International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 12, December - 2013

A Survey on Web Mining Taxonomy

K. NETHRA[1]
Post graduate student, ME(CSE)
Sri Ramakrishna Engineering College,
Coimbatore

UMASRI M. L[2],
Post graduate student, ME(CSE)
Sri Ramakrishna Engineering College,
Coimbatore.

B.SHARMILA
Assistant professor
Sri Krishna Arts and science College,
Coimbatore

ABSTRACT

The world wide web has been grown explicitly and its transformation provides many opportunities. Web mining is the application of data mining. Web mining is the integration of information gathered by traditional data mining methodologies and technique with the information gathered over www . Based on the gathered information web mining is categorized into three –Web Content Mining ,Web Structure mining ,web usage mining. In this paper we have discussed about web mining introduction (with its problems, current research and types), its terminologies, application , algorithms used and Advantage and disadvantage of Web mining.

I. INTRODUCTION

The *World Wide Web* is a rich source of voluminous and heterogeneous information and it continues to expand in size and complexity . It provides access to all people at any place and at any time. By this facility any one can upload or download relevant data's. So that valuable content in the web site can be used for both a complementary. Because of the Web data's are unstructured and semi-structured , lots of insignificant and irrelevant document are obtained as a result after navigating several links . So that data mining cannot applied directly. For effective retrieval of web information , web mining is used .

Web mining - is the application of data mining techniques to automatically discover and to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. Some of the data mining technique applied in web mining are association rule mining, clustering , classification , frequent item set .Some of the sub task of web mining are resource finding , information selection and preprocessing ,generalization and analysis.

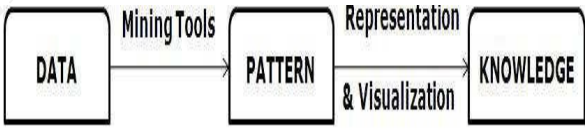

Fig.1 Process of Web Mining

Basic idea of web mining is to assist user or site owner in finding relevant information.

According to user they focus web mining for

- Discovery of documents on a specific subject.
- Discovery of semantically related documents or document segments.
- Extraction of relevant knowledge from multiple sources.
- Knowledge or information filtering.

According to Owner they focus web mining for

- Increasing content or conversion efficiency.

- Targeted promotion of goods, services and ads.
- Measuring effectiveness of site content structure.
- Providing dynamic personalized services or content.

Some of the problems when interacting with the web are finding relevant information , Creating new knowledge out of the information available on the web , Personalization of the information , Learning about the users.

Current research work in web mining are Ranking metrics - for page quality and relevance, Robot Detection and Filtering - Separating human and nonhuman Web behavior, Information scent - Applying foraging theory to browsing behavior , User profiles - Understanding how users behave, Interestingness measures - When multiple sources provide conflicting evidence , Pre-processing - making Web data suitable for mining, Identifying Web Communities of information sources, Online Bibiliometrics , Visualization of the World Wide Web , Fraud and threat analysis ,Counter Terrorism,Semantic Web mining

The techniques which are helpful in extracting data present on the web is an interesting area of research. Web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

1.1)Web Usage Mining

Web usage mining is the type of data mining process for discovering the usage patterns from web information for the purpose of understanding and better provide the requirements of web-based applications[5]. Web usage mining imitates the actions of humans as they interact with the Internet. Web usage mining is the process of extracting useful information from server logs i.e. users history.

Some of the problems while using web log are

- Identification of user is difficult.
- Post data not recorded , cookie data stored elsewhere.
- Web content may be dynamic.
- Use of spider and automated agents

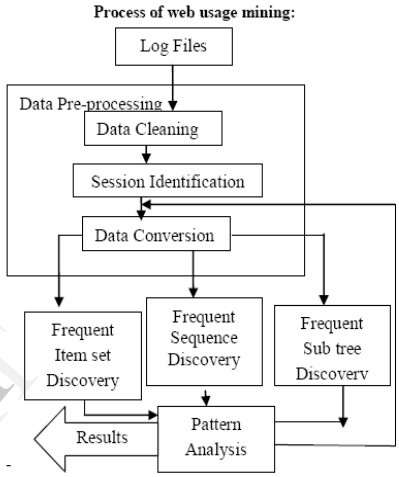

Fig 2.Process of Web usage Mining

1.2) Web Structure Mining

Web structure mining is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages[2]. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon.

Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information.

- The first of these problems is irrelevant search results.

- The second of these problems is the inability to index the vast amount if information provided on the Web.

This minimization comes in part with the function of discovering the model underlying the Web hyperlink structure provided by Web structure mining.

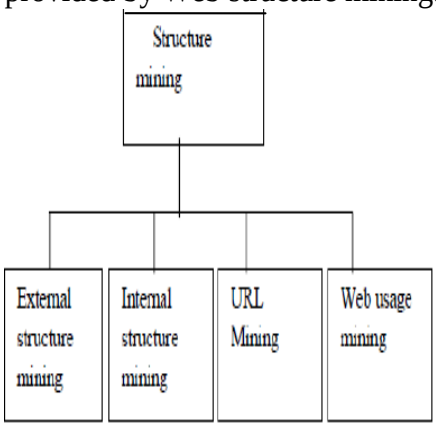

Fig 3 Types of Web Structure Mining

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site 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. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links. This enables clustering of connected Web pages to establish the relationship of these pages[3].

With improved navigation of Web pages on business Web sites, connecting the requested information to a search engine becomes more effective.Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided a into two kinds: 1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location. 2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

1.3) Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page contents. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query[4]. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query.

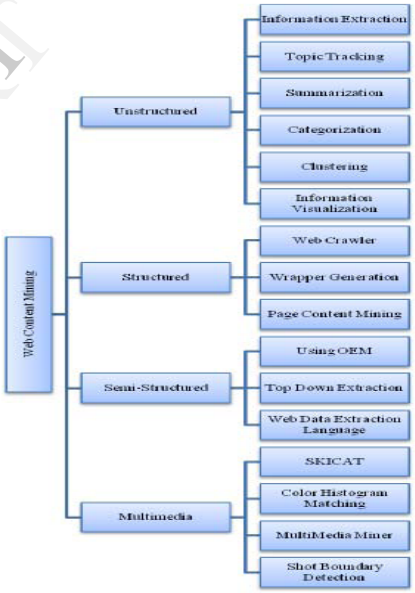

Fig 4 **Web Content Mining Techniques**

With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

In this paper , we focus on web mining and discussed the taxonomy of web mining.The rest of the paper is organized as follows.Section II introduces web mining terminologies which are used in web data extraction.Section III explains the application of web mining.Section IV suggests the advantages and disadvantages of web mining . Section V Finally , the conclusion are made .

II. WEB MINING TERMINOLOGIES

2.1 **DEEP WEB**

The deep Web (also called Deepnet, the invisible Web, dark Web or the hidden Web) refers to World Wide Web content that is not part of the surface Web, which is indexed by standard search engines.

2.2WEB DATA EXTRACTOR

A powerful web data / link extractor extracts URL, Meta tag (title, description, and keyword), body text, email, phone, and fax from web site, search results or list of URLs . High speed, multi-threaded, accurate extraction - directly saves data to the disk file. Program has numerous filters to restrict session, like - URL filter, date modified, file size, etc. It allows user-selectable recursion levels, retrieval threads, timeout, proxy support and many other options.

In analogy to search engines over the "crawlable" web, we argue that one way to unlock the Deep Web is to employ a fully automated approach to extracting, indexing, and searching the query-related information-rich regions from dynamic web pages. Extracting the interesting information from a Deep Web site requires many things: including scalable and robust methods for analyzing dynamic web pages of a given web site, discovering and locating the query-related information-rich content regions, and extracting itemized objects within each region. By full automation, we mean that the extraction algorithms should be designed independently of the presentation features or specific content of the web pages, such as the specific ways in which the query-related information is laid out or the specific locations where the navigational links and advertisement information are placed in the web pages.

- Data Collection
- Data Extraction

The Deep Web (or Hidden Web) comprises all information that resides in autonomous databases behind portals and information providers' web front-ends. Web pages in the Deep Web are dynamically-generated in response to a query through a web site's search form and often contain rich content.

Data Extraction from Web Extractors Web even those web sites with some static links that are "crawlable" by a search engine often have much more information available only through a query interface. Unlocking this vast deep web content presents a major research challenge.

2.3 WEB CRAWLER

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, Web spiders, Web robots, or—especially in the FOAF community—Web scutters. This process is called Web crawling or spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date

data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for sending spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. The large volume implies that the crawler can only download a fraction of the Web pages within a given time, so it needs to prioritize its downloads.

The number of possible crawlable URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all of which may be linked on the site. This mathematical combination creates a problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content. A crawler must carefully choose at each step which pages to visit next.

The behavior of a Web crawler is the outcome of a combination of policies:

- a selection policy that states which pages to download,
- a re-visit policy that states when to check for changes to the pages,
- a politeness policy that states how to avoid overloading Web sites, and
- a parallelization policy that states how to coordinate distributed Web crawlers.

2.4 **Wrapper Generation:**

In Wrapper Generation, it provides information on the capability of sources. The sources are what query they will answer and the output types. The wrappers will also provide a variety of Meta information. E.g. Domains, statistics, index look up about the sources

III.WEB MINING APPLICATIONS

Web mining extends analysis much further by combining other corporate information with Web traffic data. Web mining tools can be extended and programmed to answer almost any question.

It can be applied in following areas:

1. Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly.
2. The company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.
3. In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites.

4. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing
5. The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement.
6. Search engine Google provides advanced and efficient searching capabilities.

IV. WEB MINING ALGORITHMS

Some of the algorithms used in web mining are

Type	Algorthim
1.Web structure minning	a.Link analysis algorithms[1] b. Page rank c. Weighted page rank d. Hits (hyper-link induced topic search) e. Topic sensitive pagerank algorithm
2. Web usage minning	a.Clustering b.k-mean algorithm c.Latent semantic analysis d. Prefix Span Algorithm e. One pass SI and one pass AIISI Algorithm
3. Web content minning	a.Correlation algorithm for relevance ranking b.Cluster hierarchy construction algorithm(chca) c. Weighted Page Content Rank d. fuzzy c-mean (fcm) algorithm

Table with different Web mining Algorthim

V.ADVANTAGE AND DISADVANTAGE OF WEB MINING

5.1 ADVANTAGE:

Web usage mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled e-commerce to do personalized marketing, which eventually results in higher trade volumes. Government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of mining applications can benefit society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

5.2 DISADVANTAGE

Web usage mining by itself does not create issues, but this technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. The obtained data will be analyzed,

and clustered to form profiles; the data will be made anonymous before clustering so that there are no personal profiles. Thus these applications de-individualize the users by judging them by their mouse clicks. De-individualization, can be defined as a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits.

Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user's interests.

The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site.The companies which buy the data are obliged make it anonymous and these companies are considered authors of any specific release of mining patterns. They are legally responsible for the contents of the release; any inaccuracies in the release will result in serious lawsuits, but there is no law preventing them from trading the data.

Some mining algorithms might use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes. This process could result in denial of service or a privilege to an individual based on his race, religion or sexual orientation, right now this situation can be avoided by the high ethical standards maintained by the data mining company. The collected data is being made anonymous so that, the obtained data and the obtained patterns cannot be traced back to an individual. It might look as if this poses no threat to one's privacy, actually many extra information can be inferred by the application by combining two separate unscrupulous data from the user.

V.CONCLUSION

World wide web is very important to carry out our day to day activities like business, education, e-commerce etc. In this paper, we survey the introduction of web mining (with its problems, current research and types), its terminologies, application , algorithms used and Advantage and disadvantage of Web mining . So it is helpful for storing and processing in the web pages.

REFERENCE

1. gurpreet kaur and shruti aggarwal " a survey- link algorithm for web mining"

2. Preeti Chopra, Md. Ataullah " A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms"in International Journal of Engineering and Advanced Technology (IJEAT)

3. P. Ravi Kumar and Ashutosh Kumar Singh "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval" in American Journal of Applied Sciences

4. Rahul Neve, K.P Adhiya "Comparative Study of Web Mining Algorithms for Web Page Prediction in Recommendation System" in International Journal of Advanced Research in Computer and Communication Engineering

5. J Vellingiri, S.Chenthur Pandian ," A Survey on Web Usage Mining" in Global Journal of Computer Science and Technology