Asian J. Adv. Basic Sci.: **2016, 4**(*2*), 41-44
ISSN (Print): 2454 – 7492
ISSN (Online): 2347 – 4114
www.ajabs.org

Web Mining - Concept, Classification and Major Research Issues: A Review

Praveen Kumari [1*], Pooja Ranout [2], Ankush Sharma [3] & Pratibha Sharma [4]

[1, 2, 3 &4] *Deptt. of Computer Science and Engineering, Career Point University, Hamirpur,(H.P.) INDIA*
[*] *Correspondance: E-mail: kumaripraveen5@gmail.com*

(Received 31 Mar, 2016; Accepted 29 Apr, 2016; Published 11 May, 2016)

ABSTRACT: Web is a collection of inter-related data or files one or more web server whereas web mining is the application of data mining which is used to extract the valuable information from web database. Web mining also provides better performance system to end users to search for a particular thing and collect information regarding that thing by searching throughout the web server. Aim of this paper is to study about the concept, classification of Web Mining and major research issues in web mining, web content mining, web usage mining and web structure mining.

Keywords: Web Mining; Web Content mining; Web usage Mining and Web Structure mining.

INTRODUCTION: Web is taking one of the important part in human's life and now a day it increases the number of information based on the expectations of the customers or users using it. Daily updation is required to fulfill the customer's need. Web mining is used to extract the web information which is needed by the customers so that the important information can be fetched and utilized. In recent years, we have witnessed an ever-increasing flood of information in the form of digital libraries through the medium of World Wide Web. Although the web is rich with information, but gathering useful data is difficult to identify. A large number of websites, their dynamism, heterogeneity, high linkage and diversity turned the WWW into an entanglement that is hard to specify.

REASONS FOR USING WEB MINING:

- **Finding Relevant Information:** We use search engine for finding specific information on the Web server. We specify simple keywords and in response, we get a list of pages which are ranked based on their similarity with the given keywords. However, finding relevant information regarding keywords is a big problem even with the search engine because it may return some low precision pages, and these pages are not relevant to our query.
- **Discovering New Knowledge:** When we have already collect data from the Web server, we may want to extract more useful knowledge out of this.
- **Customizing Web Pages:** We may want to customize a web page differently for individual users. Every user who seeks information from the Web has its own preferences regarding the style of the contents and presentations. The information providers like to respond to user queries by aggregating information from several sources in a user-dependent manner.

WEB MINING TAXONOMY: According to kinds of data to be mined web mining can be broadly divided into three different categories. Figure 1 shows the taxonomy.

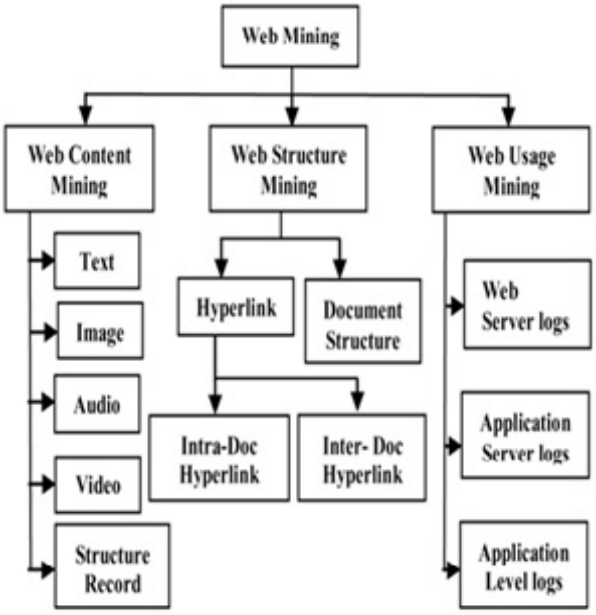

Figure 1: Web Mining Taxonomy. (Source: http://pubs.sciepub.com/ajss/3/2/3/image/fig2.png).

RESEARCH ISSUES IN WEB MINING: web is highly dynamic; lots of pages are added, updated and removed every day and it handles huge set of information hence there is an arrival of number of problems or issues.

Major Issues in Web Mining:

- Web data sets can be very large; it takes large memory to store on the database.
- It cannot mine on a single server so it needs more than one servers.
- Proper organization of hardware and software to mine large database.
- Automated data cleaning.
- Over fitting and under fitting of data.
- Difficulty in finding relevant information.
- Web is dynamic. Information available on Web changes constantly.
- Limited query and limited coverage interface to limited users.
- Over sampling of data.

WEB CONTENT MINING: Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It is the mining and scanning of **text, images, audio, video, or structured records** such as tables. This mining is completed after the clustering of web pages through structure mining and providing the results regarding the queries. With number of pages those contain information regarding query or keywords available on World Wide Web, content mining provide the results to search engine (goggle, yahoo etc) in order to highest relevance to the keywords in the query. In Web content mining there are two types of approaches 1) Agent based 2) Database Approach.

1) Agent Based Approach: These approaches have software systems such as agents that perform the content mining. In the simplest case, search engines belong to this class; perform intelligent search agents, information filtering and personalized Web agents.

- **Intelligent Search Agents** go beyond the simple search engines and use other techniques besides keywords searching to accomplish a search. For example, they may use user profile or knowledge concerning specifies domains.
- **Information Filtering** utilizes IR techniques, knowledge of the link structures and other approaches to retrieve and categorize documents.
- **Personalized Web Agents** use information about user preferences to direct their search.

2) Database Approach: This approach view the Web data as belonging to a database. There have been approaches that view the Web as a multilevel database and have been many query languages that target the Web.

One problem associated with retrieval of data from Web documents is that they are not structured as in traditional databases and there is no scheme or division into attributes. Traditionally Web pages are defined using hypertext markup language (HTML).

Major Issues in Web Content Mining:

- Information extraction concentrates on extraction of structured data from web pages such as keywords of product and search results.
- Web contain large amount of data, each website accept similar data in different ways.
- Web pages created using HTML are only semi structured, thus making query is more difficult than with well-formed database.
- Automatically segmenting web pages and detecting noise is an interesting problem in web application.

WEB STRUCTURE MINING: Web structure mining is the study of data interconnected to the structure of a particular website. It consists of web graph which contains the web pages or web documents as nodes and **hyperlinks** as edges those are connecting between two related pages. Below figure represent the web graph structure.

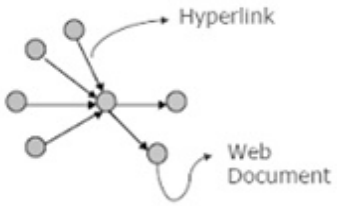

Figure 2: Web Graph Structure. (Source: http://www.expertsupdates.com/ArticleAttachments/seo/web-mining/Figure2.gif).

Based on the topology of the hyperlinks, Web structure mining categorizes the web pages and generates information about similarity and relationship between different websites. Web structure mining can be performed either at intra-page level or inter-page level. A hyperlink that connects two different part of the same page is called **intra-page level hyperlink**. A hyperlink that connects two different pages is called **inter-page level hyperlink** which is structure level.

The main purpose of structure mining is to extract previously unknown relationship between web pages. This structure data mining provides use for a business to link the information of its own Website to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining.

According to the type of Web structural data, web structure mining can be divided into two categories:

I. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects web pages to a different link.
II. Mining the document structure: analyses of tree like structure of a page structures to describe HTML or XML tag usage.

Major Issues in Web Structure Mining:

- The related page of search information becomes unavailable or unorganized due to search engine's problems and often tolerate for low precision criteria.
- Indexing information on the Web server. This also cause low amount of recall with content mining.

WEB USAGE MINING: Web usage mining is also known as **web log** mining which is used to analyze the behavior of online users or we can say that Web usage mining performs mining on Web usage data or Web logs. A Web log is a listing of page reference data. Sometimes it is referred to as click stream data because each entry corresponds to a mouse click. These logs can be examined from either a client perspective or a server perspective. When evaluated from a server perspective, mining do not cover information about the sites where the service resides. It can be used to improve the design of the sites. When evaluating client perspective, a client's sequence of clicks, information about the user or group of users is detected. This could be used to perform prefetching and caching of pages.

time	database_name	direction	class	time_period	quantity
2010-07-10 00:00:00.000	master	Ingress	External	OffPeak	8
2010-07-12 12:00:00.000	AdventureWorksLT2008R2	Egress	External	OffPeak	25162
2010-07-12 12:00:00.000	AdventureWorksLT2008R2	Ingress	External	OffPeak	5
2010-07-12 12:00:00.000	master	Egress	External	OffPeak	35
2010-07-12 12:00:00.000	master	Ingress	External	OffPeak	14
2010-07-12 18:00:00.000	AdventureWorksLT2008R2	Egress	External	OffPeak	18347
2010-07-12 18:00:00.000	AdventureWorksLT2008R2	Ingress	External	OffPeak	2
2010-07-12 18:00:00.000	master	Egress	External	OffPeak	12
2010-07-12 18:00:00.000	master	Ingress	External	OffPeak	6
2010-07-13 12:00:00.000	master	Egress	External	OffPeak	7
2010-07-13 12:00:00.000	master	Ingress	External	OffPeak	2
2010-07-13 18:00:00.000	master	Egress	External	OffPeak	24
2010-07-13 18:00:00.000	master	Ingress	External	OffPeak	12

Figure 3: Web Logs.

(Source:http://social.technet.microsoft.com/wiki/contents/articles/3398.windows-azure-sql-database-delivery-guide-forbusinesscontinuity.aspx?Sort=MostRecent&PageIndex=1).

Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data: User logs are collected by the Web server and typically include IP address, page reference and access time.

Application Server Data: Commercial application servers such as Web logic, Story Server have significant features to enable E-commerce applications to be built on top of them with little efforts. A key feature there is the ability to track various kinds of business events and log them in application server logs.

Application Level Data: New kinds of events can be defined in an application and logging can be turned on for them-generating histories of these events. However, that may end applications require a combination of one or more of the techniques applied in the above categories.

Major Research Issues in Web Usage Mining:

- Session identification problem: single IP address and multiple users are there or rotating IP address for load balancing.
- Common gateway interface (CGI) data problems: The relevant data for determining what page was accessed may not be present in CGI pair.
- Caching problems: Clients and proxy servers save local copies of pages that have been accessed.
- Wrong access timings recorded at server.
- Incompleteness of server logs due to caching of pages.
- Transaction identification: user ids are often suppressed due to security concerns.

CONCLUSION: On the basic of above study two major reason that if the keywords that we are searching in our database are inappropriate to our requirement. We also conclude that if the Common gateway interfaces (CGI) doesn't allow the keyword to search in the database where that Common Gateway Interface (CGI) is interface between the data required and database in which data present.

ACKNOWLEDGEMENT: This research paper is supported by Computer Science and Engineering Department of Career Point University Hamirpur, Pratibha Sharma (HOD) and A.P. Ankush Sharma.

REFERENCES:

1. S. Vijiyarani and E. Suganya (2015) Research Issues in Web Mining, *International Journal of Computer-Aided Technologies*, 2(3), 55-64.
2. Pooja Mehta, Brinda Parekh, Kirit Modi and Parash Solanki (2012) Web Personalization Using Web Mining: Concept and Research Issue, *International Journal of Information and Educational Technology*, 2(5), 510-512.
3. Pradeep Mittal and Monika Yadav (2013) Web Mining: An Introduction, *International Journal of Advance Research in Computer Science and Software Engineering*, 3(3), 683-687.
4. D. Jayalatchumy, P. Thambidurai (2013) Web Mining Research Issue and Future Directions *IOSR Journal of Computer engineering*, 14(3), 20-27.
5. Raymond Kosala and Hendrik Blockeel (2000) Web Mining Research: A Survey, Kotholieke University Leuven Celestijnenlaan 200A, B-3001 Heverlee, Belgium, 2(1), 01-15.
6. Qingyu Zhang and Richard S. Segall (2008) Web Mining: A Survey of Current Research, Techniques and Software, *International Journal of Information Technology and Decision Making*, 7(4), 683-720.
7. S. Balan and P. Ponmuthuramalingam (2013) A Study if Various Techniques of Web Content Mining Research Issue and Tools, *International Journal of Innovative Research and Studies*, 2(5) 508-517.
8. Reema Thareja Data Warehousing (YMCA Library Building,1, Jai Singh Road, New Delhi 110001, India) GGP IT University New Delhi, 391-395.
9. Margaret H. Dunhain Data Mining: Introductory and Advanced Topics (Dorling Kindersley (India) Pvt. Ltd., licensees of Pearson Education in South Asia) Southern Methodist University, 193-206.
10. Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi (2013) Overview of Web Content Mining Tools, *International Journal of Engineering and Science*, 2(6), 106-110.
11. Claudia Elena Dinuca, Dumitru Ciobanu (2012) Web Content Mining, University of Petrosani, Economics, 12(1), 85-92,
12. A. Jebarajratna Kumar (2005-2010) An Implementation of Web Personalization using Web Mining Techniques, *Journal of Theoretical and Applied Information Technology* JATIT, 67-73.
13. Jaideep Srivastava, Prasanna Desikan & Vipin Kumar Web Mining: Accomplishments and Future Directions University of Minnesota USA.