ISSN : 0974 - 7435 Volume 10 Issue 17

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(17), 2014 [9688-9693]

Research of web information retrieval based on data mining technology

Kun Qian
Jilin Engineering Vocational College, 136000, (CHINA)
E-mail: 8188521@qq.com

ABSTRACT

Internet makes it easier and faster to access to information, but because of the multitude of information onlineand continued rapid growth which the vast majority of users are independent, so how to find the information that they need in the network has become an important issue. Search engines are to some extent help people solve this problem, but it does not fully meet the needs of users. Based on this context,this paperdiscussed the Web data mining techniques and summarized its presentation, analyzed the current situation of Web information retrieval. For some shortcomings of Web information retrieval, this paper made a number of perspectives. This paper also introduces the data mining technology research which is applied to Web information retrieval and personalized search of online teaching resource library and improved the efficiency and quality of Web information retrieval.

KEYWORDS

Web information; Data mining; Retrieval technology.

INTRODUCTION

Web search engines partially solve the resource discovery on the problem, but it tends to return to the user to thousands of pages retrieved, of which a large part of the retrieval request with the user regardless of the user can quickly and accurately obtain the valuable information required. Furthermore, the purpose of the search engines is to find resources on the Web, on knowledge discovery on the Web, even if retrieval precision higher, the search engine is not capable. Faced with the challenge of data mining technology, this technology shows its strong vitality[1].

The so-called data mining is to extract from a large number of incomplete fuzzy random noise of raw data implicit in them beforehand unknown but is potentially useful information and knowledge. Or is found useful knowledge from databases (KDD), and data analysis, data integration and decision support process. It is the source of the data as the formation of knowledge, data mining as mining from ore, sand, like in gold mining, from the raw data in the vast ocean, scouring out a little bit of information they need. Data mining is involved in multi-structured data, in order to deal with heterogeneous, unstructured or semi-structured data on the Web, Web data mining has become an important branch of data mining research. Web data mining originated in data mining, aims to handle unstructured data, while they can use their findings to improve the accuracy and efficiency of information retrieval, discovery potentially useful model or data information in the WWW, making Web information Retrieving developed to a new stage.

THEOVERVIEW OF WEB DATA MINING TECHNOLOGY

Web data mining is a comprehensive technology, is to extract information (or knowledge) on the WWW resources from the process, is to extract the potential value of Web resources implication unknown patterns[2]. Repeated use of a variety of data mining algorithms to determine the mode or reasonable models from observational data, but also data mining techniques and theories applied to WWW resources for an emerging field of study mining. According to tap different objects can be Web-based data mining is divided into three categories: Mining Web-based content-based mining WEB structure, based on mining WEB use.

The so-called content-based Web mining is actually acquiring knowledge from the Web documents and their descriptions, Web-based document file mining and resource index or search concept Agent technology should also be attributed to this category. Web information resources of many types, the current WWW information resources have become the main network of information resources. However, in addition to a large number of people outside could crawl directly from the Internet, indexing, resource tracking, a considerable part of the information is hidden data (such as the results from the user's question and dynamically generated data exists in the database system, or some private data) cannot be indexed, so they cannot provide for effective retrieval methods, forcing us to dig out the contents. If the information from the resource point of view forms, Web content is text, images, audio, video, metadata, and various forms of data consisting of, so we said mining Web-based content for the multimedia data is a excavation.

Most studies are based on the vocabulary bag or said vector representation, but the drawbacks of vocabulary bag method is data rich free-text, vocabulary is very large, very difficult to deal with them on the basis of law. In addition, a more meaningful approach is latent semantic indexing, which by analysis of different documents shared vocabulary same subject, find their common roots, instead of using the common root of all terms, in order to reduce the dimension of space. For example: informing, information, informer, informed can be expressed, thus reducing the size of a collection of properties with their roots inform.

Web structure mining objects are hyperlinked Web itself, namely the structure of Web documents excavation. For a given set of Web documents, should be able to find useful information on the link between their situation through the algorithm, hyperlinks between documents reflect contain between documents, references or affiliation, reference documentation for instructions often referenced documents more objective, more generally, more accurate[3].

To some extent Web structure mining has benefit from the social networks and research and analysis of reference. The relationship between pages divided in coming and going connection to connect, using citation analysis method to find a link between the interior and the same site between different sites. Most notably in the field of Web structure mining algorithm is H its algorithm and Page Rank algorithm. Their common denominator is calculated quality hyperlinks between Web pages using certain methods, resulting in a heavy weight pages. Clever and well-known Google search engine is on the use of such algorithms.

In addition, Web structure mining is another attempt to dig under the Web data warehouse environment, including connecting to measure Web Structure Mining complete Web site on the local inspection by the same server, in a different Web data warehouse to help locate copies of checks mirror sites, through the discovery level attributes for a particular field of super-connected to explore how it affects the flow of information Web site design.

Web usage mining, in the emerging field of electronic commerce is important, it is by tapping the relevant Web log records to find Web pages the user access patterns, by analyzing the logging rules, can identify the user's loyalty, preferences, satisfaction, you can find potential customers, enhance service competitiveness site. In addition to using the recorded data Web server's log records also included the proxy server logs, browser logs, registration information, user session information, transaction information between possible, Cookie information, user queries, mouse click stream and all other users and sites The interaction record. Visible Web data usage record is very great, and the data type is also quite rich. According to the data sources of different treatment methods, Web usage mining can be divided into two categories, one is the Web using the

recorded data conversion and transfer into a traditional relational table, then use the data mining algorithms to the data in relational tables routine mining; Another is the Web using the recorded data directly pretreatment further excavation. Web usage mining is an interesting problem is to use a proxy server environment with how to identify a user across multiple users, how to identify that belong to the user's session and use of records, this problem seems small, but in a very large extent, affect the mining quality, so it was special in this regard were studied. Generally speaking, the classical data mining algorithms can be directly used in Web usage mining up, but in order to improve the quality of mining, researchers at the extended algorithm efforts, including composite association rules algorithm, the improved sequence discovery algorithm[4].

THE SITUATION OF WEB INFORMATION RETRIEVAL

Web information retrieval, refers to find a given number of queries related to the appropriate subset of documents from a large collection of Web document. The web is the most important part of the Internet, but also the most important source for people to obtain network information, in order to facilitate people to find the information they need in a large number of complex web pages, these search tools developed rapidly. Generally believed that web-based information retrieval tools have major Web search engines and classified two kinds of networks. Web search engine spiders and other web pages through the automatic search software search page and then automatically give some or all of the characters on the page for the index, the formation of the target summary format files, and database network accessible for people to retrieve network information retrieval tool. Network directories and search engines is completely different, it will not all pages across the network for each site are into them, but by professionals to carefully choose the site's home page, put it in the appropriate category.

However, several existing network information retrieval model, while ordinary users to meet the needs for higher retrieve information needs of professional users, there are still a number of problems.

As the network information navigation only supports single-step information on positioning, navigation system only allows step by step tracking information[5], the client received a lot of redundant information in the tracking process, resulting in reduced efficiency. On the other hand, also because there is a deviation problems locating information: Because it is not yet customized information WWW, WWW cannot exclude a document on the user's part useless, causing customers to get a small part of the relevant information to get a lot of irrelevant information Since there is no unified standard document customization, so this bias is often not easy to control.

In the retrieval process, users often with great blindness, sometimes just luck, lack of clear objectives. Web-based search engine based retrieval with hypertext / hypermedia browser, users easily retrieved to produce a clear impression of the content. When the user retrieves a network address and retrieve the entrance as you can along the chain and the line step by step browse, but this process continues to be the subject of fresh searchers out to attract attention, the user may go astray, completely forgotten Retrieving the target, and finally nothing.

Robot automatically indexed by search engines indexing database software, but not high degree of intelligence Robot. When using Google, Baidu and other search engines to search, enter a search query resulting output is a lot of web addresses, users can only view one turn on filtering, which filter out content in line with their retrieval goals. In addition, a wide variety of search engines, search strategies vary, some search engines offer search interface coupled with no building tips and syntax to retrieve search results are not accurate, appropriate levels of detail in the summary information, all of which add to the burden on the user.

THE APPLICATION OF WEB DATA MINING TECHNOLOGY ON WEB INFORMATION RETRIEVAL

Knowledge Retrieval Web information retrieval in order to solve the current data exist, a lot of information but little knowledge of the problem of inefficient retrieval proposed a new kind of information retrieval concept. It is a comprehensive application of information science, artificial intelligence, cognitive science and linguistics and other disciplines of advanced theory and technology, based on knowledge and knowledge organization, the integration of knowledge processing and other multimedia information processing methods and technology, is a fully express advanced information retrieval methods and optimize the user requirements, efficient access to all types of knowledge sources of media (text, images, video, sound, etc.) and the selection of the user needs accurate results. Network data mining applied to Web-oriented knowledge retrieval, data mining is mainly reflected in the network to achieve efficient retrieval of knowledge acquisition, knowledge base and provide "knowledge model" for knowledge retrieval, which assisted retrieval system retrieves accurately grasp the needs of users, adjust and knowledge to optimize the system's own environment and problem-solving skills, accurate search results. By mining various types of information sources on the Web content and structure of users and experts using records of the excavation, to achieve the associated knowledge, clustering, classification, and the establishment of appropriate sources of information knowledge base, user knowledge model, expert knowledge model, real-time access to retrieve all kinds of knowledge and guidance to achieve retrieval process.

Mining Web information sourcesto form a knowledge base of information sources mainly for Web content mining and information source structure mining. Sources of information on the Web content mining, mainly the use of statistics (including natural language processing) method, machine learning, neural networks and other methods of information contents of the Web information sources for analysis, and found sources of information on the subject, subject distribution, structure and content on the context, to extract knowledge. Specifically, for each data type by mining such as text, sound,

images, graphics, multimedia and other features of web content, web content based feature extraction to achieve information sources, clustering, classification, identify potential data inside knowledge form information source knowledge base[6]. Source of information on the Web structure mining refers to the relationship between the Web page hyperlinks, document structure, URL address of the structure of the documentation of the excavation, in order to discover knowledge from the Web's organizational structure and link relationships. Because the Web is very much the type of source data, both text and non-text data (such as sound, graphics, images, multimedia information, etc.), both structured data, there are semi-structured and unstructured data so for different types of data, data mining methods are not identical. Text types of information sources, typically using statistical analysis methods for feature extraction, using k-nearest neighbor method, fuzzy pattern recognition method based on feature correlation method, based on the concept reasoning methods classification, clustering. For non-text mining complex information sources, it is usually rule-based reasoning technology concepts. Web structure mining is often used Proprietary algorithms, Page rank algorithm, Hits sorting algorithms and other web mining methods.

Information source content mining, can reveal features of interrelated themes of knowledge and information between knowledge networks, to achieve a reasonable classification of knowledge, the formation of an information source knowledge model (including classification knowledge and content knowledge network information objects); through mining information source structure can reveal the authority associated with the page, the authority of the link between these documents reveal the inherent structure information in a useful model, help from multiple dimensions and levels provide access channels[7].

Mining Web users access to records, the formation of user knowledge model. Mining user access records, that is, we often say that the network usage mining. Its essence is to use data mining techniques from cognitive psychology point of view of the user's awareness of the situation analysis of the Web environment, mainly through mining the user's access to records. Users access the original data recorded from the user registration information for each Web server and Cookies reservations, access to records, as well as information about user interaction with the system, the original data on which the information is not Web network, but to extract the user and network interactions out of secondary data. Mining user access records usually have two kinds of ways, one is to track user access patterns, one is tracking the user's record. Analysis of user access records mainly includes the following aspects: user context analysis, classification analysis of user groups, user preference analysis, analysis of user satisfaction, etc. retrieval. User access records mining method is usually used in mining statistical probability analysis, path analysis, classification, clustering, association rules analysis. Commonly used classification algorithms including decision trees, neural networks, k nearest neighbors and so on. Typical clustering algorithm commonly used cluster analysis, the main cluster, such as neural network method of self-organizing feature mapping method through distance clustering method based on random search, clustering feature tree method. Common association rules discovery algorithm Aprioriclassical algorithm and AprioriBest improved algorithm.

Web users type information widely, the background is complex, personalized information needs strong, so its mining and analysis using records with the aim of obtaining information about the user's law, the formation of user knowledge model. Because users to save and manage knowledge model with user knowledge, new knowledge is added to the user model, support and respond to the needs of retrieval systems function, and therefore, access to mining and analysis of information helps the user to accurately understand and grasp the user's request, reduce thinking man-machine differences lead to bias retrieval of user intent to understand. If the system is difficult to understand the real-time requests of users, you can call the user model, auxiliary judge[8].

THE APPLICATION OF WEB DATA MINING TECHNOLOGY IN PERSONALIZED INFORMATION RETRIEVAL OF ONLINE TEACHING RESOURCE LIBRARY

Differences in the status quo of education, economic development, cultural background and other factors make the construction of teaching resources at home and abroad there is a big difference. Our information service, particularly in the field of higher education, still shows the supply situation. In today's rapidly growing popularity of the network, concentrate manpower, material and use of the network at any of this, a wide range of information delivery tool, to carry out large-scale construction of teaching resources to take full advantage of the network in teaching and learning, is the field of education essential a daunting task.

(1) Network teaching resources repeat now many schools online teaching system is jointly sponsoring the school and network andthe network company out of capital and technology, school of brand intangibles school, teaching force and teaching management. This school is for profit is essentially a kind of market behavior, in this case, cooperation between the various schools is more difficult, competition is inevitable, so it is easy to cause repeated on each school teaching resources construction. Many schools in the conduct of teaching[9], the teaching system support platform, the learning management system to each course are all re-development. In fact, some network currently teaching universities are similar in the teaching content and teaching methods, such as mathematics, computer literacy foundation, university English courses, so there is no need to go to each school to develop their own teaching support platform and management system, on the same network programs spend multiples of manpower and financial resources. So for the construction of online distance education resources in the education department should be carried out under unified planning, in order to avoid duplication.

(2) Lack of good online teaching resources. For the current status of online learning, the country has built many network schools, but suffers from a lack of teaching resources online that can give full play to the role of network teaching. The lack of online teaching resources are mainly good online course material library and network less, the lack of good

network courseware, some schools are teaching content of the lesson plans written material in the form on the web page, so that a single rigid forms of teaching, students less attractive, teaching poor, minority attached to classroom instruction video images, but due to the current network transmission speed is slow, continuous video image is poor, there is little visible value. If you do not have good online teaching resources, network would be empty, cannot play its due role in the teaching process.

(3) Teaching resource library construction heavy material library, light resource integration tools. Although now many online teaching resource library, but most just a teaching material library. And its carrier to the disc-based, although some companies can also provide relevant material on the Internet, but because of restrictions on domestic network development, the scope of its impact is still relatively limited. And in many repositories, in addition to providing a resource search tool, but does not provide additional information integration tool, teachers receive information in the future, need to use other authoring tools, such as web authoring tools, multimedia presentation software to combine information many teachers feel that there is a certain degree of inconvenience and difficulty, thus affecting their use of the resource base, reducing the resource library to play the role of teaching in the.

(4) resource classification is not standardized, low resource utilization library. Different educational information resource library, a resource library is usually divided into lesson plans, test database, teaching software libraries, libraries and other multimedia material, apparently a large part of this classification is based on the resources that may arise in the teaching activities carried out, but this does not meet the data classification methods and standards, there will be a message to multiple definitions, or information contained excessive and overlapping each other circumstances, lead to relatively low utilization repository.

To solve the above problems, the development trend of online teaching resource library will be like this:

(1) The existing network teaching resource library construction norms and standards promotion, the trial will greatly facilitate the construction of the repository standardized, scientific and systematic. According to the same standards development and construction of the repository system due to follow the same definitions and criteria, it is possible to easily achieve data exchange and sharing of resources, effective problem solving resource library extension.

(2) XML technology will play a very important role in building networks among teaching resource library. Many of the organizations in the world to develop standards and norms purpose to achieve data exchange and sharing between libraries and other resources. XML is a data markup language, which is used to describe the data itself. Data and style separation of data is more flexible and free. Using XML files, need to exchange data between different databases were exported and imported, you can achieve data exchange platform without considering the issue, so XML can be used as a cross-platform data exchange standards in different databases, XML that is used as an intermediate layer virtual database[10]. Will occupy a very important position in future data transmission and exchange, the use of XML for data transmission and management is the trend.

(3) The future of online teaching resource library building will toward universal, professional, local, personalized with the characteristics of the direction. People repository-based learning is in the event of a significant change a resource-based and learning-based teaching resource. Therefore, we at the same repository construction, information technology education should be integrated into the repository using them to improve the user's IT capacity to ensure that educators can use a variety of tools that can be useful for a variety, easy to use multimedia teaching resource material for creative, personalized intelligent combination of multimedia teaching creative at the same time give the learner to provide a good learning environment.

(4) "Service" as the center of the network teaching resource library for construction, will be a major feature of the future of online teaching resource library. Now many educators who proposed the "online education is service" concept, Leave aside this sentence is correct or not, this has reflected the people for the "service" Call of consciousness in the current network education. Resource library building as a basis for online education, the same should also have this awareness.

(5)Intelligent Resource Library with intelligent analysis and intelligent search feature is one of the development trend of network resource library construction. There is now dependent on the repository is not very high, as people repository deepening understanding and reliance growing, the construction of the repository is bound to put forward higher requirements, and only some intelligent technology into to the repository under construction, in order to meet the growing needs of the user. Intelligent network teaching resource library will offer people a more efficient, more convenient, more personalized service.

CONCLUSION

Network technology has penetrated into every aspect of social life, with the rapid development of the Internet in the world, a growing number of databases and information systems continue to join the network, making the development of the world's Internet home most wide, the largest repository of information. WWW bring people to the new network in the world, but also the people on the network into a complex network maze. Faced with the colorful complex Web space, how to quickly and efficiently explore the vast network required information from the HTML document has become a major issue of concern, Web data mining technology, however, is an effective solution to this problem.

With the continuous development of WWWand the proliferation of WEB information, how quickly, efficiently and accurately retrieve network information has become increasingly important, the WEB growing need for the development of information retrieval through a variety of techniques to further promote. As data mining is an important research branch of

the WEB data mining, it has a higher than WEB information retrieval technology level andit has a very close relationship with WEB Information retrieval, so it has been an important reference for information retrieval on the WEB, so you can apply WEB data mining techniques to research in the field of WEB information retrieval to improve information retrieval intelligent processing capabilities and makeWEB information retrieval development to a new and high level.

ACKNOWLEDGMENTS

The authors wish to thank the National NaturalScience Foundation of China for contract 61100194, the foundation of education department of Zhejiang province of China for contract Y201120520 and the research foundation of Hangzhou Dianzi University for contract KYS105612008 and YB1205, under which the present work was possible.

REFERENCES

[1] Jiawei Han; *Computer Research and Development*, November, **1(4)**, 405-415 **(2001)**.
[2] Xiaohu Wang, Gang Liu; *Journal of Information*, June, **53(10)**, 3-6 **(2012)**.
[3] Xin Chen; *Computer Application*,February, **11(2)**, 33-36 **(2006)**.
[4] Zhengyan Liu; *Microcomputer Development*, April, **189(4)**, 163-169 **(2010)**.
[5] Jinqing Yao; *Modern information technology*,April, **27(1)**, 8-11 **(2012)**.
[6] Xiaolin Ma; *Journal of The software herald*, June, **12(1)**, 36-42 **(2011)**.
[7] Chun Zeng, ShuYueqing; *Computer Research and Development*, March, **12(1)**, 14-23 **(2000)**.
[8] Xunleiwang, Chunlai Li; *Computer Research and Development*, March, **5(3)**, 149-158 **(2000)**.
[9] Weifeng Zhang, Qinzhi Ma; *Modern information technology*, February, **2(3)**, 90-97 **(2009)**.
[10] Xiekang Lin, Baili Tong; *Journal of E-education Research*, February, **2(3)**, 148-150 **(2003)**.