Research On Web Data Mining Concepts, Techniques And Applications

K.Jayamalini
Research Scholar
Department of Computer Science & Engineering
Bharath University
Chennai, India
malini1301@gmail.com

Dr.M.Ponnavaikko
Vice-Chancellor
Bharath University
Chennai, India
vc@bharathuniv.ac.in

***Abstract*— WWW is, by far, the enormous database in the world, holding humongous amount of data of various types that would be consumed by users for various needs. It contains priceless data for businesses, if mined effectively. Web mining is a process that targets to find useful information or knowledge from the Web page contents, hyperlink structure, and usage or sever logs of websites. Web data mining is divided into three major groups - Web Content mining, Web Structure mining and Web Usage mining. This survey paper reports the basic concepts of each type of web mining, techniques for data extraction or knowledge detection and uses of different types of web mining methods.**

Keywords* - *Usage Mining; Structure Mining; Content Mining;

I. Introduction

World Wide Web (WWW), simply the Web, which is considered the world's largest database, holding enormous amount of various types of data that would be consumed by us for various needs. Web data is unstructured raw information. It is a virtual gold mine for business and people, if it is mined properly. Web mining [1] is a process that targets to determine useful data from web page contents, hyperlinks organization, and server logs of websites. Web mining [2,3] processes are categorized into Web Content Mining(WCM), Web Structure Mining(WSM) and Web Usage Mining(WUM) established on the type of statistics used in the mining methods, which is given below in Fig.1.

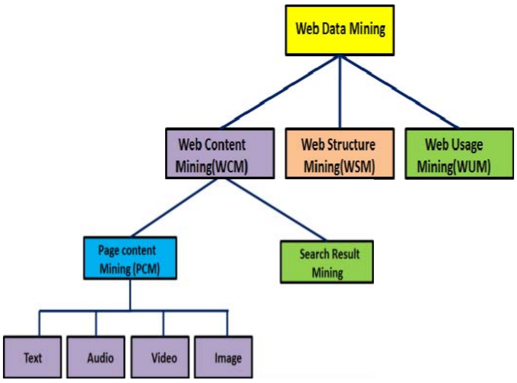


Fig.1. Web Data Mining Basic Classifications.

Web content mining is used to mine different web data like text, image, audio and video contents. Web structure mining is used to obtain the hyperlink organization of the web site which delivers improved search outcomes for the user. Web usage mining is a method of obtaining valuable knowledge from the webserver log files, log files of users and frequently accessed paths.

II. Web Content Mining - Concepts And Its Techniques

A. *Web Content Mining - Concepts*

WCM [4] is method of discovering valuable information such as text, images, audio, and videos from web pages i.e. it identifies the useful information from the Web contents. It is classified into Page Content Mining (PCM) and Search Result Mining as shown in Fig 1.

Page content mining discovers useful data like text and multimedia contents from the webpages. Search result mining discovers hidden data or deep data from the Search Engine Result pages (SERs). SERs are special web pages which are generated dynamically when the user executes the web query through Web Query Interface.

B. *Page Content Mining –Techniques*

Data mining techniques like summarizing, classification and clustering are used to discover useful and interesting patterns from web contents as per user needs. Various types of WCM techniques [5, 6] are given below:

- Unstructured data mining techniques - refer to the process of finding facts from unstructured webpages. It is further divided into two types:
  - Text Mining – refers to knowledge discovery from textual databases or method of extracting interesting patterns from textual documents.
  - Topic Tracking – is a technique by which an authorized customers can trace their topic of

interest. Registered users can get intimated by messages about their topic of interest, whenever there is an update regarding the interest of the user.

- Semi-structured data - refers to a set of data, in which some implicit structure is generally followed e.g. Web documents and database
- Structured data mining [7] techniques are processes which refer to extracted data from the semi structured webpages. When the user query is executed through "search interface" provided on the website, the matching data is fetched from database of the website and displayed to the user in the form of webpage with a fixed template. These pages are called as data rich pages. "Web Wrappers" is a software program which is used to extract data from search engine result pages.
- Multimedia mining - refers to the procedure of discovering useful facts from multimedia data such as audio, video, image and text. By basic queries these data are not accessible.

III. WEB STRUCTURE MINIG - CONCEPTS AND ITS TECHNIQUES

A. *Web Structure Mininig – Concepts*

WSM [8] is method of determining useful information using the hyperlink organization of the Website. While WCM mainly emphases on facts that is provided by Web document contents, the Web offers extra information using hyperlinks through which various web documents are linked to each other. The Web is represented as a directed graph in which vertices are html pages and the edges are the hyperlinks which are used to connect other pages. The hyperlink which points towards a particular web document from other web documents is called as in-links and the hyperlinks which are generated from particular web documents to point other web pages are called as out-links, which is depicted in Fig 3. The two terms which are used in link analysis are:

- ○ In-degree: Total number of hyperlinks pointing to particular node.
- ○ Out-degree: Total number of hyperlinks generated from particular node.

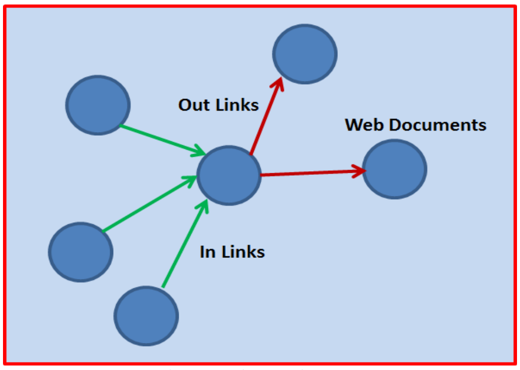


Fig.2. Web Graph Structure

B. *Page Structure Mining –Techniques*

Web link analysis techniques are mainly used in

- ○ web search engines
- ○ social network analysis (SNA)

A web search engine is a software application which is used to search for information on the WWW. SNA [9] is the measuring and mapping of flows and relationships between organizations, computers, people, groups, URLs, and other connected information entities. The people and groups in the network are considered as nodes while show relationships or flows between the nodes are considered as the links.

SNA is used in various applications like

- ○ aggregation of data
- ○ mining of data,
- ○ development of recommender systems
- ○ sampling and network modeling,
- ○ user behavior analysis,
- ○ community detection,
- ○ location-based interaction analysis
- ○ social sharing and filtering

The PageRank and the HITS algorithms are to analyze the link organization of the web sites.

- PageRank Algorithms [10] are used to increase the effectiveness of search engines. It also identifies the most important and useful pages as per user query It computes rank for each pages and assigns the values to a rank vector. Page rank for each page is computed by calculating the number of pages that are connecting to it and called as backlinks. More the count of the backlinks to a web page, that page is considered as of high importance. If pages having a higher rank or it is pointed by many pages then that page gains a higher rank.
- HITS [11] stand for Hypertext Induced Topics Search, which is developed by Jon Kleinberg. This algorithm is applied on a sub graph after a search is done on the complete graph. It uses hubs and authorities to define a relationship between web pages and to calculate the page rank.
    - ○ An authority page is a webpage which contains valuable information, which is pointed to by a many number of hyperlinks.
    - ○ A hub page is a webpage that points to many authority pages.

IV. WEB USAGE MINIG - CONCEPTS AND ITS TECHNIQUES

A. *Web Usage Mininig - Concepts*

WUM [12, 13] is a process that exercises data mining techniques to examine and find valuable information from user's log files. The log files contain the user's activities when the user surfs on the website. The web servers record the details of user activities automatically in server log files. Organizations often collect huge volumes of such data and analyze those data

- to find the value of individual customers
- to plan effective promotional campaigns
- to find the cross promotional marketing across products

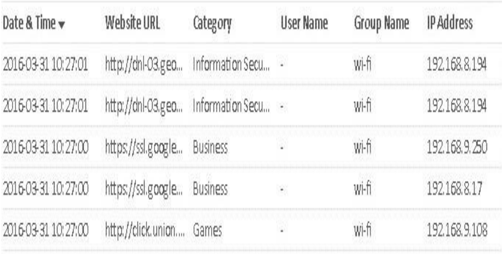
Different kinds of usage data which is used in the process of Web usage mining are:

- **Web Server Data** - User details are automatically recorded in the Web servers, which contains user information like address of Internet Protocol, page reference, agent name and time of accessing website.

- **Application Server Data -** E-commerce applications use Commercial application servers. It stores various kinds of business functions and logs related to those.

- **Application Level Data** – logs generated and recorded at application level

Example server log is shown below in fig 3.



Date & Time ▾	Website URL	Category	User Name	Group Name	IP Address
2016-03-31 10:27:01	http://dnl-03.geo...	Information Secu...	-	wi-fi	192.168.8.194
2016-03-31 10:27:01	http://dnl-03.geo...	Information Secu...	-	wi-fi	192.168.8.194
2016-03-31 10:27:00	https://ssl.google...	Business	-	wi-fi	192.168.9.250
2016-03-31 10:27:00	https://ssl.google...	Business	-	wi-fi	192.168.8.17
2016-03-31 10:27:00	http://click.union....	Games	-	wi-fi	192.168.9.108

Fig.3. Example for Web Sever Log

B. *Web Usage Mining –Techniques*

In WUM, three main phases are used to examine and determine interesting information from user logs.

- Preprocessing - converts the server log file contents, which normally is in text format, into a structured format, which can be further processed by data mining algorithms. Some of the steps [14, 15] performed in preprocessing of web server logs.

    - Data Cleaning – elimination of irrelevant information from server logs.
    - User Identification – recognition of unique users by combing IP addresses, user agents and referrers.
    - Session Identification - used to find the session. It is actual sequence of activities performed by one user during one visit to the site. This information is identified from click stream data. Timeout is the simplest way to detect session. If the time for page request exceeds certain limit, then it is assumed that the user started a new session.
    - Formatting – After applying appropriate preprocessing steps to the log files, a complete preparation module is made, which contains properly formatted details so that data mining algorithms to be accomplished on those data.
- Pattern discovery - applying the data mining techniques like frequent pattern discovery techniques on the server log data to find the interesting patterns. The three main operations carried out at this phase are:
    - Association – shows the relationship between pages which are accessed together
    - Clustering - discovery of group of users, dealings, pages, etc.
    - Sequential analysis – deals with the order of web pages to be accessed
- Pattern analysis - understanding the outcomes found by data mining algorithms on server logs and describing final decisions. In this phase, the motivation is to keep only interested patterns and strain out uninterested rules or information found in earlier phase.

The steps carried out in WUM technique [16] is shown in Fig 4.

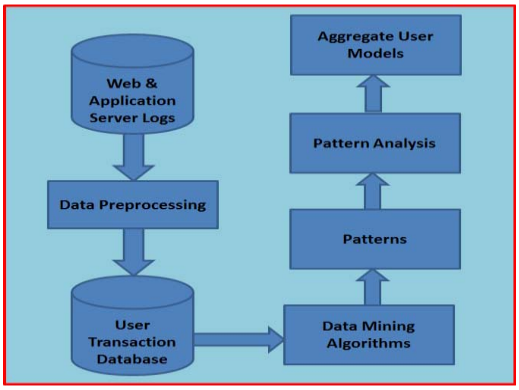


Fig.4. Steps in Web usage Mining

V. WEB DATA MINING – PROBLEMS, ISSUES AND APPROACHES

Web data mining approaches are used to mine various kinds of data, which are used in various applications. Various mining algorithms are used in the web mining. The unstructured web data should be converted into structured data before using the data mining algorithms. The major problems and issues [17,18] in the field are web data mining are:

- Data Cleaning and Integration
- Preprocessing - Html pages on Web is usually unclean, ill-formed and not suitable for further processing. The semi-structured nature of HTML language make the HTML files non-standardized. It is necessary to first clean them up before any serious use of these documents. Several html parsers like HtmlCleaner, NekoHTML, Html Parser, Jsoup and JTidy that can be used to clean up the html pages.
- Duplicate Data elimination methods are still a tedious task to perform. No standard tools are available.
- Over Fitting of data
- Under Fitting of data
- Oversampling of data
- Scaling up high dimensional data
- Mining of time series data
- Altering data mining algorithms for huge amount of data
- Dealing with unstructured and unbalanced data
- Mining text data streams, hyperlink structures and networks
- Difficult to execute queries on heterogeneous data
- Security in web data mining

VI. APPLICATIONS OF WEB MINING

Web mining is used in several fields [19]. Some of the applications that use web mining techniques are:

1. Cloud user makes use of web mining procedures to mine the information from the cloud servers.
2. Ecommerce websites make use of web mining methods to get the detailed info about products and specifications of products.
3. Search engines allows the user to search over 2 billion data. It assigns the rank to each page. Based on page rank it orders the pages and publish the pages based on the user query.
4. Web mining methods used to track individual web sessions effectively. It provides valuable information about user behavior.
5. Group of users of same interest called Web community can be clustered and maintained, to connect with each other and share information among them through the network.
6. Web page personalization plays important role now a days. It is used to maintain the personal and confidential information about the user. Web mining is used to maintain personalized data about the users.
7. Automatic citation and indexing is performed in Digital library using web mining techniques.
8. E-services like on-line knowledge management, e-banking, search engines, blog analysis, on-line auctions, social networking, e-learning, recommendation systems and personalization are analyzed and enable recommendations to customers.
9. Analysis of text documents at the disposal of a company (like e-mails, human resource records, technical and legal documentation)
10. Crawling of Social Web platforms like Facebook, YouTube or Flickr - to verify sociological theories on a large scale or to check whether mathematical models developed in the field of complex networks are able to correctly explain human behaviors.
11. Web mining plays important role in Business and Competitive Intelligence, which is process of collecting actionable information on your business based on competitive environment.
12. Web mining methods are used in Comparison Shopping where business filters and compares products of competitors based on price, features, and other criteria
13. Terrorism is controlled through web mining [20]. Terrorists use websites, forums, Blogs, social networking sites, and virtual worlds for their information exchange. Careful analysis of data collected from these helps to control the terrorism.
14. In E-commerce era, web mining is used to find various requirements of customers, their opinion

about products and services, in order to ensure competitiveness.

15. Companies use Opinion mining technique to extract reviews of about customers and their products and its specifications.

VII. WEB DATA MINING APPROACHES – DRAWBACKS

The following points are considered as some of the drawbacks of web data mining approaches are:

1. Too long response time by user.
2. heavy traffic on networking due to massive growth of web
3. Huge amount of web resources and web servers on www.
4. Any solution to improvise the quality and efficiency of web services would increase the bandwidth, which in turn increase the cost.
5. In web caching scheme, if the proxy server details is not properly updated,
    a. stale data to users
    b. increases the the number of users
    c. bottleneck in servers

VIII. CONCLUSIONS

This survey paper explains the concepts of Web mining and its classifications. It also explains the details about the techniques and data mining algorithms used to discover useful information from each category. Finally this paper explains the applications of web data mining in various fields. Now-a-days, web contents give new challenges to the traditional data mining algorithms that work on flat data. This paper also deals with how the traditional data mining algorithms have been modified or new algorithms have been developed to work on the Web data. This research paper simply generalizes the concepts, techniques and applications in the area of web mining.

References

[1] Kavita, P. Mahani and N. Ruhil, "Web data mining: A perspective of research issues and challenges," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 3235-3238.

[2] G. Chen, "Application of Web Data Mining Technique to Enterprise Management of Electronic Commerce," *2014 Seventh International Symposium on Computational Intelligence and Design*, Hangzhou, 2014, pp. 154-157.

[3] Sunena and K. Kaur, "Web usage mining-current trends and future challenges," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 1409-1414.

[4] K. Hassani and W. S. Lee, "Adaptive animation generation using web content mining," *2015 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, Douai, 2015, pp. 1-8.

[5] A. Dutta, S. Paria, T. Golui and D. K. Kole, "Structural analysis and regular expressions based noise elimination from web pages for web content mining," *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, New Delhi, 2014, pp. 1445-1451.

[6] Govind Murari Upadhyay, Kanika Dhingra," Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013, pp.610-613.

[7] Kolkur, Seema, and K. Jayamalini. "Web Data Extraction Using Tree Structure Algorithms–A Comparison." International Journal of Recent Technology and Engineering (IJRTE) volume 2,Issue 3,July 2013,pp.35-39.

[8] D. Gupta and D. Singh, "User preference based page ranking algorithm," *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, 2016, pp. 166-171.

[9] http://www.orgnet.com/sna.html accessed on May 2016.

[10] Ms.M.Sangeetha, Dr.K.Suresh Joseph," Page Ranking Algorithms used in Web Mining", ICICES, IEEE 2014,pp1-7.

[11] W. Yang, "An Improved HITS Algorithm Based on Analysis of Web Page Links and Web Content Similarity," *2016 International Conference on Cyberworlds (CW)*, Chongqing, China, 2016, pp. 147-150.

[12] Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah,"Web Usage Mining: A Review of Recent Works", Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference,IEEE 2014,pp.1-5.

[13] Bhupendra Kumar Malviya, Jitendra Agrawal," A Study on Web Usage Mining: Theory and Applications", Fifth International Conference on Communication Systems and Network Technologies,IEEE2015,pp.935-939.

[14] Y.Han and K.Xia, "Data Preprocessing Method Based on User Characteristic of Interests for Web Log Mining," *2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, Harbin, 2014, pp. 867-872.P.

[15] V. Anitha and P. Isakki, "A survey on predicting user behavior based on web server log files in a web usage mining," *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Kovilpatti, 2016, pp. 1-4.

[16] www.cs.famaf.unc.edu.ar/~laura/llibres/wm.pdf.gz accessed on May 2016.

[17] D. Jayalatchumy, Dr. P.Thambidura," Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE),Volume 14, Issue 3, Oct. 2013,pp. 20-27.

[18] J Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology .Volume 11,Issue 4 , March 2011.pp.1-7.

[19] Ananthi.J, "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Volume 5, Issue (3, 2014, pp.4091-4094.

[20] T.Anand, S.Padmapriya, E.Kirubakaran," Terror Tracking Using Advanced Web Mining Perspective", Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference (lAMA), IEEE 2009,pp.1-4.