International Journal of Trend in Research and Development, Volume 3(5), ISSN: 2394-9333
www.ijtrd.com

A Study on Web Mining Types and Applications

K.Harish Kumar
Assistant Professor, Department of Computer Science & informatics, M.G. University,
Nalgonda, Telangana, India

***Abstract*—** Internet or World Wide Web emerged rapidly from past two decades which became crucial part of every ones life. Internet is a supreme fountain of data for all online realms and serves as important communications channel. Web becomes primitive repository for extracting data online. Because of intensifying growth data available on line and continuous growth in size of data available online it is difficult to extract desired information as per query of user. Web Mining is a part of data mining which is used to extract information from web servers based on user's wishes. It is an amalgamation of WWW and data mining. Web Mining furthermore organized into three segments-Web Content Mining, Web Usage Mining and Web Structure Mining. Web Content Mining is an activity, in which extracting convenient information from the contents of Web documents. Web Structure Mining, we determinate the structure of the data from the Internet. Web Usage Mining is the disclosure of significant information from data induced by client-server events on one or more places. Web Mining measured by data mining ways, by means of association rules, classification and clustering. It has a few advantageous areas or applications such as E-governance, Business Computing, digital libraries, digital learning, digital policies, data analytics, E-democracy, online security, criminal forensics and Electronic commerce. The focus of our paper is to study the types of web mining techniques and applications.

***Keywords*—** *Web mining; Web usage mining, Web content mining; Web structure mining;*

I. INTRODUCTION

Web is a huge, widely heterogeneous, humongous, knowledgeable, discoverable, accessible and interconnected information repository. It is difficult for user to find desired information form enormous data without any knowledge. Web Mining serves as knowledge factor for user.

Web Mining is the application of data mining technique which is an unstructured or semi structured data and it inevitably locates and extracts potentially useful and previously unknown information or knowledge from the web [1]. Etzioni who is given the first time definition of Web Mining in 1996, says that various kinds of information are available on the World Wide Web and the information is structured on their respective resource [4]. Etzioni refers to that the various kinds of information is most usefully and the information is previously unknown from the others websites.

Web Mining Process: The Web Mining Process used for getting useful knowledge from web data is given in below Figure [5].

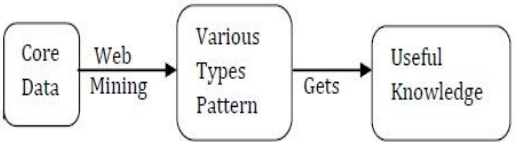

Figure 1: Web Mining process

Steps in Web Mining:

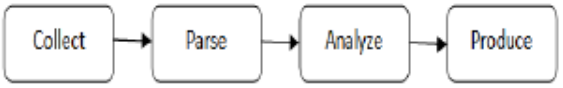

Figure 2: Steps in Web Mining

- Collect - fetch content from web
- Parse - extract data from formats
- Analyze - tokenize, rate, classify, cluster
- Produce - useful data

Web mining techniques are decomposed into the following subtasks[1]:

Resource Discovery

It is responsible to find the web information from various sources.

Information selection and pre-processing

It pre-processes the data collected from web automatically.

Generalization

Patterns are automatically discovered at both the different sites and distinct sites.

Analysis

It certifies the data which is mined. According to Praveen Kumari [6] the reasons for using Web Mining are

Finding Relevant Information: We use search engine for finding specific information on the Webserver. We specify simple keywords and in response, we get a list of pages which are ranked based on their similarity with supplied tokens. Nevertheless, detecting admissible 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: We have already collect data from the Web server, we may want to derive more helpful knowledge out of this.

Customizing Web Pages: We may want to customize web pages differently for individual users. Every user who seeks information from the Web has his/her own priorities concerning the way of the contents and presentations. The information providers respond to user queries by accumulating information from several sources in a user-dependent manner.
Web mining can be roughly categorized into three types[7]:

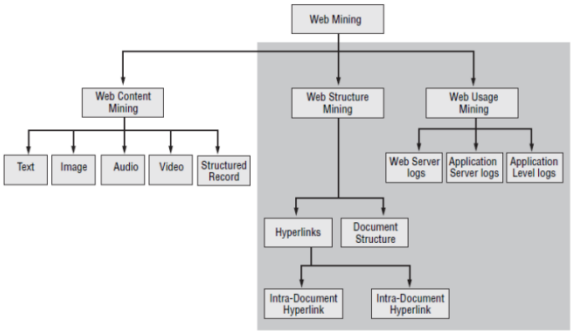

Figure 3: Web Mining types

A. Web Usage Mining

It provides the information that describes the usage patterns of Web pages, such as IP addresses, page references, date and time of accesses, other information depending on the log format, free texts, HTML Files, XML Files, Dynamic Content, and Multimedia Files [9].

B. Web Structure Mining

Web Structure Mining is about extracting knowledge from the hyperlinks. Significant web pages can be identified; also users that have common interests, i.e. using the identical clusters of linked pages. Till 1996, pages were fetched based on content similarity. From beginning i.e. from 1997 the largely used hyperlink search algorithms were PageRank and HITS (Hypertext Induced Topic Search). These algorithms are influent to social network analysis (measures of the degree of prominence of an actor in a social network). Pages are rated pursuant to their prestige or authority. The Web is considered a virtual social network pages being the hyperlinks, social actors and the relationships. In this way models and techniques from social networks analysis can be sent to web structure mining[8].

C. Web Content Mining

Content means the perceptible data in the web pages or the data which was denoted to be make known to the users. A vital part of it encompasses text and graphics (images). Since a text document put forwards as not machine-readable semantic, so some ways have been suggested which are used to restructure content of document in a description so that machines can recognize it such as Free texts, HTML Files, XML Files, Dynamic Content, Multimedia Files. The accustomed approach to make use of familiar structure in documents is to use wrapper to point the documents to certain data model. There are principally two categories for web content mining strategies, one directly digs the document's contents and other refines on the search contents of other tools like search engines[9]. These three categories discussed followed by next sections.

II. WEB USAGE MINING

The web usage mining has emerged as the essential tool for realizing more personalized user pleasant and business optional web services based on data logs of user interaction with the web including citing pages and user recognition, web log may be proxy server logs, web server logs, and browser log. In web mining data usage is site content data of visitors, web server logs gathered from external channels future application data. It is helpful practice which can give assistance in constructing a web site so that, top class service may be provided. A big part of web usage mining involves dealing with usage stream of data, subsequent to that different data mining algorithms can be applied in three phases[10].

- Preprocessing
- Pattern discovery
- Pattern analysis

A. Preprocessing

The preprocessing is an action to rearranging the web log data before processing. This is also called as cleaned data. The undesirable data are trashed and curtailed log file is obtained. Preprocessing has different sections like:

Data cleaning: Clears deviated and irrelative data.

Session identification: Divide all page accessed by a user into session.

Data conversion: It is converting the log file data into the format needed by the mining algorithms.

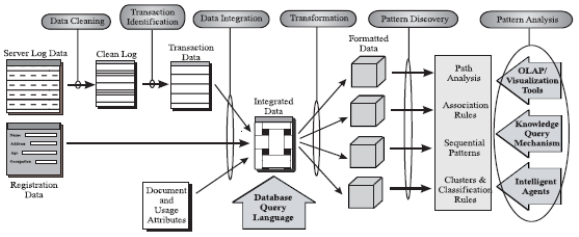

Figure 4: Web Usage Mining Process

B. Pattern Discovery

Data mining techniques are applied to find the engrossing characteristics in the hidden pattern. Subsequent to preprocessing of log file into well-organized data the pattern discovery process is undergone, existing data of the log files may useful patterns are detected with some files.
Pattern discovery has different sections like

- Path analysis
- Association rules
- Special patterns
- Cluster and classifications

C. Pattern Analysis

Pattern analysis is the endmost phase of web usage mining which can verify interested patterns from the result of pattern discovery that can be used to user behavior.

Knowledge query mechanism: SQL is pretty regularly used language for knowledge query mechanism. The language is applied in order to extract the useful patterns from discovered patterns.

Visualization tool: OLAP provides a desegregated frame work for analysis is which allows change in aggregate level.

Intelligent agent: Various agents are also devised that helps in examining the pattern in web usage mining, these agent perform the work of analysis the discovered patterns[10].

III. WEB CONTENT MINING

It gives detailed account about the finding of useful information from Web documents. Basically, Web content consists of several types of data like metadata, text, audio, hyperlinks, image as well as videos. Research in mining multiple types of data is now entitled as multimedia-data mining. We could contemplate multimedia-data mining as an object of Web-content mining. The Web content data embodies sloppy data such as free text, semi-structured data such as HTML documents, and complex structured data such as tables and HTML pages generated by database. The goal of Web-content mining is predominantly to give assistance or to ameliorate information-finding or filtering the information. Structuring a new model of data on the Web, further knowledgeable queries other than the keyword-based search could be asked. Topic discovery, clustering of Web documents, extracting association patterns and classification of Web pages are some of research issues in text mining. These actions use approaches from other regulations - IR, IE (information extraction), NLP (natural language processing) and others [11] [12]. Automatic extraction of semantic relations and structures from Web is a growing application of

Web content mining. Varying types of algorithms are used in Web content mining: Hierarchical clustering algorithms on terms in order to create formal concept analysis, concept hierarchies, and association rule mining to learn. Primitive Steps of Web Content Mining are

• *Collect* – fetch the content from the Web
• *Parse* – extract usable data from formatted data (HTML, PDF, etc)
• *Analyze* – tokenize, rate, classify, cluster, filter, sort, etc.
• *Produce* – turn the results of analysis into something useful (report, search index, etc)

Some of the prominent web content mining techniques are:

- Unstructured data mining techniques
- Structured data mining techniques
- Semi structured data mining techniques
- Multimedia data mining techniques

A. Unstructured data mining techniques

One of a kind of web content mining is unstructured mining. In this number of the web pages is in the proportion of text. Corresponding to this technique the data is searched and retrieved. It is not obligatory that the data which is extracted is significant data, it may be unknown information. We have to use some tools or methods to get pertinent data/information from that data. It is categorized into two types.

Text Mining for Web Documents

Text Mining is a sub-part in the domain of data mining techniques. Retrieval of knowledge from dynamic HTML pages in itself is a challenging task. However, with the advent of modern tools such as SVM, Decision trees and IEPAD, the results that are coming out is of much high accuracy [13].

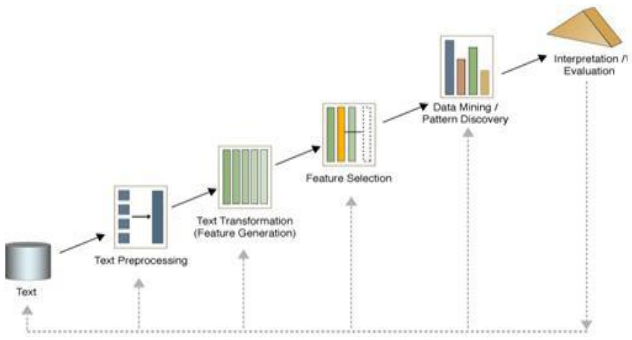

Figure 5: Text Mining for Web Documents [13]

Topic Tracking

Topic tracking is the technique by which registered users can track the topic of their will. User need to register with the topic, whenever there is any update in correspondence with the interest of the user, he will intimate by the message.

B. Structured data mining techniques

An advancement of retrieving information from web pages is structured data extraction. A program for extracting such data is usually called as wrapper. Structured data are normally data records retrieved from concealed database and exhibited in the web pages following a few templates. Erstwhile, the template is a table. Sometimes, it is a form.

Intelligent Web Spiders

Web spiders are certainly known as crawlers which look for the information across the WWW. Web crawlers are principally used to create a replica of each visited pages for later handling by search engine that will archive downloaded pages to furnish fast searches. Spiders have number of applications i.e. personal searches, building-up the search databases, web site backups and so on.

C. Semi structured data mining techniques

Semi-structured data is a point of confluence for the Web and database circles: the former deals with documents, the latter with data. Emergent renditions for semi- structured data (such as XML) are contrasts on the Object Exchange Model (OEM). In OEM, data is in the sort of atomic or compound instances: atomic instances may be integers or strings; compound objects refer to other objects via tagged edges. HTML is a precise example of such "intra- document" structure [14].

D. Multimedia data mining techniques

Multimedia data mining defined as an affair of finding interesting patterns from media data such as text, video, image and audio that is difficult to access by basic queries and associated results. The motivation for Multimedia data mining is to use the founded patterns to enhance decision making. Multimedia data mining has consequently enticed notable research efforts in developing methods and tools to organize, manage, search and accomplish domain specific chores for data from domains such as surveillance, meetings, broadcast news, sports, archives, movies, medical data, furthermore individual and online media collections.

IV. WEB STRUCTURE MINING

The structure of a typical Web graphs contains Web pages as nodes, and hyperlinks as edges, connecting between two related pages. Web Structure Mining is determined as the techniques of discovering structured data from the Web. This type of mining can be performed either at (intra-page) document level or at (inter-page) hyperlink level The research at the hyperlink level is also called Hyperlink Analysis. Hyperlinks serve two main purposes. Pure Navigation and Point to pages with authority on the same topic of the page containing the link. These can be used to retrieve useful information from the web.

Structure Mining narrows the two major barriers of the web which occurs because of the abundant amount of data available [15]. The two major problems can be defined as following:

Unrelated conclusion of search: distortion occurs for rectified search duly a result of search engines allow for precision method.

Availability of the abundant data: One more issue is indexing of humongous data available on internet. The atop reduction is a functional bit of determining the model below the web hyperlink structure given by the web structure mining[16].

This mining serves usage of link information of one's own website endowing navigation as well as chunk of data into site maps. Hence relevant data can be guaranteed with the use of tokens. Hyperlink command chain is decided to obtain related knowledge with in the sites as a liaison between competitor links and connection by means of third party co-link and search engines[2]. Web Structure Mining also helps in establishing the similar structure of web pages with the help of clustering technique. Whenever there are massive web crawlers there will be more beneficial desired results to the related search [17]. If the web pages are directly linked with

each another or web pages are neighbor we could detect the connection among those pages. The relations may fall in category of ontology, they may have similar contents. Web Structure Mining as well steers to generalize the sequence or networks of hyperlinks in the Websites in some specific domain. This steers to a judgment of flow information in sites and this leads to easy and efficient query processing [18]. In the course of 1997-1998, two ultimately influential hyperlink–based search algorithms Page Rank and Hits were introduced, which are Page Rank Algorithm and HITS, and then improvement of Hits came by adding content information to the links structure and by using exceptional refining. These practices primarily used for calculating the quality rank of each webpage. Hyperlinks primarily act as useful resources in the following areas[18]:

- Exploring the real pages link.
- Suggesting the pages with authority on the similar subjects the page containing the link.

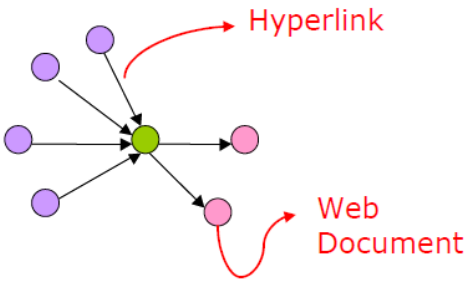

Figure 6: Web Graph Structure [15]

V. WEB MINING APPLICATIONS

There are many of applications which describe the use of web mining strategies in contempt of the association and made-up the analogous technologies did not regard as it as such. Web Mining has become very popular in commercial applications and is very much in demand in specific areas like e-commerce and e-business. The e-commerce and e-business also runs efficiently with the applications like text mining and data mining but web mining is considered to be best among them [19]. Some applications of web mining are given below [20][21]:

E-Commerce

Major test for e-commerce is to grasp the visitors or customer's needs and to value orientations as such as possible.it can improve the quality of service for consumers and competitive benefits. Web Mining generates individual user's profile to understand the needs of users. It checks for fraud. Helps in internet advertising and also provides retrieval of similar images.

Information Retrieval

Search engines on the web use this application of web mining to generate topic hierarchies. Also, it is used to extract schemas for XML documents.

Digital Libraries

Digital libraries services provide precious information distribution to all over world and eliminating the requirement of physically present at several libraries in different parts of globe. Web Mining provides us the privilege to get access to all the different books in different parts of the world at one place without being physically present there.

Network Management

Network Management helps to deliver the content to users reliably in a short duration of time. This is one by traffic management and fault management.

E Business Functional Features

Web mining applications can support online electronic business to improve online marketing, customer support and sales. Web Mining and E-Business correlation for a number of years AI in the form of Data Mining has been used: Mobile phone firms, to stop customer attrition. Financial services firms, for risk management and portfolios. Credit card companies, for identifying fraud and set pricing Mail catalogers, to life their response rates, retailers for market analysis. Business Intelligence itself is major application area of the Web Mining. In this, information on the customer's usage of website is critical information for marketers of E-Tailing the business.

E-Learning

Web mining can be used for improving and enhancing the E-learning environments. In E-learning applications of web mining are usually web usage based. Machine learning techniques and web usage mining enhance web based learning environments.

E –Government

Organizations that interact with the citizen of the country lead to better social services. The main distinctiveness of the e-government systems is related to the use of technology to deliver services electronically, concentrating on the citizen needs by providing better information and enhanced service to support government. This system may provide customized services for citizen, outcomes user satisfaction and quality of services and support in citizen's decision making, which leads to social benefits.

E-Politics and E-Democracy

E-politics provides political information and politics on demands to the citizen results in enhancing political transparency. Election information, parties, members of parliament, members of state governments on the web are part of e-political services. Despite of the importance, e-politics in democracy there are limited web mining strategies to meet citizen needs.

Security and Criminal Investigations

Web mining techniques are also used for securing user system or logs against such cyber-crimes as hacking, internet fraud, fraudulent sites, phishing, illegal online gambling, virus spreading, child pornography and cyber terrorism. Clustering and classification methods of web mining can expose identities of cyber criminals. Neural network, decision trees, genetic algorithm and support vector machines can be used to trace criminal patterns and network visualization on websites.

CONCLUSION

As the web and its usage are continually increasing, increases the chances of analyzing web data and extracting all manner of useful knowledge from it. Since past five years have seen the appearance of web mining as rapidly spreading area, because of the of the research society as well as different organizations that are working on it. In this paper we studied briefly about concept of web mining, its types, discussed various beneficial areas of web mining and applications of web mining.

References

[1] Mr. Dushyant B.Rathod, Dr.Samrat Khanna, "A Review on Emerging Trends of Web Mining and its Applications". ISSN: 2321-9939.

[2] Anmol Kaur, "Comparable Analysis of Web Mining Categories", The International Journal Of Engineering And Science (IJES), Volume 5, Issue 5, PP -27-31 2016, ISSN: 2319 – 1813, ISSN: 2319 – 1805.

[3] B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaiane, ed. Proceedings of "WebKDD2002 –Web Mining for Usage Patterns and User Profiles", Edmonton, CA, 2002.

[4] Saravaiya Viralkumar M. "Web Mining: A Survey on Various Web Page Ranking Algorithms", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 03 Issue: 04| Apr -2016 p-ISSN: 2395-0072. (www.irjet.net)

[5] Aggarwal, B. B. D. S., and Shivangi Dhall, "Web mining: Information and pattern discovery on the world wide web.", International Journal of Science, Technology & Management (2010).

[6] Praveen Kumari "Web Mining - Concept, Classification and Major Research Issues: A Review" Asian J. Adv. Basic Sci.: 2016, 4(2), 41-44 ISSN (Print): 2454 – 7492 ISSN (Online): 2347 – 4114.

[7] Neha Sharma, "A Hand to Hand Taxonomical Survey on Web Mining" International Journal of Computer Applications (0975 – 8887), Volume 60– No.3, December 2012.

[8] IOANA MOISIL, "Advanced AI Techniques for Web Mining" Mathematical Methods, Computational Techniques, Non-LinearSystems, Intelligent Systems.

[9] Simmi Bagga, " Ethos of Web Usage Mining - A Survey" IJA-ERA) ISSN: 2454-2377 Volume – 2, Issue – 1, May – 2016.

[10] S.R.SriAbirami and A.P.Ponselvakumar (2015), "Restructuring the User Search Results Using Feedback Sessions and Evaluation Methods", In Proc. National Conference on Intelligent Computing (NCIC 2015), Pondicherry Engineering College, pp.1240-1246.

[11] Han, J., Kamber, M. "Data Mining:Concepts and Techniques", Second edition p. 628-648.

[12] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2005. "Tapping into the Power of Text Mining. Communications of the ACM - Privacy and Security in highly dynamic systems: Vol. 49 Issue-9.

[13] Govind Murari Upadhyay, " Web Content Mining: Its Techniques and Uses" IJARCSSE Volume 3, Issue 11, November 2013 ISSN: 2277 128

[14] V. Bharanipriya & V. Kamakshi Prasad "Web Content Mining Tools: A Comparative Study"

[15] Raymond Kosala and Hendrik Blockeel," Web Mining Research: A Survey" ACM SIGKDD, July 2000

[16] Yan Wang," Web Mining and Knowledge Discovery of Usage Patterns" CS 748T Project (Part I), February, 2000.

[17] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava," Data Preparation for Mining World Wide Web Browsing Patterns" Supported by NSF Grant, Oct 1998.

[18] Jaideep Srivastava, Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, volume-1 Issue-2 Jan 2000.

[19] S. Yadav, K.Ahmad, J.Shekar , "Analysis of Web Mining Applications and Beneficial Areas" in proceedings of the IIUM Engineering Journal, Volume 12 no. 2 2011.

[20] Navneet Goyal "Web Mining" http://www.slideworld.com/slideshow.aspx/WEB-MINING Prof- Navneet-Goyal-BITS-Pilani-ppt-681474

[21] S.Vidya, K.Banumathy, "Web Mining- Concepts and Application" IJCSIT Vol. 6 (4) , 2015, 3266-3268 ISSN 0975-9646.