INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

AN APPROACH IN WEB MINING: WEB-BASED LEARNING ENVIRONMENT

VIJAY KHANDAR[1], PROF. P. D. DESHMUKH[2]

1. M.E Student Department of Computer Science and Engineering, P. R. Patil College of Engineering, Amravati.
2. Asst. Professor Department of Computer Science and Engineering, P. R. Patil College of Engineering, Amravati.

Accepted Date:

27/02/2013

Publish Date:

01/04/2013

Keywords

Web Mining,

Web Content Mining,

Web Structure Mining,

Web Usage Mining,

Teaching Resource.

Corresponding Author

Mr. Vijay Khandar

Abstract

This ongoing research focuses on how data mining techniques, if incorporated into web learning environments, can enhance the overall qualities of learning. Web mining techniques, including clustering and association rules mining, could be applied to extract hidden and interesting knowledge to facilitate instructional planning and student diagnosis. In this paper mainly discuss how to make use of web mining technology to improve distance education platforms. It will introduce web mining and its application in distance education platforms and propose a model of web mining process.

Introduction

With the rapid development of computer network and multimedia technology, modern distance education has become a new kind of teaching model, it breaks the multiple limitations in terms of personnel, time and space in traditional education mode. What's more, it provides an abundance of open teaching resources, so that people can teach and learn anytime and anywhere. Through the distance education platform, the interaction between students and teachers can be realized, such as video courses, answering, questioning, discussing, testing, homework submission and other teaching activities. However, at present, there are also many shortages in distance education platforms: 1) The use of network teaching resources is insufficient. That is the system can't provide the necessary teaching resources to students efficiently and accurately, and can't serve students' personalize learning. 2) Education platforms mode are very single, and low-smart. These models are system self-centred, rather than student-centred, the system cannot provide different learning resources according to different students' learning, and realize students' personalize learning [1]. Web mining in education is not new. It has been applied to mine aggregate paths for learners engaged in a distance education environment to recommend relevant words to students based on text mining from their browsed documents to recommend e-articles for students based on key-word-driven text mining, and to analyze learners' learning behaviours. The research proposed here will go beyond *usage* mining to consider the content of the pages that have been visited. In an e-learning system, both learners' browsing behaviours and course content are important to derive learners' learning levels, intentions, goals, interests, or abilities. Incorporating course content can aid in an understanding of learners' browsing habits. In particular, understanding the learners' browsing behaviours can facilitate, say, the personalization of course contents delivered.

1. Technology in Web Mining

Web mining is a comprehensive technology, related to web, data mining, computer linguistics, information theory 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, and the resources, to find the knowledge which is effective, novel, potentially valuable, and ultimately understandable, including the non-trivial process of concepts, patterns, rule, regularities, constraints and visualizations and so on [2].

2.1. The Classification of Web Mining

At present, there are three kinds of main researched web mining technology; they are web content mining, web usage mining and web structure mining. Its detailed structure is illustrated as Fig.1.Web content mining refers to the process of mining from the content of web pages or its description, and extracting the knowledge. There are two kinds of web content mining according to the objects of mining: 1) Text documents mining, including the text format, HTML format and so on. 2) Multimedia documents mining, including image, audio, video and other media types. Web text mining can associate analyze, conclude, classify, cluster the content of a large number of documents on the web, and make use of web documents to forecast the trend, etc. Multimedia documents mining on web mainly uses multimedia tools for the extraction of features, and then associate analyze and classify these features. Web structure mining refers to derive knowledge from the organizational structure of WWW and the relationship of links. As a result of the interconnection of the documents, WWW can provide the useful information besides the content of documents. Making use of this information, it's 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, and so on. Web usage mining 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 kind 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 ServerLogData.

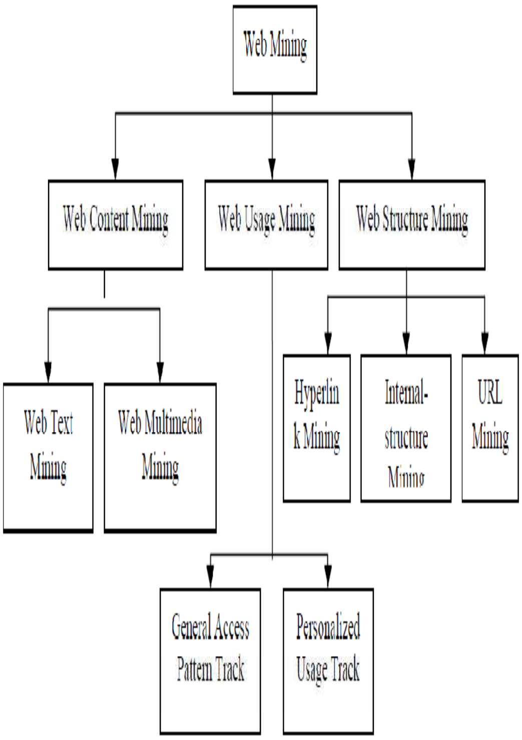

Figure 1 – classification on Web Mining

1.2. The Realization of Web Mining

Web mining is a branch of data mining. 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 model mainly based on machine learning, the methods used include neural network, natural law calculation method, etc.

- The Technology of Web Content Mining [4] 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.

➢ Text conclusion. 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. For example, when a user want to search the content about some knowledge through the distance education platform, the result of the search can be the interception of the first few lines of the description about the knowledge, through the first few lines description about the knowledge, users can get the view to the whole content, so that it is more convenient for users to search the resources they need. And now many search engines also use this method. Current related technology:

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 classification. It 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.

➢ Text clustering. 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.

➢ Associate rules. 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.

- The Technology of Web Usage Mining 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) pre-treatment, 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 behaviour can be discovered through the collection and analysis of users' click events.

2.3. The Process of Web Mining

Compared with data mining based on relation database or data warehouse web data mining is much complex. 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.[5] For example, in the distance education platform we can regard the resources of each course as an independent web site, and integrate the independent course into a comprehensive resource through web mining. Typical web mining process is as follows:

1) 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.

2) Selection and pre-treatment 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.

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

4) 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. The Application of Web Mining in Distance Education Platform

3.1. Web Content Mining in Distance Education Platform

The teaching resources in traditional distance education platform mainly include courseware, course plans, exercises, multimedia materials, courses, papers and other fixed contents. These contents have been categorized on the server, when the users want to search the relevant information, because of the dispersion of resources, it is necessary to choose and comprehensively analyze resources of all parts, and this process will use up a lot of server resources and consume a large amount of time. At the same time because of the differences between the resource developers and users such as knowledge structure, expression manner, understanding ability and so on, the two kinds people may understand the same statement of resources much differently, that cause the problem that users can't quickly find the information they needed in the mass of educational resources. In addition, users' different cultural backgrounds, language barriers, thinking manners, the capacity of knowledge acceptance and the age will also cause the same problem. It is known to all that distance education platforms tend to have these resources such as BBS, test library, exercises submitted by students or users. While in traditional platforms always treat these functions as the interactive way between students and teachers, actually in these modules there are a large number of targeted knowledge. For example, when a student can't understand a definition, he/she may write it down on the BBS for help, and someone will answer it sooner or later. In fact, this is a very specific knowledge, the educational resources on the web will be increased, and significantly expanse the entire knowledge base. At the same time, the analysis of the test, the correct views to students' homework, etc. all of these can be treated as a kind of knowledge which haven't been treated as knowledge in present platforms. Web content mining can mine information from web pages and their descriptions, and web content mining mainly bases on texts information. The function of web content mining is that it can associate analysis, conclude, classify, cluster the content of a large number of documents on the web, and make use of web documents to forecast the trend, etc. So it can solve the problems mentioned above. [4]

3.2. Web structure mining in distance education platform

There will always be some relationships between different knowledge, while little relationships have been identified in current model, even if have been identified by complex methods, it is inevitable that there will be omissions. What's more, the relationships between knowledge points will become more complex when knowledge increases, so that it is difficult to maintain. The platform can classify and tide up the knowledge automatically by web content mining, then form a more structured knowledge base by web structure mining, and the system will automatically constitute all of the association between knowledge, and don't need to consider the link between knowledge when add new knowledge points. So that it not only makes the large amount of maintenance work be reduced, but also make the entire knowledge base be integrated into a whole.

3.3. Web usage mining in distance education platform

Another prominent issue of current distance education model is that the system always regards resources as the core of education, it shows the same resources to different students to achieve the aim of teaching, and ignores the differences including interests, learning styles, learning needs as well as learning ability between different students. Making use of web usage mining, in one hand can be used as a reference of knowledge association, so that the link between knowledge points will be more closely and reasonable. Meanwhile through the analysis of access frequency and access time of different knowledge that the students have been accessed, it's get a more intuitive understanding about the knowledge students are interested in, as a result of this, it's recommend their interested knowledge to them for learning. On the other hand, when a student study a course, and spends much time on learning the knowledge is interested in and ignores the knowledge is not interested in but very important to learn this course well, the system can remind the student properly, and recommend the missing knowledge to the student, to some extent, it realizes the students personalized learning. While distance education platform in general only recommend the knowledge students interested in.

4. The model of web mining process in distance education platform

Web mining in distance education platform can be divided into two main parts, the first is mining from the resources (include fixed resources and dynamic resources) in the distance education platform; the second part is mining from the web logs.

(1) Web mining based on resources

- Resource search module Collecting data from course wares, course plans, exercises, materials, teaching videos, papers and other teaching resources in the distance education platform to form a original resource base.
- Information selection and pretreatment module Removing the some repetitive resources, redundant tag format, automatic identify paragraphs or field from the original resource base to form a classified resource base.
- Pattern discovery and analysis module Finding the relationships between all kinds of knowledge in the classified resource based on pattern discovery and analysis.

5. Data Clustering for Web Learning

Among mining techniques of particular interest in webbased learning environments is data clustering. It can, for example: Promote roup-based collaborative learning Traditionally, learning systems focus more on how to individualize course contents and delivery. However, in web-based learning environments where both the number of students and the size of the information can be huge, to reduce the cost and the computational burdens on the system, group-based learning will also be useful. Data clustering is a powerful tool to find clusters of students with similar learning characteristics based on their path traversal patterns and the content of each page they have visited. The clusters of students can be used to promote effective group learning, e.g., assigning students from different clusters so as to form effective learning groups for collaboration. In addition, after we find a cluster of learners with similar browsing paths, we could extract course contents along the paths to create fragmented contents (group-based course content delivery). These fragmented course contents can also be selected for recommendation, and the clustered paths can be used to sequence the curriculum for other students in the future (group-based instructional planning). Provide incremental learner diagnosis Incremental clustering can be performed to help diagnose learners as they browse through the system.

6. Conclusion

From the view of data analysis, it can mine, extract and analyze the massive data in the distance education platform by web mining technology, and also able to identify and extract the implicit information. To some extent, students and teachers can use this information for their own learning strategies and teaching strategies adjustments. From the perspective of curriculum design, using web text content mining to organize the teaching resources in the platform. For example, using the classification technology to organize the resources and form different levels; using text clustering to achieve reasonable classification of text content, in order to facilitate the text browse and retrieval; at the same time, through web usage mining we can grasp users interests which may contribute to personalized teaching and attract more users. From the perspective of the function of information retrieval, web mining technology is a key of the development of network information retrieval, it can improve the search results in the distance education platform.

References

1. Yang Dingzhong, "The Personalized Research of Modern Distance Education Based on Web Mining", Journal of Yangtze University (Nat Sci Edit), Vol. 5, No. 3, 2008, 205.

2. Wang Qijun and Shen Ruimin, "Studies on Web Mining Based Intelligent and Personalized Distance-learning Environment", Computer Engineering, Vol. 26, No. 12, 2000, 158.

3. 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.

4. Tu Chengsheng, Lu Mingyu and Lu Yuchang, "Research on Web Content Mining", 2003, 6-8.

5. Li Jian, Xu Chao and Tan Shoubiao, "Design and Research of a Web Data Mining Sytem", Computer Technology and Development, Vol. 19, No. 2, 2009, 70-72.

6. Agrawal, R., and Srikant, R. 1995. Mining Sequential Patterns. In Proc. of the Eleventh International Conference on Data Engineering (ICDE), 3-14, Taiwan.

7. Chen, M.S.; Park, J.S.; and Yu, P.S. 1998. Efficient Data Mining for Path Traversal Patterns. IEEE Trans. Knowledge and Data Engineering 10(2): 209-221.

8. Cooley**,** R. 2000. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD diss., Dept. of Computer Science, University of Minnesota.

9. Gaul, W., and Schmidt-Thieme, L. 2000. Mining Web Navigation Path Fragments. In Proc. of 2000 Workshop on Web Mining for E-Commerce—Challengers and Opportunities, Boston.