Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 4, Issue, 07, pp.066-070, July, 2012

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

FEATURES AND CHALLENGES OF WEB MINING SYSTEMS IN EMERGING TECHNOLOGY

A. Pappu Rajan[1] and S.P. Victor[2]

[1]Department of Computer Science and Research Center, St.Xavier's College (Autonomous), Palayamkottai, Tirunelveli, Tamil Nadu, India
[2]Department of Computer Science & Director of Computer Science and Research Center , St.Xavier's College(Autonomous), Palayamkottai, Tirunelveli ,Tamil Nadu, India

ARTICLE INFO

Article History:
Received 16th April, 2012
Received in revised form
17th May, 2012
Accepted 28th June, 2012
Published online 30th July, 2012

Key words:
Data mining ,
Web mining,
Challenges of web mining,
Web Log mining.

ABSTRACT

Web mining is a new developing research discipline and also it is subdivision of Data Mining , it has attracted a great deal of attention in the Information Technology and in society as a whole in recent years, due to the wide range and availability of huge amount of heterogeneous data. The web has become versatile tool for almost all application today. Mine this available huge data to make it proper use and presentable, giving right solution to a particular problem is a big real challenge . In this paper deals with a introductory idea about the data mining, web mining , web log mining and challenges of mining the web data.

INTRODUCTION

In the late 1990s, low cost personal computers and an extensive, relatively easy to use Internet helped computers spread to the majority of households in may developed countries .Many of the activities for which people use the internet are long standing and well rooted in our social system. The web is a system of information distribution using the internet. The internet revolution is transforming to economics in the world wide. No business sector, no company, will be left untouched. Globally, Internet users have grown from 39 million in 1995 to 315 million in 2000. The number is projected to grow to 716 million users in 2005. There were an estimated 2,459,846,518 internet users world in the February 2012, it is representing about 30.2 % of the population worldwide, according to Internet world stats data updated in February 2012 [10] All are involved in a wide range of decision about technology, decision that are vital to the success of the organization. How is web technology changing organizations? Every business sectors dealing huge volume of data so that Web technology today is a vehicle for making substantial changes in organization, markets, and the economy. In multi-tier architecture often referred to as n-tier architecture is a type of Two Tier or client–server architecture in which there are three layers such as the presentation, the application, and the data management. Each layer logically separated by their unique features and processes. There are computer systems or programs that use different learning intelligence techniques to solve problems that ordinary require a knowledgeable human, whenever the person is taking decision that time data is required, so data should be extracted from the wide availability of huge amounts of data. Data can be extracted or mining knowledge from large amounts of data refers data mining. The overall ultimate goal of the data mining system is to extract knowledge from an existing data set and transform it into a suitable structure for future use.

Web Data Mining

Web data mining can be defined as the discovery and analysis of useful relevant information from the World Wide Web data. WWW having lot of useful or useless meta data, the web log data regarding the users who retrieved the multiple web pages and the web structured and unstructured data. Over the past Fifteen years, we have already faced lot of explosion type of information resources available over the web. Web mining can be applied all the field of artificial intelligence system, human interaction, cloud computing, neural data mining, geographical data mining , information retrieval and so on to the web data and traces user's visiting characteristics and then extract users' pattern are very important issues. Web applications, which can be focused to extraction of knowledge from the web, extraction of knowledge from the user's behavior, getting information from the web , providing information to the web, downloading and uploading data over the web. This paper aims this, pays close attention to the development trend and characteristics of the web mining, web log mining, and major challenges of web mining.

**Corresponding author:* ap_rajan2001@yahoo.com

II. Categories of web mining

Thus, the Web data mining should focus on these three issues: Web Structure Mining, Web Content mines and Web usage mining. All of the three categories focus on the process discovery for unknown and potentially very useful information from the web. Though each of them focuses on same attribute but each might be different mining objects of the web.

Web Structure Mining

Involves mining the web document's structures and links. In some insight is given on mining structural information on the web. Web structure mining is very useful in generating information such visible web documents , luminous web documents and luminous paths , a path common to most of the result returned , use linkage information to improve search engines ,hyperlink structure analysis, link analysis ,graph, categorization, mining the document structure .

Web Content Mining

Describes the automatic search of information resources available on-line. It represents structured, unstructured, semi structured documents and model to interactive retrieval view and DB View. All the above it is a Mining, extraction and integration of useful data, information and knowledge discovery from Web page contents. Web content mining examines the contents of web pages as well as results of web searching. Basically it can give two different major approach: Agent based approach, Data base approach. First approach is on improving the information finding and filtering which is incorporate with using intelligent search agents, information filtering or categorization, personalized web agents. Second approach is on modeling the data on the web into more structured form connection with multilevel data bases and web query systems.

Web Usage Mining

Focuses on several techniques that could learning or predict user behavior and navigation pattern because user using the web round the clock. It includes the data from server access logs, user registration or profile, user sessions or transactions, , etc. It also depends on the collaboration of the user to allow the access of the web log records.

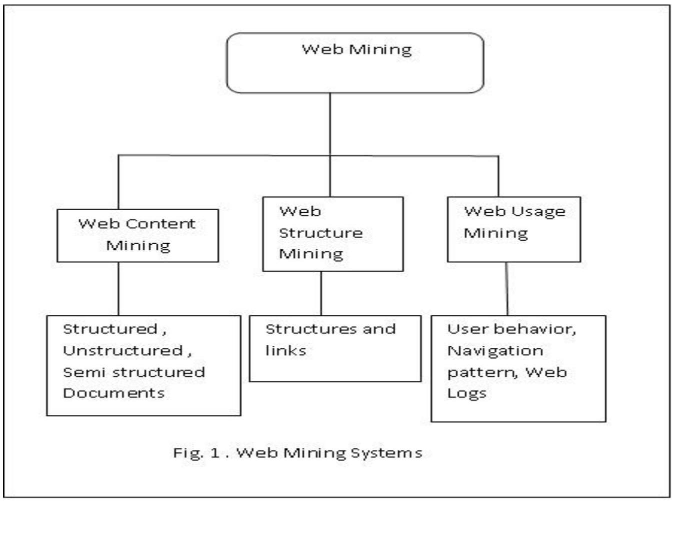
Web Mining
Web Content Mining
Web Structure Mining
Web Usage Mining
Structured , Unstructured , Semi structured Documents
Structures and links
User behavior, Navigation pattern, Web Logs

Fig. 1 . Web Mining Systems

The main objective of the web mining is to provide data mining algorithms which improve the content structure performance, categorization of web documents, snippets and user sessions.

Web Multimedia data bases

Older day's web system mainly deals large collection of documents whereas recently documents always linked or embedded with lot of collection of multimedia documents with different types of files, heterogeneous data such as images, videos, audios. Millions of web pages are added every day and million of others are modified or deleted. The main aim of web content mining is knowledge discovery. While retrieval of data is difficult so that we need to find different new algorithms used to improve the system and also recent days we are using multilevel data bases , multidimensional data bases and web query systems. Most of the popular algorithms are used only numeric, character, text data types. The use of multimedia data is complicated or not suitable for current system, all the above most of the proposed algorithms are not fulfilling the requirements for web technology. All present web pages are Dynamic nature so data base cannot be assumed to be static. Most of the previous Content management support algorithms more flexible to static, while we use dynamic it's require new suitable algorithm. The traditional way of retrieving images from data bases is to assign text annotations to image data. Presently we are using different types of data bases like Web site image database, web site text data base, web site video data base, web site image text data bases. On this finding or improving new filtering approaches is necessary to improve web multimedia meta data bases otherwise it leads poor interpretability of mining result.

Information Retrieval

IR involves retrieving desired information from textual data. The historical development of IR was based on effective use of libraries. Many universities and public libraries use IR systems to provide access to books, journals and other documents. To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things: 1) A document collection 2) A test suite of information needs, expressible as queries 3) A set of relevance judgments, standard a binary assessment of either relevant or non relevant for each query-document pair. [11]. Automated information retrieval systems are used to reduce what has been called "information overload". Web search engines are the most visible IR applications. In work of searching , retrieving data from web, we are naming several other words such as data retrieval, document retrieval, information retrieval, text retrieval .There are no of words overlapping but meaning and tasking almost similar. For the information retrieval to be efficient, the documents are typically transformed into a suitable representation. There are several representations like theoretical, probabilistic, future based retrieval model are available. As a result, some traditional data mining methods are not applicable to web mining data retrieval. The key problem of information retrieval in web mining is how to improve comprehensive and correlation of information accessed from the web data base and to make efficient information retrieval for classification of different web page and to retrieving relevant. The challenges for web structure mining are to deal with the structure of the hyperlinks within the web itself. Link analysis is an old area of research. However with the growing interest in web mining, the research of the structure analysis increased and these efforts had resulted in a new emerging research are link mining accurate information without spreading spam, irrelevant or unwanted web pages.

Search Engine Result Pages

A web search engine is designed to search for information on the World Wide Web. Performance of search engines is following steps: Crawling, Indexing, Searching. A Web crawler is one type of software agent. The functionality of web crawler is starting to find list of URLs to visit called the seeds. After visits these URLs, the crawl frontier using different set of polices to finds all the hyperlinks in the page and listing the list of URLs to visit. The large volume of load implies that these crawler can only download limited number of the Web pages within a given time, Though we already employing various number of page ranking algorithms so it needs to prioritize its downloads. The high rate of change implies that selection, re visit and parallelization policies wherever the pages might have already been updated or even deleted. Search engines, although they are continually getting better, tend to be rather low tech.

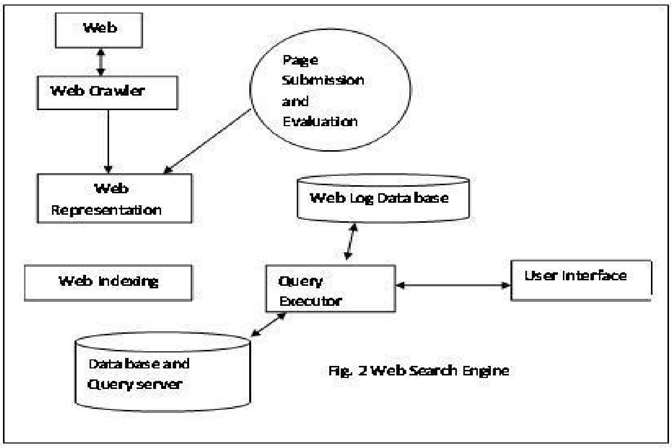

Fig. 2 Web Search Engine

The Web indexing includes back-of-book-style indexes will fallow to individual websites and the creation of keyword metadata to provide a more useful vocabulary for Internet or onsite search engines. How web index is performing indexing? As we aware of this the fundamental concept of indexing is a sequence or order of items. In general, arrangement of the indexes are usually forming either alphabetically, numeric or chronological. Indexing is not only just like ordering or sequencing of listing of items but it is a well defined structured , and also it can have lot relationship with other close related objects, so it can be lead users to more exact or related topic that might meet their information needs more closely. Metadata is data about data, it leads Metadata management involves storing structured information about other information. Metadata web indexing involves assigning keywords or phrases to web pages or web sites within a meta-tag field, so that the web page or web site can be retrieved with a search engine. Web indexing involves assigning keywords or phrases to web pages or web sites within a meta-tag field, so that the web page or web site can be retrieved with a search engine that is customized to search the keywords field. This may or may not involve using keywords restricted to a controlled vocabulary list. Popular Google search engine is also followed inverted file structure for indexing. A robots.txt is a file placed on our server to tell the various search engine spiders not to crawl or index certain sections or pages of our site. One can use it to prevent indexing totally, prevent certain areas of our site from being indexes or to issue individual indexing instructions to specific search engines. Sometimes search engine indexing problems can be caused by Robots.txt file errors. Robots.txt is a small text file which is uploaded to the root directory of a web server to tell search engine robots which web pages and website assets such as folders, images should be excluded from search engine indexation. A simple syntax error in the Robots.txt file could totally prevent Googlebot (Google's search spider or crawler) from indexing your website [13].

In general, there are no of common factors behind search indexing: merge factors, storage techniques, index size, lookup speed, maintenance, fault tolerance. The major key problem of the web indexing in the web mining is how to improve or construct a new system in the following areas : Word Boundary Ambiguity, Language Ambiguity ,Diverse File Formats , Faulty Storage ,Tokenization, Language recognition. If we find new technique that will be meet many other problems in the web search engine indexing. Web searching is a service for finding information on the World Wide Web. The Dictionary says meaning of the Searching is examining carefully or thoroughly. So searching is not at all a ordinary task, which is highly complicated task in the view of fulfilling the user requirements. Web search is very different from a normal information retrieval because of the following factors: bulk, diversity, growth, dynamic, demanding users, duplication of data, hyperlinks, indexed pages, complex queries. We need to find the solution to the above issues. The major research is meeting to give the solution on the above issue. Bulk of data is very important factor, the web is very large. Bulk import large objects are major problem of web searching. Many proposed approach relies on the observation that a web searching software sometime fails for certain values when the system deals bulk of data. Diversity means many things to many people. Diversity of the web, it consists of text, images, movies, audio , power point presentation , pdf files and all other multimedia data. The people can be from anywhere in the world with different domain knowledge , the web pages or files built using many latest software, and store on a variety of high end server. Each and every day million of million web pages developed by people, so that growth of web page and web application architecture are having large scope it should not be the end of the develop.

We are coming to the important issues of web searching , which is irrelevancy of retrieving of unwanted information . In most search engine , some read meta tags, some read only the first few hundred words of text. Some read both and link them together, so that if we have meta tags which don't relate to the words on the page, it won't rank our page very highly. The meta tag element used in search engine optimization can be improve the search engine which are popularized by keyword , description, language , robots attributes. However, when no other ranking signal is present, unique words that only appear in the meta keyword tag section of documents can still be used to recall these documents.[14]. The meta tags problem should be solved to improve the search engine optimization. Search engine optimization is the process of improving the visibility of a website or a web page in a search engine's and search results. SEO may target different kinds of search, including image search, local search, video search, academic search, news search and industry-specific vertical search engines. On a global scale India made the maximum number of requests to google through executive and police agencies for removal of content from the company's online services but achieved a low rate of compliance from July to December 2011 . By contrast there were five court orders from removal of content, four of them on grounds of defamation, with a higher compliance rate by the company than in the case of executive orders [THE HINDU, ISSN 0971-751X, Vol.135 No.146]. We need to maintain Laws surrounding these issues vary by country, and the requests reflect the legal context of a given jurisdiction. Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide. These biases could be a direct result of economic and commercial processes, and political processes. So the popularity of social networking we wouldn't be avoid this types issues but web content hosting must be regularize without business motive, the complete transparency also needed in this context.

One of the advertising strategic, followed by the financial basis of the search industry is Pay-Per-Click targeted advertising. Some observers think that click fraud is approaching a tipping point that poses an existential threat to the entire industry. We are now getting spam messages submitted to our comments web forms mainly by people in India and other off-shore locations. Each message had to be individually, manually, submitted because of our image interpretation requirements. That this much labor can be expended to send one spam message to one person is bad news for the search industry. Similar efforts expended in click fraud would be much more lucrative. The solutions to the click fraud problem are inevitably going to require progressively more invasive tracking of individual web users, which in turn represents an ever increasing privacy issue, Pages indexed ,Daily direct queries , personalized result , tracking, Privacy Sharing these are major resulting challenges in search Engines.

Web Log Mining

Web analytics is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage. Web log analysis software is also called a web log analyzer is a simple kind of Web analytics software that parses a log file from a web server, and based on the values contained in the log file, derives indicators about who, when, and how a web server is visited. Usually reports are generated from the log files immediately, but the log files can alternatively be parsed to a database and reports generated on demand. There are two methods we are having for Web analytics, first one is page tagging, and second one is web log file analysis. In the early 1990s, web site statistics consisted primarily of counting the number of client requests or hits made to the web server. The first commercial Log Analyzer was released by IPRO in 1994. [16]. Two units of measure were introduced in the mid 1990s to gauge more accurately the amount of human activity on web servers. These were page views and visits. A page view was defined as a request made to the web server for a page, as opposed to a graphic, while a visit was defined as a sequence of requests from a uniquely identified client that expired after a certain amount of inactivity, usually 30 minutes. The page views and visits are still commonly displayed metrics, but are now considered rather rudimentary [17]. Web log file it consists of two files, first one is server log, second one is client log. A server log is a log file automatically created and maintained by a server of activity performed by it. Always web server log file maintains a history of page requests. Web usage mining is used to find out the interrelated information from web log file which is involved all of users' browsing behavior is completely recorded in the web log file. Which contains users' name, IP address, date, and request time etc., This file also keeping the Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent. These are types of server log file: Access Log, Agent Log, Error Log and Referrer Log. Ordinary internet users can not accessible these files, only web master or administrator can have the ability to access these files. A statistical analysis of the server log may be used to examine traffic patterns. Which may be noted the following information: time of the day, day of the week, referrer, and user agent.

Web log file is saved as text (.txt) file Web usage mining technique avoiding irrelevant information usually cannot support directly use of original log file. Therefore the preprocessing of web log file is complicated one. If web site want to maintain effectively proper analysis is required in the case of web log maintaining. Here also we are having number of technique available for preprocessing of web usage mining, but that will not meet improvement of preprocessing of web usage mining therefore we need to improve the quality technique for improving web log preprocessing in web usage mining. Server log files are in keeping much relationship of web pages. Any one web site can be visited by many user as well as users' behavior also recorded. In general Web pages contain lots of files such as images, voice, animation and advertisements. Whenever user behavior identified based on click analysis or usage, only minimum no of log records related to the web page. Such analysis are useless because without filtered there is no use for keeping the log records related to the requests for image files. In the packet sniffers, data can be transferred using any type stateless or statueful protocol. When stateless protocol is used between a server and the client, the server does not remember anything. It treats any message from a client as the client's first message and responds with the same effects every time. Stateful protocol means the server remembers what a client has done before. The data being detected includes much more information than Web logs, not only the data transferred by using HTTP, but also the data transferred by using other protocols, such as FTP, PTP and etc., Therefore, it is too difficult to extract useful relevant information about the particular title by detecting the sequence of signal in the data streaming transferred through the networks system from the Websites designed in different ways. In session layer a session token is a unique identifier that is generated and sent from a server to a client to identify the current interaction session perfectly. In the networking system lot of problem is already founded by users. In the case of session dealing is also difficult to identify users and user sessions because of the stateless nature of HTTP, it creates very big difficulty to identify users and user sessions from Web logs. Session identification as well as problem retried may be lead many other problem.

Client log file are most accurate to represent in words or pictures the user behavior. It also highly authentic to find the user behavior. Without user corporation or collaboration to modify the browser for each user or client is difficult task in real time. It is in the form of one-to-many relationships of client and web sites visited by that particular user. In addition, log files may contain information supplied directly by the client, without escaping. Therefore, it is possible for malicious clients to insert control-characters in the log files, so care must be taken in dealing with raw logs. Proxy server log files are most complex and more vulnerable to user access data in log file. In this capturing user access data using proxy log file will not give real user behavior. One side we can have unique user login and another side many user using the same IP address.. Proxy server is a type of many-to-many relationships. Many of the user can visit one site and once user can access many site at a time. In this case identifying and capturing the real user and users' browsing behavior is too difficult. In addition to change log file location, to change the log reporting level, to change the log size also create many problem to the administrator.

In web usage mining preprocessing of web log file is necessary . Pattern mining and pattern analysis are paying important attention to pre processing system. Describes novel pattern analysis techniques and applications , details new technology and methods for pattern recognition and analysis in applied domains, examines the use of advanced methods are important recent issue of parent mining in web usage mining. When we enter in to the analysis of web server logs it is necessary to find knowledgeable web site administration, adequate hosting resources and other resources. Many of the present solution of web server log file analysis , only stands theoretically ,in order to give theoretical solution ,much better to have practical solution , that will be very useful to meet many business problems.

Conclusion

This research introduce the theoretical basic of web based data mining and surveys the state of the art of web data mining. With the information overload, Web log mining is a new and promising research issue to help users in gaining insight into overwhelming information on the Web. In this paper, we present a preliminary discussion about Web mining, including the definition, Concepts, and the functions. There still remain many areas for further research, such as the design of efficient algorithms for Web Log miningr for large document collections, and so on. We can do research in Some new techniques can provide the user with the opportunity to analyze the log file at different level of abstraction such as user sessions.

REFERENCES

1. Jaideep Srivastava, Robert Cooley, Mukund Deshpande,pang-Ning Tan, Web Usage Mining : Discovery and applications of usage, patterns from web data. SIGKDO, Explorations, Vol.1.Issue 2, Jan.2000.
2. Zhong Xue Ling, "Semantic web in the core layers of technical analysis",[M],South Chinna Financial Computer Applications Technology,2007,10.
3. Wen- Wei , "Data Warehouse and Data Mining Tutorial ,"[M],Beijing Tsinghua University Press, 2008,
4. W H Inmon.Building the Data Warehouse,the fourth edition. New York: John Wiley, 2005
5. Leise, Fred, "Improving Usability with a Website Index". Archived from the original on 2010-12-28. Retrieved 2010-12-28.
6. Tatsuya Ushioda ,Shigeru Fujita,The RDF – based Information Capturing System from web pages,2010 IEEE International Conference on P2P, Parallel , Grid, Cloud and Internet Computing.
7. Zhang Hui, ed, "Ontology-based Semantic Web Mining Technology."[D], computer development and applications, 2009
8. Qing Gao, Bo Xiao, Zhiqing Lin, Xiyao Chen, Bing Zhou A HIGH-PRECISION FORUM RAWLER BASED ON VERTICAL CRAWLING, Proceedings of IC-NIDC2009.
9. Shi Jing. GONG. Personalization based on Web Mining Service Technology [J]. Computer Science, 2006 (8) :34-36
10. http://www.howmanyarethere.org
11. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press.
12. Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, Cambridge, Mass., 2010. http://www.ksl-consulting.co.uk/google_ indexing_problem.html
14. Yahoo's Senior Director of Search Got It Wrong, Yahoo Uses Meta Keywords Still" SEO Roundtable, October 16 2009, retrieved April 22 2011
15. Srivastava, A., and Sahami. M. (2009). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: CRC Press. ISBN 978-1-4200-5940-3
16. Clifton, Brian (2010) Advanced Web Metrics with Google Analytics, 2nd edition, Sybex
17. Kaushik, Avinash (2009) Web Analytics 2.0 - The Art of Online Accountability and Science of Customer Centricity. Sybex, Wiley. http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-.html.

****