International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.3, pp.79-85, June (2017)

Review Paper

E-ISSN: 2320-7639

Web Mining an Approach to Evaluate the Web

AshirrK Kashyap[1*], Iflah Naseem[2], Dheeraj Mandloi [3]

[1*]Computer Science, IET-DAVV, Indore, India
[2]Computer Science, IET, DAVV, Indore, India
[3]Chemistry, IET, DAVV, Indore, India

**Corresponding Author: 247ashkash@gmail.com*

Available online at: www.isroset.org

Received 24th Apr 2017, Revised 07th May 2017, Accepted 20th May 2017, Online 30th Jun 2017

***Abstract*—** Web mining is a field of quite importance a s it include two of the most activated research fields, the data mining and the world wide web. The web mining is an application of data mining which extract the indexed information from the web and sorts them accordingly. In the absence internet, the advancement in technology would not have been possible. The information on the web is so voluminous and heterogeneous that it becomes a necessity to evaluate this available information and present it into a useful form according to the particular problem. Web mining helps in figuring out various patterns obtained in the activities of users and dealing with these interesting patterns developing usable abstracts from various sources. The paper will deal with the various categories of web data mining, its effectiveness in current arena of internet, its pros and cons along with the future it holds in the field of computer science.

***Keywords*—** Web Usage Mining, Web Structure Mining, Web Data Mining, Web Content Mining

I. INTRODUCTION

Internet is evolving at a very rapid pace recently, where WWW (World Wide Web) is its primary key. The World Wide Web is an enormous field of knowledge and information where the URLs act as the analyzers. It is an essential part for the technical development of society. The web being so vast needs to be handled properly and hence needs a computational way of detecting patterns in bulky data sets, which will result in ease in maintaining of data.

A subfield of computer science whose comprehensive goal is to excerpt information of data sets and transform it into coherent structure for future use has been introduced which is known as Data mining. When this technique of sorting and assembling of data based on patterns is applied to web, it is known as Web mining[1].

Web mining is subset of data mining where documents and services are extracted and used naturally. The subtasks of web mining include resource mining which is the task of reclaiming expected web documents, information selection and preprocessing which includes automated selection and prior processing of precise information from the web, generalization of data which automatically discovers the usual patterns at web sites individually and across the multiple sites and finally analysis of information obtained which is affirmation and interpretation of quarried patterns.

Web mining is further subcategorized namely into the web content mining, the web usage mining and the web structure mining. Here the web content mining is the process of extracting necessary data components of web document. Web structure mining is the course of evolving structured data from the web, this category can be further divided into hyperlinks and structured documents. The Web usage mining is another application where a better understanding is established and web based applications could be properly served, this is again divided into three, the web server data, the web application server data and the web application level data. There exists a system where data could be filtered known as WebSIFT.

Everything comes with pros and cones so is in the case of web data mining. It has been proved useful by helping government agencies against crimes online and creating awareness but at the same time privacy threatened by it remains an area of real concern and acts as major negative aspect of web mining. The efforts are being made to devise possibly effective solutions for the same. The web mining happens to be fast growing field of computer science, and has a wide range of opportunities in the field of research and development[2].

II. DATA MINING

Data mining is a multidisciplinary subfield of studies of computers. It is the computational way of detecting patterns in hefty data sets I involving methods at the crossings of artificial intelligence, statistics database systems and machine learning. The comprehensive goal of the data

mining process is to excerpt information from a data set and mutate it into a coherent structure for further use. Aside from the coarse analysis step, it involves database and data management facets, data pre-processing, model and inference considerations, alluring metrics, intricacy of considerations, post-processing of identified structures online updating and visualization[3][4].

III. THE WORLD WIDE WEB

The World Wide Web (WWW) is an information field where documents and Uniform Resource Locators commonly called as URLs analyze web resources, hypertext links intertwine them, and can be approached through the Internet. The World Wide Web has been essential to the advancement of the Information Age and is the elementary tool use by billions of people to interact via Internet. Web pages are mainly text documents annotated and formatted with HTML (Hypertext Markup Language). In extension to formatted text, web pages may contain video, audio, images and software fragments that are provided in the web browser of the users as consistent pages of multimedia content. Ingrained hyperlinks grant users to navigate the web pages. Multiple web pages with a universal domain name, a common theme or both, builds a website. Website content can majorly be provided by the publisher or interactive where user's grants matter or the content relies on the customers or their response. Websites may be predominantly informative, mainly for entertainment, or primarily for governmental, or non-governmental or commercial organizational purposes.

IV. THE WEB OF DATA MINING

Web is a set of files on one or many Web servers which are inter-related. Web mining is an extension of data mining. Data mining techniques are used by web mining to naturally discover and extract data from Web services and documents. This area of research is so enormous presently slightly due to the interests of diverse research communities, the prodigious growth of information sources available on the Web and the newly established keenness in e-commerce. This phenomenon partially initiates turmoil when questioned about the consistency of Web mining and when correlating research in this field. Broad distribution of web mining's sub tasks is elaborated as follows:

A. Resource finding:

It is the task of reclaiming expected Web documents. Resource finding signifies the task of retrieving the information which is online or offline using the text sources accessible on the Web comprising electronic newswire or the newsgroups, electronic newsletters, the text contents of HTML documents procured by eliminating HTML tags, and also the picking manually of Web resources. It comprises of text sources that formerly were not available from the Web but are now present, for example online texts made exclusively for research purposes, text databases and much more.

B. Information selection and pre-processing:

It is the automated selection and prior processing of precise information from Web resources which are retrieved. The selection of information and prior-processing step is a type of conversion process of the actual data obtained in the IR process. These conversions could be a type of pre-processing such as eliminating stop words, stemming, etc. or a pre-processing focused at achieving the aimed portrayal such as searching phrases in the training corpus, conversion of the representation to relational or first order logic form, etc.

C. Generalization:

It automatically finds usual patterns at Web sites individually and also the across multiple sites. Data mining or Machine learning techniques are used in a typical way for the generalization. Humans play a vital role in the discovery process of information or knowledge on the Web since the Web is bilateral medium.

D. Analysis:

It is the affirmation or interpretation of the quarried patterns. Collective query-triggered information discovery is critical as the more automated data-triggered information discovery. However, we ostracize the data discovery done manually [5].

V. THE WEB TAXONOMY

Web mining can be typically divided into three categories, in accordance to the kinds of data to be mined.
These are mainly web content mining, web usage mining and web structure mining, which are further divided into sub categories.
Web content mining illustrates the discovery of appropriate data from the Web contents or web data or web documents. The contents of the Web could encompass quite large range of information. Formerly the Internet comprises of various types of services and data sources such as Usenet, Gopher and FTP. In present scenario, most of the data is either transferred to available from the Web. In the last few years the advancement in the amount of government data has been astounding. There exist Digital Libraries that are also available from the Web. Also quite a few companies are changing their services and business electronically. As a consequence many of the company databases that earlier resided in the legacy systems are being transferred to or made available from the Web. Hence the company database

can now be accessed by partners, employees and sometimes even the customers.

One of the consequences of the transformation is the presence of Web applications, via web interfaces users are enable to use these web applications. Numerous applications and systems are being moved to the Web and many kinds of applications are evolving in the Web environment. Although specific contents of web data remain hidden, this cannot be indexed. Dynamically as a result of queries, this data is either generated and remains stored in the DBMSs or is of private form. Already the Web contains diverse types of data such as audio, images, textual contents, metadata, video and even hyperlinks too. Another term coined for the recent research on multimedia and mining of data is multimedia data mining. Hence multimedia data mining can be called as an instance of web content mining.

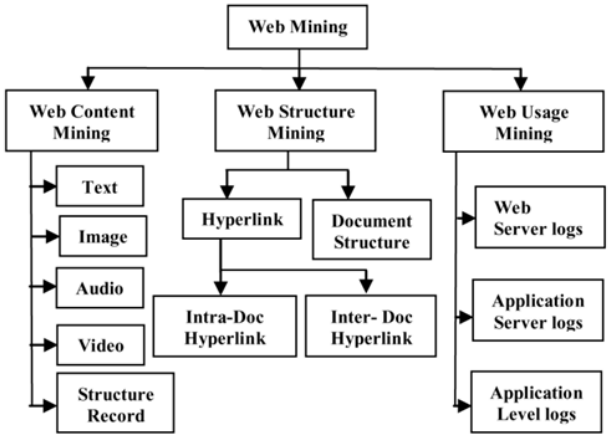

Fig. 1 Web Mining Taxonomy

A. Web Content Mining

The process of extracting necessary data from the components of web documents is termed as web content mining. Content data is the assemblage of details a web page is ought to contain. It comprises of structured records such as lists and tables and text, images, video or audio. Arguments described in text mining consist of topic diagnosis and tracking, deducing clique patterns, chunking of documents from web and regulation of web pages. Research activities in this reign rely on techniques advanced in other disciplines such as Information Retrieval and Natural Language Processing (commonly called as IR and NLP). There exists an important body of work in excerpting data from images in the reign of computer vision and image processing, in web content mining the application of these techniques is limited.

Web content mining characterizes the automated search of data resource which is available online, and comprises mining of web data contents. In the domain of Web mining, the Web content mining particularly is an analogue of data mining techniques for databases and their relations, since there exist the probability of discovering similar types of data from the unstructured data lodged in Web documents. The Web document comprises of mostly many types of data, including text, image, audio, video, metadata and hyperlinks. Among them few are semi-structured like HTML documents, and some are more structured data such as the information in the tables or database generated HTML pages, still maximum of the data is unstructured text data. The unstructured aspects of Web data influence the Web content mining in more complicated direction of approach. In case of the semi-structured data, all the works use the HTML structures in documents and a few used the hyperlink structure within the documents for document representation. According to the database view, in order to have the improved data management and querying on the Web, the mining always make efforts to derive the structure of the Web site leading the Web site to transform into database. Multimedia data mining is a component of the content mining, which is committed to mine the remarkable data and information from huge online multimedia sources. Multimedia data mining on the Web remains area of keen interest for researchers. Working in direction of consolidating framework for representation, problem settling, and learning from multimedia is a real challenge.

B. Web Structure Mining

The architecture of a regular web graph comprises of web page sand hyperlinks such that web pages are nodes and hyperlink are edges which provides connectivity between the nodes that is the web pages. Web structure mining is the course of evolving structure data from the web. This can be again divided into domains namely hyperlinks and documented structure.

Maximum of the Web data restoration tools use only the information which is textual, ignoring the linked information which could be quite valuable. The main objective of Web structure mining is to create structural summary regarding the Web site and Web page. The Web content mining focuses precisely on the structure of inner-document, while Web structure mining make efforts to find the link structure of the hyperlinks at the level of inter-document. In accordance of the hyperlinks, Web structure mining will segregate the Web pages and set up the data, like the resemblance and relationship of various Web sites.

Web structure mining can also have another perspective of devising the structure of Web document by its own. This form of structure mining can be utilized to declare the structure or schemes of Web pages, this would be valuable for navigational purposes and make it viable to correlate and integrate Web page schemes. This form of structure mining will aid introduction of database techniques for accessing data in Web pages by facilitating reference schemes.

The structural information developed using the Web structure mining includes the data measuring the frequency of Web fragments in a Web table that consists of links that are global and the links that stretch through different Web sites.

When Web pages are linked to other Web page directly or are neighbors, then it would be likely to discover relationships among those web pages. They may be related by synonyms or ontology, their content may be similar, both may reside on the same Web server hence developed by the same person. Another assignment of Web structure mining is to find the essence of the hierarchy or network comprising hyperlinks in the Web sites of an appropriate domain. This may lead to generalize the stream of knowledge or data in Web sites that may epitomize some peculiar domain; hence the inquiry processing will be straightforward and more productive. Since the Web structure mining has relations with the Web content mining, therefore it is quite reasonable that the Web document comprises of links, and they both utilize the actual or elementary information on the Web. It's very usual to associate these two mining tasks in the application[6].

1) Hyperlinks

The structural unit that links a location in a particular web page to a various locations, may be within the same web page or on another web page is termed as hyperlink. An intra-document hyperlink is that hyperlink that connects to a various part of the same page; where as an inter-document hyperlink is that hyperlink that connects two separate pages. The document comprising a hyperlink is termed as its source document.

2) Documented Structure

The content in a Web page can also be standardized in a tree structured format, based on the various HTML and XML tags in the page.

C. Web Usage Mining

Web usage mining is another application of data mining techniques to find curious usage patterns from data of web usage, so that a better understanding could be established and web based applications could be properly served. Usage data finds the identity or root of web users also their browsing attitude towards a web site. Web usage mining itself can be divided further based on the type of usage data considered. The web server data, the application server data and the application level data are the categories of web usage mining. Web usage mining tries to devise the appropriate data from the secondary data derived from the communication of the users with web while surfing on it. It targets the techniques that could anticipate user practices while the user interacts with Web.

There are no distinct difference between the Web usage mining and rest of the categories of web mining. During the data formation of Web usage mining, the Web content and Web site are used as the knowledge source which further associates with Web usage mining along the Web content mining and Web structure mining. Also the rounding up in the process of pattern discovery is a link to Web content and structure mining via usage mining. The work which has been done in the IR or Database or Intelligent Agents and Topology lays a base for the Web content mining.

1) Web Server Data

In this category of web usage mining, user logs are gathered by the web server and precisely include IP address, page reference and access time.

2) Application Server Data

Application server data is another aspect of web usage mining where commercial application servers like Weblogic, have symbolic features to facilitate E-commerce applications to be constructed on top of them with meager effort. Another feature of this domain is the ability to track different types of business affairs and hence logging them in application server logs.

3) Application Level Data

Recent types of events can be characterized in an application, and further logging can be switched on for them. This generates history of the events. But many end applications needs a sequence of one or more of the techniques used in the described categories.

4) WebSIFT: The Web Site Information System

Another framework of Web usage mining is The Web Site Information Filter System, which utilizes the components and structure data from a Web site, and finally determines the interesting outcomes from mining usage data. The WebSIFT system is fashioned to execute usage mining from the server logs in the protracted NSCA format. The pre-operational algorithms comprises of diagnosing users, server sessions, and surmise cached page references via the usage of the cited reign. Besides formation of the server sessions, WebSIFT system also exhibit content and structure preprocessing, and facilitate the option to transform into server sessions into episodes. The server session or episode files can be carried out via sequential pattern analysis, union rule discovery, rounding up or usual statistics algorithms.

The WebSIFT system is associated with the WEBMINER prototype and fragments the Web usage mining process into three primary parts that corresponds to the various phases of usage. The input of the mining process comprises of three server logs – the access server log, referrer server log, and agent server log. The HTML files that build up the site; and

the alternative data like registration files and remote agent logs.

The WebSIFT system has been enforced using a relation database, procedural SQL, and Java language where Java Database Connectivity (JDBC) drivers are used to interact with the database[7].

VI. PRIVICY ON THE WEB

Due to the enormous growth of the e-commerce, privacy turn into a sensitive topic and invited quite a large attention recently. The primary goal of Web mining is to excerpt data from data set for business use, which regulate its application, is profoundly customer-related. There happens to be unavoidable struggle between the Web user and the administrator according to the perspective of privacy. The administrator's opinion, quite a few uses of data mining are innocuous, like the data analysis to detect undisclosed style of patterns to grant supermarkets to organize items
in ways that will inspire customers to buy more of specific products. But from individual perspective, many users trust that some of the applications of Web mining, may be questioned when privacy is concern, like junk mails stuck mail account or personal data divulged in course of online shopping. Recently, the matter of privacy has become the most demanding involvement not only for the Web user but for the e-commerce developers too.
The inadequacy of regulations in the usage and deployment of Web mining systems along with the extensively spread privacy exploitation reports associated with data mining has made privacy a major issue[8].

VII. PROS AND CONS OF WEB MINING

A. Pros

Web usage mining necessarily has many assets which makes this technology inviting to corporations comprising the government agencies. This technology has permitted e-commerce to do personalized marketing, this results in much higher trade volumes. Government agencies are utilizing this technology to distinguish threats and combat against terrorism.

The predicting features of mining applications can be profitable for the society by recognizing criminal activities. The companies can facilitate better customer relationship by giving them precisely what they want. Companies can figure out the needs of the customer in a better manner and they can reply to customer needs quickly. The companies can discover, attract and maintain customers; they can save on management costs by using the gathered insight of customer requisites. They can boost profitability by targeting pricing based on the profiles made. They can even discover the customer who may default to an opponent the company will try to keep the customer by granting promotional offers to the precise customer, thus decreasing the risk of losing a customer.

B. Cons

Web usage mining by its own does not devise issues, but the technology used on data of personal format may cause concerns. The most condemned ethical issue including web usage mining is the intrusion of privacy. Privacy is considered vanished when data concerning an individual is fetched, utilized, or disseminated, particularly if this occurs without their consent. The obtained information will be checked, and gathered to form profiles; the information will be made anonymous prior to grouping so that personal profiles are absent. Hence these applications make the users ordinary by making judgments based on their mouse clicks. Generalizing can be defined as a capability of judging and handling people on the grounds of group of characteristics instead of on their separate individual characteristics.

Another necessary concern is that the companies gathering the information for a definite purpose may utilize the information for a completely different objective, and this substantially violates the user's interests.

The expanding tendency of selling personal information as a commodity inspires website owners to trade personal data collected from their site. This trend has heightened the amount of information being apprehended and traded rising the chances of the invasion of privacy of an individual. The companies which buy the information are bound to make it unidentified and these companies are considered authors of a precise clemency of mining patterns

C. Arguments in Defense of web Mining

All the assets exhibit that web-data mining is a highly important technique, which has been developed and practiced on an enormous and growing scale. Though the threats to a few crucial values tend to be rather serious, and may create stress in the web data mining field. Many professionals handling web-data mining in a business context do not acknowledge any moral threats in web-data mining. To achieve some insight into recent web-data mining practices and the behavior of web data miners to the ethical issues involved, the professionals were interviewed.

D. Possible Solutions

There are several ways to solve some issues with respect to privacy in the ethical context of web-data mining. Solutions can be further categorized into individual and at a collective level.
The solutions at an individual level, refers to actions an individual can take in order to protect himself against

possible threats. Some of the possible solution at individual level includes making use of privacy enhancing technologies (PETs), being aware while providing data online which is personal, and evaluating privacy policies on web sites.
The solutions at a collective level include things that can be done by society to prevent web-data mining from resulting in any harm. Some solutions at collective level includes further development of PETs, publishing privacy policies, auditing web mining activities, legal measures, establishing awareness among individuals and web data miners. A combination of technical and non-technical solutions at all of the levels, individual and collective is probably essential to fight against the problems[9].

VIII. THE FUTURE OF WEB DATA MINING

There is an imminent phase of 'irrational despair' succeeded by a phase of 'irrational exuberance' in the commercial capability of the Web, the usage and acceptance of the Web continues to emerge as a persistent reign. This trend is expected to continue similar to Web services continuing to grow. As the Web and its usage advance, it will continue to raise even more content, usage data and structure and the rate of Web mining will keep expanding.
Temporal progression of the Web Society's communication with the Web is transforming the Web along with the manner the society interacts. Accumulating the history these interactions in a place is undoubtedly too astounding a task. The modifications to the Web are being stocked by the pioneering Internet Archive projects. Research obligated to be carried out in obtaining temporal progression models of how various fields like Web content, Web communities, Web structures, authorities and hubs are emerging. Large organizations usually get usage data from the Web sites. Along these origins of data accessible, there is a huge opportunity of research to establish techniques for evaluating the evolution of web over the course time
The optimization of Web services is experienced as services over the Web continue to evolve; there will be requirement to compose them into robust, scalable, efficient, etc. Web mining can be practiced to get better understanding of the behavior of these services, and the information extracted can be used for several kinds of optimizations[10].

IX. CONCLUSION

We are in the world where technology plays a vital role and with advancement of technology, the internet and World Wide Web are emerging as primary key. Being primary key, which has enormous depths the research in the field of their patterns across the web, comes out to be a necessity. Web data mining proves to be of utmost importance since it helps business organizations in marketing, it also helps in predictive analysis for the future use. Despite of having a so many applications, there a few drawbacks too, including privacy threats and illegal use of information. Sincere efforts are being made to overcome the drawbacks sustained in web data mining since it holds a useful and a valuable reign in near future.

ACKNOWLEDGMENT

We have made our best efforts to present this research as simple as possible utilizing basic terms that we hope will be comprehended by the widest spectrum of researchers, analysts and students for further studies.

We have completed this study under the able guidance and supervision of Dr. Dheeraj Mandloi .We would like to thank esteemed scholarly, assistance and knowledge we have received from him towards fruitful and timely completion of this research paper.

We would also like to thank our Institute of Engineering and Technology (IET DAVV) and faculties who provided us with ne.

REFERENCES

[1] R Kosala, H Blockeel, "*Web mining Research: A survey*", ACM SIKGDD, Vol.2, Issue 1, pp. 1-2, 2000.

[2] D. Jayalatchumy, P.Thambidurai, "*Web Mining Research Issues and Future Directions – A Survey*", IOSR Journal of Computer Engineering (IOSR-JCE), Vol.4, Issue.3, pp.20-27, 2013.

[3] V. Krishnaiah, G. Narsimha, N.S. Chandra, "*Survey of Classification Techniques in Data Mining*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.9, pp.65-74, 2014.

[4] N Jain, V Srivastava, "*Data mining techniques: a survey paper"*, IJRET: International Journal of Research in Engineering and Technology, Vol.2 ,Issue.11, pp 116-119, Nov-2013.

[5] P Mehtaa, B Parekh, K Modi, P. Solanki, "*Web Personalization Using Web Mining: Concept and Research Issue",* International Journal of Information and Education Technology, Vol.2, Issue.5, pp.1-7, 2012.

[6] R. Munilatha1 , K. Venkataramana, "*A Study on issues and techniques of web mining*", International Journal of Computer Science and Mobile Computing, Vol.3, Issue.5, pp.331-341, 2014.

[7] J Srivastava, P Desikan, V Kumar, *"Web Mining—Concepts,Applications, and Research Directions", Foundations and advances in data mining*, Berlin, pp.275-307, 2003.

[8] Siddu P. Algur, Prashant Bhat, "*Abnormal Web Video Prediction Using RT and J48 Classification Techniques*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.6, pp.101-107, 2016.

[9] M. Aldekhail, "*Application and Significance of Web Usage Mining in the 21st Century: A Literature Review*", International Journal of Computer Theory and Engineering, Vol. 8, No. 1, pp.1-6, 2016.

[10] Rajesh Shah and Suresh Jain, "*Web Mining Using Cloud Computing Technology*", International Journal of Scientific Research in Computer Science and Engineering, Vol.3, Issue.2, pp.21-25, 2015.

Authors Profile

Mr Ashirr K Kashyap is pursuing his Bachelor of Engineering in Computer Science from Institute of Engeneering and Technology ,Devi Ahilya Vishwa Vidyalaya ,Indore. His research interests include Network Security,Web Security,Dark Web,Web–data mining and IOT.

Miss Iflah Naseem is pursuing her Bachelor of Engineering in Computer Science from Institute of Engeneering and Technology ,Devi Ahilya Vishwa Vidyalaya ,Indore. Her research interests Web–data mining and IOT.

Dr. D Mandloi pursed Phd from Devi Ahilya Vishwa Vidyalaya ,Indore in 2004. He has nearly 14 years teaching experience. Teaching Engineering Chemistry, Environmental Chemistry, Material Science and Engg,Analytical Instrumentation Techniques.His research interests include Engineering Chemistry, Environmental Engineering Computational Chemistry.

