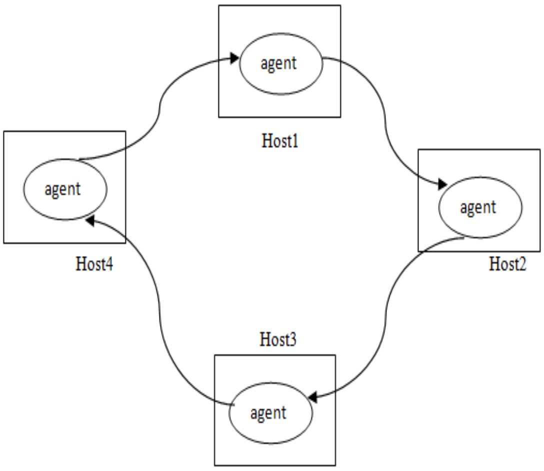

Fig.4 The mobile agent travel from host to host [14]

- 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 cannot 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.

4. System framework

4.1. Agent adapter

In mobile agent framework, Adapter agent used to start Agent, to realize the communication between the Agents, as well as data mining and long-range communications systems. Each Adapter implements the Adapter Agent interface. The Adapter Agent communicates with the local host and provides a consistent interface to the other system components. The Adapter Agent maintains a local table with available services provided by the host and known services that the host uses. Another adapter is a function of decomposition mining request, and then sent to the appropriate Data Mining Agent, in Data Mining Agent completion of the excavation process, the complete results of the excavation Data Mining Agent longer to process applications.

4.2. Agent search engines and Web data set

Agent uses multi-search engine, Agent architecture of the Web to achieve the parallel, distributed processing, in order to solve large-scale Internet to group information and develop the accuracy of information retrieve [3]. Agent between the knowledge-based reasoning 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.

4.3. Agent Server

A mobile agent system includes several parts and the Mobile Agent Server is the core part. Mobile agent administrator manages services provided by Mobile Agent Server and can cooperatively work with the services. The operations on it are accomplished through user's interface

It keeps the status of agents moving in the network, gives the mechanism of queue and management for the execution of agents for mining, maintain agents. It is responsible for distinguishing users and authenticating their agents, and protects servers' resources and ensures the security and integrity of the agents and their data objects when moving in the network. It also controls dynamic load needed of Java class library to agents.

It provides a mechanism for developers, which makes it possible to add some related services, and to access local services on the destination server. Remote management is executed through Remote Administration API.

5. Conclusion and Future work

Based on the research of web mining, XML is used to convert semi-structured data to well structured data This paper ”Web Data Mining using XML and Agent Framework” bring forward a kind of XML-based distributed data mining architecture. There are so many data mining algorithm for knowledge existing on the web to be discovered ,shared and utilized but there are quite a few problems existent, such as the inadequate utilization of network resources and the lack of individuation of the existed platforms so future work is: how to improve the effectiveness of data mining methods; dynamic data and knowledge of the data miming; network and distributed environment, such as data mining; the development of adapt more types of data to allow for the noise of the dig methods.

References

[1] Beale Russel1.Supporting serendipity Using ambient intelligence to augment user exploration for data mining and webbrowsing[J]. International Journal of Human ComputerStudies, 20007, 65(5) : 421-433.

[2] S.Chakrabari. Mining the Web: discovering Knowledge from Hypertext Data. Morgan Kaufmann, an Francisco, CA, 2003

[3] J.D. Velasquez, H. Yasuad, T. Aoki, etc., A generic Data Mart architecture to support Web mining, Management Information Systems, Vol.7,2003, 589-599.

[4] Ian H.Witten, Eibe Frank.Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Elsevier Inc, 2006

[5] P.s.Bradley.J.Gehrke, R.Ramakrishnan, and R.Scaling algorithms to large database.Communications of the ACM, 45(8):38- 43, 2002.

[6] R.Agrawal and R.Srikant. Privacy-preserving data mining. In Proc.of 2000 ACMSIGMOD intl conf.on Management of data, Pages239-245

[7] CHEN Yu-ru,HUNG Ming-chuan, Don-Iin YANG.Us|ng data mining to construct an intelligent web search system[J].International Journal of Computer Processing of Oriental Languages,2003,16(2)

[8] N.R. Raghavan, Data mining in e-commerce: A survey, Academy Proceedings in Engineering Sciences, Vol. 30, No.2, 2005, 275-289.

[9] Wang Qijun and Shen Ruimin, “Studies on Web Mining Based Intelligent and Personalized Distance-learning Environment”, Computer Engineering, Vol. 26, No. 12,2000, 158.

[10] Zhang Tao and Deng Jun, “The Research of Modern Distance Education Personalized Web Mining”, Science Technology and Engineering, Vol. 7, No. 5, 2007, 742-743.

[11] Tu Chengsheng, Lu Mingyu and Lu Yuchang, “Research on Web Content Mining”, 2003, 6-8.

[12] Li Jian, Xu Chao and Tan Shoubiao, “Design and Research of a Web Data Mining System”, Computer Technology and Development, Vol. 19, No. 2, 2009, 70-72.

[13] M. Jelena, K. Regina, Data mining technique for collaborative server activity analysis, WSEAS Transactions on Information Science and Applications, Vol. 2, No. 5, 2005,530-533.656.

[14] M.L.Liu, “Distributed Computing Principle and Applications”, Dorling Kindersley Pvt. Ltd., Licenses of Pearson Education, 2007 pages (413-414).

S. Mukthyar Azam received the Bachelor of Technology in Information Technology from Jawaharlal Nehru Technological University, Hyderabad, India in 2006, and Master of Technology in Computer Science & Engineering from Jawaharlal Nehru Technological University, Kakinada, India in 2009.He did Diploma in Embedded Systems Design from Centre for Development of Advanced Computing (Scientific Society of Department of IT, Ministry of Communications and IT, Govt. of India), Noida, India. He is currently an Assistant Professor at the Department of Computer Science in S.C.E.T., Hyderabad, India. His main research interest are Data mining, Information Security, Biometrics, web Information Retrieval, Programming Language Design.

Macharla Kiran Kumar received the Bachelor of Technology in Computer Science & Engineering from Acharya Nagarjuna University, Guntur, India in 2004, and Master of Technology in Computer Science & Engineering from Acharya Nagarjuna University, Guntur, India in 2009. He is currently an Assistant Professor at the Department of Computer Science in R.V.R & J.C College of Engg, Guntur, India. His main research interests are Data mining, Information Security, Biometrics, web Information Retrieval, Image Processing, Discrete Structures, and Network Security.

Shaik Rasool is received the Bachelor of Technology in Computer Science & Engineering from Jawaharlal Nehru Technological University, Hyderabad, India in 2008. He is currently pursuing Master of Technology in Computer Science & Engineering from S.C.E.T., Hyderabad, India. His main research interest includes Data mining, Network Security, Biometrics, Information Security, Cloud Computing, Programming Language and security and Artificial Intelligence.

S. Jakir Ajam is received B.Sc. from Nagarjuna University, Guntur, India, and Master in Computer Application from Indira Gandhi National Open University,Delhi ,India. He is a software engineer in Datawarehousing domain. His main research interest includes Data mining and Data Warehousing, Network Security, Biometrics, Software Engineering and Environments, Large-Scale Distributed Systems and Networking.