IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

XML Based Distributed Web Mining

Prof. Mamta gehlot[1], Siddhesh Deorukhkar[2], Rishit Kotian[3], Ajay Kemble[4], Anuj Rana[5]

Lecturer, Information Technology, Atharva College of Engineering, Mumbai, India[1]

Student, Information Technology, Atharva College of Engineering, Mumbai, India[2, 3, 4, 5]

Abstract: With the rapid growth of the Web data, for the phenomenon of data explosion but poor knowledge, as the data mining based on database table has been unable to meet the needs of Web application, a distributed Web mining is becoming a research hot spot presently. Firstly, based on the research of web mining, XML is used to transform semi-structured data to well-structured data, and a distributed web mining model based on XML is deeply discussed. Then, a mobile agent technology is introduced into the design of Web mining system, and a distributed Web mining structure based on data mining agent is built. Finally, an architecture design of network teaching platform based on XML and agents is realized, which can provide intelligent aids for personalized learning, thus greatly improve the teaching efficiency of network teaching platform. 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 introduced the process and the general methods 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 and a model of web mining system which has basic data mining function and faces multi-data on the Web is built. With the development of information technologies, web data mining has been put forward and in wide research. It is defined as the discovery, extraction and analysis of useful and potential information from the World Wide Web. But much of inhomogeneous and anomalistic and dynamic updated semi-structured data in web pages makes web data mining difficult. To solve this problem, on the basis of analyzing the characteristics of XML, the paper presents a web data mining model on XML, introduces the method to implement the model with XML.

Keywords: XML Database, ASP.net, SQL server.

I. INTRODUCTION

.Web mining is used to automatically discover and extract information from web related data sources such as documents, log, services and user profiles. Although standard data mining methods may be applied for mining on the web, many specific algorithms need to be developed and applied for various purposes of web based information processing in multiple web resources, effectively and efficiently. At present many websites are built with HTML, which is difficult to achieve real effective and accurate web mining with the development of computer and network technology data mining based on database table has been unable to meet the practical applications. The data resources often exist in the geographical distribution database; they can be structured, semi-structured and unstructured.

Modem decisions require information from various aspects. The face of "Information Ocean" on the internet, we need to extract useful knowledge which can be guide to decision- making. Because XML (extensible Markup Language) can make it easier together for structured data from different sources, which makes it possible to search for diversification of incompatible databases so as to brought new opportunities for Web data mining[l]. Through the Application Research on WEB Data Mining for XML and agent, the distributed data mining based on XML is proposed. Therefore, at effectively enhancing intelligent education service, Network Teaching based on XML and Data Mining Agent (DMAgent) has been seen as a solution to solving the bottlenecks of existing Network Teaching Platform.

II. LITERATURE SURVEY

In "Research of Web mining Technology based on XML" by Lilanand Rong Qiao-mei described 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 introduced the process and the general methods 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. According to the Web targeting a different group of mining, Web data mining is divided into: Web content mining, 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

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

on the hyperlink topology, Web mining structure can be classified pages, summed up the page.

III.PROPOSED SYSTEM

The proposed system is XML based distributed web mining. In this system, User searches query regarding the medical management. XML database will provide with the appropriate result which has been stored after the Data mining. Fig 3.1 is the Block Diagram of XML based distributed web mining.

Description of XML based distributed web mining system: Admin interface:

- HTML web page: Various web pages of medical management.
- Data cleansing and Data Extraction: Process of pruning web pages and extracting features. Parsing web page using XML.
- XML Database: Storage of parsed XML web pages.

User Interface:

- User: Client/patient who needs to retrieve information.
- Query box: User can type his/her queries.
- Output: Displaying the final result (web page).

Implement it for fabricating other feedbacks. However, some users can give already summarized feedbacks that can directly be included in the knowledge data base.
Actually, before sending the users feedback and appreciation about the product to the trust reputation system, system have to verify the concordance between them in order to avoid and eliminate contradiction or malicious programs attacking the system. In the redirected interface, system will display several feedbacks from different types. However, the user can specify the number of feedbacks to be liked or disliked. Of course, the system can also specify the minimum and the maximum number of feedbacks to be displayed by the user.
By redirecting the user to prefabricated feedback page the system tries to detect and analyse the user intention behind his intervention on the E-Commerce application. The system examines and evaluates the user intention using prefabricated feedbacks of different types. The trust score 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 (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].
A DMAgent model based on XML proposes a web mining data structure of adopting the XML technology in the

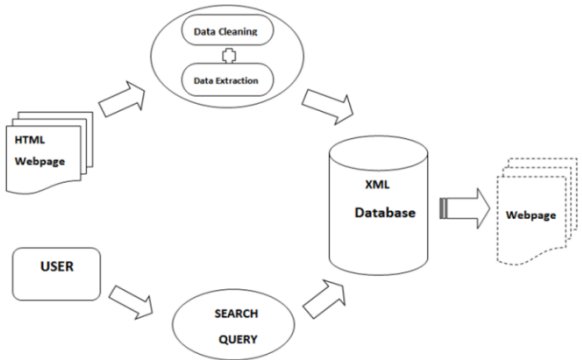

Fig.1. Block Diagram of Proposed System

application service layer to sample users' access data, and analyzes its advantages. The model overcomes a series of problems previously encountered in data preprocessing, high accuracy, ease of use in mining algorithms, and high application value. We can take advantage of XML technologies for data preprocessing, establish a semi-structured data model, extract meta-data on behalf of its characteristics, and save as a structured database. Mining-driven engine is a method of selecting experts. According to requirements for mining method selection policy, we can choose the most appropriate and efficient algorithm for combination of several algorithms for sequence in order to perform mining tasks.

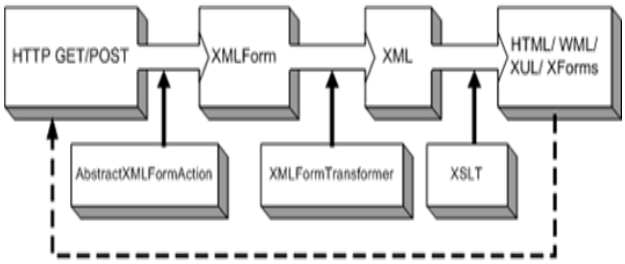

Fig.2. DFD Level 1 Diagram of Proposed System

The technologies and algorithms to be used in achieving this goal are explained in the next section.

IV. METHODLOGY

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {onion,potatoes} => {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

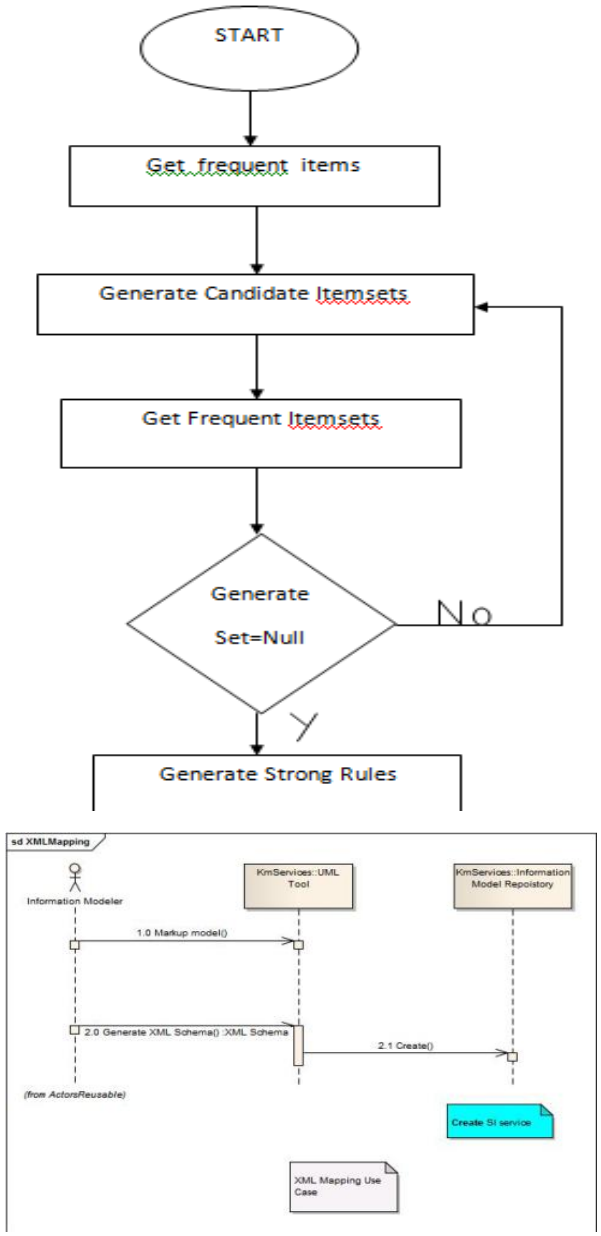

Fig.3. Flow Chart of Trust Reputation System

mining, intrusion detection and bioinformatics. In computer science and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps. Association rule generation is usually split up into two separate steps: 1. First, minimum support is applied to find all frequent itemsets in a database. 2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules. While the second step is straight forward, the first step needs more attention. Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations). The set of possible itemsets is the power set over I and has size $2^n - 1$ (excluding the empty set which is not a valid itemset). Although the size of the powerset grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all its supersets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent itemsets.

V. CONCLUSION

We are proposing XML based distributed web mining. In which XML gives us the flexibility to create customizable user interface architecture. It allows us to create structured content that we can manipulate in different educational context. XML help to normalize the network information, so that developers and computers can easily recognize the web information and create the open data that is not independent on platforms, languages or limited in formats. XML's flexibility and scalability is to allow XML to describe different types of application in the data, which describes the web page to collect the data records.

ACKNOWLEDGMENT

It gives us great pleasure in presenting this project report titled: "**XML based distributed web mining**". On this momentous occasion, we wish to express our immense gratitude to the range of people who provided invaluable support in the completion of this project. Their guidance and encouragement has helped in making this project a great success. We express our gratitude to our project guide **Prof. Mamta Gehlot,** who provided us with all the guidance and encouragement and making the lab available to us at any time. We also would like to deeply express our sincere gratitude to Project coordinators. We are eager and glad to express our gratitude to the Head of the Information Technology Dept**. Prof. Neelima Pathak**, for her approval of this project. We are also thankful to her for providing us the needed assistance, detailed suggestions and also encouragement to do the project. We would like to deeply express our sincere gratitude to our respected principal **Prof. Dr. Shrikant Kallurkarand** the management of Atharva College of Engineering for providing such an ideal atmosphere to build up this project with well-equipped library with all the utmost necessary reference materials and up to date IT Laboratories

REFERENCES

[1] S. Mukthyarazam and M.Kiran Kumar, "Web data mining Using XML and Agent Framework", IJCSNS, VoLlO No.5, May 2010.
[2] lieShen and Xueguirong, "Web data mining model based on XML", Systems Engineering Theory& Practice, Vo1.9, 2002, pp:75-77.
[3] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, VoU3, 2007,
[4] Tuncay Sevindik and Necmi Demirkeser, "Digital mining applications in Web-based education Environments", Scientific Research and Essays Vol. 5 (21), 2010, pp.3213-322I.
[5] Zhou Xiaomei and WangQianping,Su Lin, "Design of web mining model based on XML", Computer Engineering and Design, Vo1.28
[6] Liu Tangjian and Wu Xiaoning, "A J2EE-Based Network Education Platform", In Proceedings of the 2th International Workshop on Education Technology and Computer Science, 2010, pp: 659 - 662.