

Web Mining, Process and Applications using Cloud computing

N P PATNAIK M

Assistant Professor,
Department of Computer Science and Engineering,
Viswanadha Institute of Technology and Management, Visakhapatnam,
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India.

Abstract— Due to the rapidly growth in the World Wide Web, there has been large amount of web data generated in various factors. This web data can be mined to provide results which can greatly improve the performance and help business to grow. In this paper, we have discussed about the type of web mining. Which uses the web server logs as data sources known as web usage mining. We have surveyed about the process of web usage mining and the different techniques which includes clustering, classification, association rule mining etc. We also studied about the various applications of web usage mining.

Keywords— Association Rule Mining, Clustering, Pre-fetching, Web Mining, Web Usage Mining.

1. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web Structure and web usage data [1]. The World Wide Web is a huge repository of data which develops increasingly. This data produced by WWW can be discovered to extract information and acquire knowledge using the data mining techniques to further enhance the web.

The web provides different type of information that can be discovered using mining techniques:

- Web user activity, from server logs and Web browser activity tracking.
- Web graph structure, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

Extracting knowledge from the content of documents or their descriptions is known as web content mining. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

This paper gives an overview of web mining, the current state and possible future directions for the Information retrieval. The Web is simply defined as the universe of global network-accessible information. It is an abstract space with which people can interact, and is currently chiefly populated by interlinked pages of text, images and animations, with occasional sounds, three dimensional worlds, and videos. Its existence marks the end of an era of frustrating and debilitating incompatibilities between computer systems.

The explosion of advisability and the potential social and economical impact has not passed unnoticed by a much larger community than has previously used computers. The commercial potential in the system has driven a rapid pace of development of new features, making the maintenance of the global interoperability which the Web brought a continuous task for all concerned. At the same time, it highlights a number of research areas whose solutions will become more and more pressing, which we will only be able to mention in passing in this paper. Let us start, though, as promised, with a mention of the original goals of the project, conceived as it was as an answer to the author's personal need, and the perceived needs of the organization and larger communities of scientists and engineers, and the world in general.

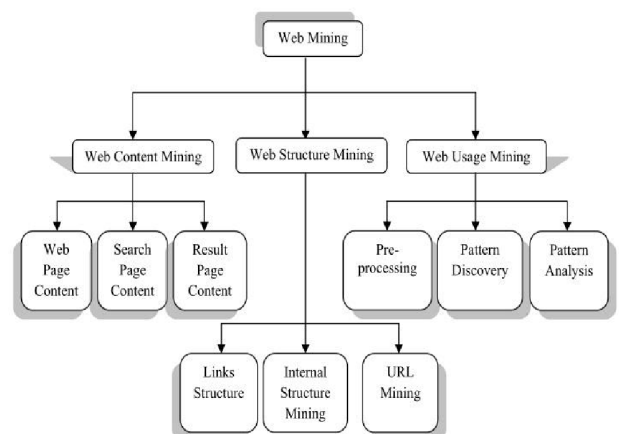


Fig.1. Web Mining Categories

The World Wide Web has involved from an academic and Technical, Scientific network to a board of commercial platform. It has become an integral and indispensable part of our daily life, economic operation, and need for our society. Many technical and non-technical challenges have emerged during this web page transactions in Internet layer.

Technically, the current Internet was design principle. Its containing success has been hindered by more sophisticated network attacks due to the lack of security embedded in the original architecture. web is a clean-slate architecture design. Web is one of the part of Internet. Internet is like a Ocean. In this manner Internet is using gateway through web browser. Browser is created by using software programming languages. The current web is owned and controlled by multiple stakeholders and their network risk of experimentation.

2. BACKGROUND

The bellow Descriptions are expressed in detailed in inner points about main content.

2.1. Internet: He Internet itself is a global, interconnected network of computing devices. This network supports a wide variety of interactions and communications between its devices.



Figure 1: Internal Design of Internet

2.2. World Wide Web: The World Wide Web (WWW) is an open source information space where documents and other web resources are identified by URLs, interlinked by hypertext links, and can be accessed via the Internet.



Figure 2: Universal resource Locator

2.3 Authentication Protocol: a legal user can access its own data fields, only the authorized partial or entire data fields can be identified by the legal user, and any forged or tampered data fields cannot deceive the legal user.

2.4 Internet Protocol: The Internet protocol, or IP for short, is the underlying system of communication for all data sent across the Internet. It is a simple protocol for one computer to send packets of data to another. Suitable example

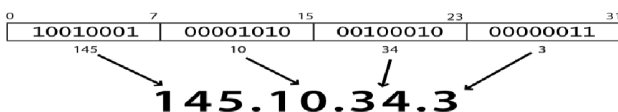


Figure 3: Example of IP Address

2.4.1 Types of IP Class Network

2.4.2 IP Address

2.4.3 Secure gateway Connected Nodes

TCP Port	Services
21	File transfer (FTP)
22	Secure shell (SSH)
25, 110	EMAIL (SMTP, POP3)
80	Web (HTTP)
443	Secure Web (HTTPS)
993, 995	Secure mail

3. RELATED WORK

WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications [2] [3]. A web user whenever interacts with the web server leaves behind traces of information which when analyzed helps improving the design of the system. Web servers record these user interactions which captures the identity or origin of web users with their browsing behavior in the access logs. The three main tasks for performing Web Usage Mining are Data Gathering, Data Preprocessing, Pattern Discovery and Pattern Analysis. Preprocessing consists of converting the usage, content and structure information contained in the various available data sources into the data abstraction necessary for pattern discovery [3].

Web content mining: Web content mining is the process of extracting useful information from the content of Web documents. The contents of a web document are corresponding to the concepts that the document sought to transfer it to users. This content can include text, image, video, sound or records such as lists and tables. The text mining has been studied more than other areas.

Web structure mining: The web can be represented as graph which its nodes and edges are the links between documents. Web structure mining is the process of extracting structural information from the web.

Web usage mining: Web usage mining is the application of data mining techniques to discover patterns using the Web to better understand and meet the needs of the user. This type of web mining explores data relating to the use of web users. It should be noted that there are no clear boundaries between web mining groups. For example, web content mining techniques can use user information in addition to using the documents. It can also be achieved to better results by the combination of above techniques [6].

WEB USAGE MINING PROCESS

Web usage mining Business web usage mining uses statistical methods to explore. But researches focus on developing knowledge extraction techniques that are used analyze the web usage mining data. Three main methods that are used to web usage mining include: Association rules, sequential patterns, and clustering.

Web usage mining applications

The main objective of web usage mining is to collect data about the user's navigation patterns. This information can be used to improve the Web sites in the user view.

Three main applications of this mining are studied in this section [1].

The privatization of web content: Web usage mining techniques can be used for personalization of web users. For example, user behavior can be immediately predicted by comparing her current survey patterns with survey patterns extracted from the log files. Recommendation systems which have a real application in this area are, suggest links that direct the user to his favorite pages. Some sites also organize their product catalogues based on predicted interests of specific user and represent them.

Pre – recovery: The results of web usage mining can be used to improve the performance of Web servers and Web-based applications. Web usage mining can be used for retrieving and caching strategies and thus reduce the response time of Web servers.

The improvement of Web site design: Usability is one of the most important issues in designing and implementing web sites. The results of web usage mining can help to appropriate design of web sites. An adaptive web site is an application of this type of mining. Website content and structure are dynamically reorganized based on data derived from user behavior in these sites.

Data Gathering: Most of the web usage mining systems use the log data as their primary source of data. A Web log file records activity information when a Web user submits a request to a Web server. A log file can be located in three different places:

- Web servers,
- Web proxy servers
- Client browsers.

Data Preprocessing: Preprocessing consists of converting the usage, content and structure information contained in the various available data sources into the data abstraction necessary for pattern discovery [3].

This step includes: i) data cleaning,
ii) Identifying user,
iii) Building sessions and
iv) Path completion.

Pattern Discovery: This step involves using statistical method to carry on the analysis and mine the processed data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

Pattern analysis: Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. Pattern analysis requires the knowledge of query languages such as SQL. The summarized

data is loaded into a data cube and OLAP operations are performed to get better results.

Web usage mining techniques

Pattern discovery phase of web usage mining can be achieved using different techniques such as association rule mining, sequential patterns, clustering and classification.

Association rule mining: Association rule mining is a data mining technique which helps to predict the future events by studying the history of events. It is used to find the frequently accessed pages by the users which are referenced together. Association rule mining using Apriori algorithm [3] may find correlation between users who visited a page containing electronic products to those who access a page about sport related products. A typical example of frequent item set mining is market basket analysis. This process analyzes customer buying habits by finding association between different items that customers place in their “shopping carts (baskets)”. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers [3].

Clustering: The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [2]. In the Web Usage Mining, there are two types of interesting clusters to be discovered, usage clusters and page clusters. Clustering of users tends to find groups of users showing similar browsing patterns. Such knowledge is useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. Also clustering of pages will discover groups of pages having related content. This information is very useful for Internet search engines and Web assistance providers.

Classification: It is the task of categorizing data items into one of several predefined classes. It can be done by using supervised learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc [3]. Weblog information can be integrated with Web content and Web linkage structure mining to help Web page ranking, Web document classification and the construction of multilayered Web information base as well. In a particular discipline, the documents need to be classified based on subject index classification standard such as to classify a set of Web documents automatically, Web linkage information to improve the quality of such classification, use Web usage information to improve the quality of such classification [4].

Sequential Patterns: Sequential patterns discovery is to find the inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. Web log files can record a set of transactions in time sequence. If the web-based companies can discover the sequential patterns of the visitors, the companies can predict users' visit patterns and target market on a group of users.

APPLICATIONS

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize web users). This information can be exploited later to improve the web site from the users' viewpoint. The results produced by the mining of web logs can use for various purposes :

- (i) To personalize the delivery of web content;
- (ii) To improve user navigation through perfecting and caching;
- (iii) To improve web design; or in e-commerce sites;
- (iv) To improve the customer satisfaction.

Personalization of Web Content: Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users [4]. Personalized Site Maps [5] are an example of recommendation system for links. [6] Proposed an adaptive technique to reorganize the product catalog of the products according to the forecasted user profile. A survey on existing commercial recommendation systems, implemented in e-commerce web sites [7].

Pre-fetching and Caching: The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time, as done in [8].

Support to the Design: Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. [9] Uses stratograms to evaluate the organization and the efficiency of web sites from the users' viewpoint. [10] exploits Web Usage Mining techniques to suggest proper modifications to web sites. Adaptive Web sites represents a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behavior [11].

E-commerce: Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure [12].

4. CONCLUSION



Web Mining is the process of extracting knowledge from the huge data repository the World Wide Web. In this paper, firstly we gave an introduction about the web mining and also studied the categories of web mining. Web Mining a part of data mining in terms of Web includes three categories, Web Content Mining, Web Structure Mining and Web Usage Mining. In the next sections, we focus on mainly Web Usage Mining and the process of web usage mining. Then we studied the techniques and different applications of web usage mining.

REFERENCES

- [1]. Web Mining— Concepts, Applications, and Research Directions Jaideep Srivastava, Prasanna Desikan, Vipin Kumar.
- [2]. Srivastava J., Cooley R., Deshpande, M. and Tan, P. N., Web Usage Mining: Discovery and applications of Usage Patterns from web data', SIGKDD Explorations 2000.
- [3]. J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd Edition, Elsevier.
- [4]. S. Shiu C. Wong and S. Pal. Mining fuzzy association rules for web access case adaptation. In Case-Based Reasoning Research and Development: Proceedings of the Fourth International Conference on Case-Based Reasoning, 2001.
- [5]. Fergus Toolan and Nicholas Kushmerick. Mining web logs for personalized site maps.
- [6]. Hye-Young Paik, Boualem Benatallah, and Rachid Hamadi. Dynamic restructuring of e-catalog communities based on user interaction patterns. World Wide Web, 5(4):325–366, 2002.
- [7]. J. Ben Schafer, Joseph A. Konstan, and John Riedl. E-commerce recommendation *Data Mining and Knowledge Discovery*, 5(1-2):115–153, 2001.
- [8]. Bin Lan, Stephane Bressan, Beng Chin Ooi, and Kian-Lee Tan. Rule-assisted prefetching in web-server caching. In Proceedings of the ninth international conference on Information and knowledge management (CIKM 2000), pages 504–511. ACM Press, 2000.
- [9]. Bettina Berendt. Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*, 6(1):37–59, 2002.
- [10]. Yongjian Fu, Mario Creado, and Chunhua Ju. Reorganizing web sites based on user access patterns. In Proceedings of the tenth international conference on Information and knowledge management, pages 583–585. ACM Press, 2001.
- [11]. Tapan Kamdar. Creating adaptive web servers using incremental web log mining. Master's thesis, Computer Science Department, University of Maryland, Baltimore County, 2001.
- [12]. Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, 3(1):1–27,