International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 3 Issue: 1 13-Aug-2014,ISSN_NO: 2321-3337

A Survey of Web Content Mining Tools and Future Aspects

Mrs.C.Menaka M.C.A.,M.Phil.,[1] Dr.N.Nagadeepa M.Sc., M.Phil.,M.C.A.,Ph.D[2]
Research Scholar, Department of Computer Applications, Bharathiar University,Tamilnadu,India[1].
Professor,Department of Computer Applications, V.S.B Engineering College,Tamilnadu,India[2].

ABSTRACT *−As the data on the web is continuously increasing day by day so, web mining become necessary to draw an inference from the huge data available on web. In web mining non trivial pattern and useful information are retrieved from the web data. Web mining consists of three types namely Web usage mining, Web content mining and Web structure mining.Today, they are several billions of HTML documents, pictures and another multimedia files available on the Internet. There is a need of methods to help us extract information from the content of web pages. One answer to this problem is using the data mining techniques that is known as web content mining, which is defined as "the process of extracting useful information from the text, images and other forms of content that make up the pages". Web mining implies the application of data mining techniques to extract knowledge from Web content, structure, and usage - is the collection of technologies to fulfill this potential. Interest in Web mining has grown rapidly in its short existence, both in the research and practitioner communities. This paper provides a brief overview of web mining and various content mining tools and the accomplishments of the field - both in terms of technologies and applications.*

Keywords: Web mining, Web content mining,Types of web mining, Content mining tools,.

I. INTRODUCTION

Web mining is the implementation of data mining techniques to extract knowledge from Web data - including Web documents, hyperlinks between documents, usage logs of web sites etc., Internet has became an indispensable part of our lives now a days so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques helps to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process (with or without other types of Web)[2]. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. With the extensive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-business[1]. A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behaviour. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future[2][3].

A. Web Mining Process

The complete process of extracting knowledge from Web data is follows in Fig.1:

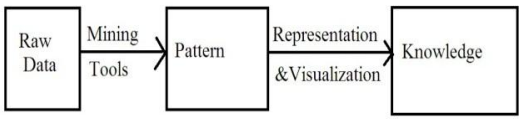

Fig.1: Web Mining Process

The various steps are explained as follows.

1.Resource finding:

It is the task of retrieving intended web documents.

2.Information selection and pre-processing:

Automatically selecting and pre- processing specific from information retrieved Web resources.

3.Generalization:

Automatically discovers general patterns at individual Web site as well as multiple sites.

4. Analysis:

Validation and interpretation of the mined patterns.

II. CATEGORIES OF WEB MINING

Web mining has three operations of interests - clustering (finding natural groupings of users, pages etc.), associations (which URLs tend to be requested together), and sequential analysis (the order in which URLs tend to be accessed). As in most real-world problems, the clusters and associations in Web mining do not have crisp boundaries and often overlap considerably. In addition, bad exemplars (outliers) and incomplete data can easily occur in the data set, due to a wide variety of reasons inherent to web browsing and logging. Thus, Web Mining and Personalization requires modeling of an unknown number of overlapping sets in the presence of significant noise and outliers, (i.e. bad exemplars). Moreover, the data sets in Web Mining are extremely large. **Web mining -** is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are **Web usage mining**, **Web content mining** and **Web structure mining[1][6]**.

A.Web Usage mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications[6]. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site[5]. Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data: correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users[6].

Application Server Data: Commercial application servers, e.g. Weblogic [BEA], BroadVision [BV], StoryServer [VIGN], etc. have significant features in the framework in order to enable E-commerce applications to be built on top of them with little effort. A key aspect is the ability to trace the different kinds of business events and that can be logged into application server logs.

Application Level Data: New kinds of events can always be defined in an application, and logging can be turned on for them – generating histories of these specially defined events.

The usage data can also be split into three different kinds on the basis of the source of its collection: on the server side, the client side, and the proxy side.

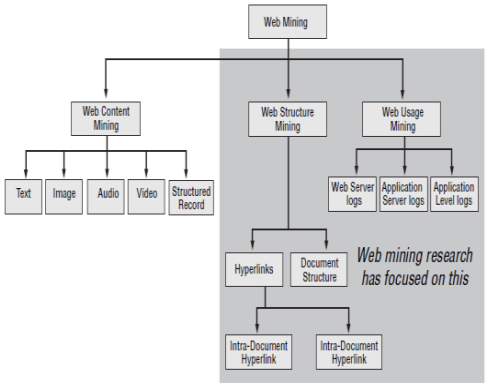

Fig.2 Taxonomy of web mining

The key issue is that on the server side there is an aggregate picture of the usage of a service by all users, while on the client side there is complete picture of usage of all services by a particular client, with the proxy side being somewhere in the middle.

B.Web structure mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. Based upon the kind of structural data used structure mining can be classified into two types:

Hyperlinks:

A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis, of which provides an up-to-date survey[6][8].

Document Structure:

In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structure out of documents[5][8].

C.Web content mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched[8].

Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

III. WEB CONTENT MINING TOOLS

Web Content Mining is mining data from the content of web pages. Web pages consist of text, graphics, tables, data blocks and data records. Web Content Mining uses the ideas and principles of data mining and knowledge discovery process. Using the Web for providing information is more complex than when working with static databases, due to Web dynamics and the large number of documents. Many researches have been made to cover web content mining problems to improve the way that pages are presented to end users, improving the quality of search results and extract interesting content pages. Required information and data can be derived from the eb, with the help of the following content mining tools. This following various tools can help us to download the essential information that one would require. They collect appropriate and perfectly fitting information. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and Web Content Extractor.

A.Automation Anywhere 6.1 (AA)[11] :

AA is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining[11].

Aspects:

- Automated tasks can be created in a quick manner by recording keyboard and mouse operations or use of point and click wizard.
- Unique SMART Automation Technology for fast automation of complex tasks.
- Web record and Web data extraction.
- More than 300 plus actions are included: Internet, conditional, loop, prompt, file management, database and system, automatic email notifications, task chaining, etc.

B.Screen-scaper[10]:

Screen-scraping is a tool for retrieving and extracting/mining information from web sites. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements[10]. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper.

Aspects:

- Graphical interface allowing the user to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data.
- Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked.
- One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet.

- A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time, after which the results are displayed in a single interface.

C. Mozenda[12] :

This tool enables users to extract and manage Web data. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, and mashup the data to be used in other applications or as intelligence[12].
There are two parts of Mozenda's scaper tool:

Mozenda Web Console:

It is a Web application that allows user to run agents, view & organize results, and export publish data extracted.

Agent Builder:

It is a Windows application used to build data extraction project.

Aspects:

- Easy to use.
- Platform independency. However, Mozenda Agent Builder only runs on Windows.
- Working place independence:
  Tune the scraper, manage the scraping process and get scraped data from any computer connected to
  the Web.

***D.Web Info Extractor (WIE)** [9]:*

This is a tool for data mining, extracting Web content, and Web content analysis. WIE can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server[9].

Aspects:

- No need to learn boring and complex template rules, and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database[9].
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file.
- Can deal with Web page in all language.
- Running multi-task at the same time.
- Support recursive task definition.

***E.Web Content Extractor (WCE)** [13]:*

WCE is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet. It offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner[13]. This tool allows users to extract data from various websites such as online stores, online auctions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source[13].

- Helps to extract/collect the market figures, product pricing data, or real estate data.
- Helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- Assists users in automate extraction of auction information from auction sites.
- Assists to Journalists extract news and articles from news sites.
- Helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- Extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

Comparison of WCM Tools

The following table represents the web content mining toolsand their respective tasks [13].

Name of Tool	Tasks			
	Records the data	Extract Structured data	Extract Unstructured data	User friendly
Automation Anywhere	Yes	Yes	Yes	Yes
Web Info Extractor	No	Yes	Yes	Yes
Web Content Extractor	No	Yes	Yes	Not for Unstructured data
Screen Scraper	No	Yes	Yes	No
Mozenda	No	Yes	Yes	Yes

Table 1: Comparison of WCM Tools

In the above table we have explored some of the popular web content mining tools and provided their comparisons and
differences. The analysis results that the Screen Scraper tool is not user friendly among the different web content mining tools discussed. Also we observe that some of these tools seem to be applicable for E-mail Data Mining.

IV. USES OF WEB CONTENT MINING

Following are the uses of Web Content Mining:

- To gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information.
- To determine the relevance of the content to the search query.
- Improve the navigation of information on the web provides productive marketing.
- Produce a higher quality of information to the user.
- Understand customer behavior, evaluate effectiveness of a particular web site, and help quantify the success of a marketing campaign.
- Business intelligence. Competitive intelligence. Pricing analysis. Product data. Reputation.

V.WEB MINING APPLICATIONS

Web mining extends analysis much further by combining other corporate information with Web traffic data[6][7]. Practical applications of Web mining technology are abundant, and are by

no means the limit to this technology. 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[6].

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[11].

VI.FUTURE DIRECTIONS

As the Web and its usage grows, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that we continue to develop Web mining technologies that will enable this value in real field.

A.Web metrics and measurements

From an experimental human behaviourist's viewpoint, the Web is the perfect experimental apparatus. Not only does it provides the ability of measuring human behaviour at a micro level, it (i) eliminates the bias of the subjects knowing that they are participating in an experiment, and (ii) allows the number of participants to be many orders of magnitude larger. However, we have not even begun to appreciate the true impact of a revolutionary experimental apparatus. The WebLab of Amazon [AMZNa] is one of the early efforts in this direction[6]. It is regularly used to measure the user impact of various proposed changes - on operational metrics such as site visits and visit/buy ratios, as well as on financial metrics such as revenue and profit - before a deployment decision is made. For example, during Spring 2000 a 48 hour long experiment on the live site was carried out, involving over one million user sessions, before the decision to change Amazon's logo was made. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, so that various Web phenomena can be studied[14][15].

B.Process mining

Mining of 'market basket' data, collected at the point-of-sale in any store, has been one of the visible successes of data mining[7][8]. However, this data provides only the end result of the process, and that too decisions that ended up in product purchase.

Click-stream data provides the opportunity for a detailed look at the decision making process itself, and knowledge extracted from it can be used for optimizing the process, influencing the process, etc. Underhill has conclusively proven the value of process information in understanding users' behaviour in traditional shops[15].

Research needs to be carried out in (i) extracting process models from usage data, (ii)understanding how different parts of the process model impact various Web metrics of interest, and (iii) how the process models change in response to various changes that are made - changing stimuli to the user[14].

C.Web services optimization

As services over the Web continue to grow there will be a need to make them robust, scalable, efficient, etc. Web mining can be applied to better understand the behaviour of these services, and the knowledge extracted can be useful for various kinds of optimizations[7]. The successful application of Web mining for predictive pre-fetching of pages by a browser has been demonstrated. Research is needed in developing Web mining techniques to improve various other aspects of Web services[6].

D. Fraud and threat analysis

The anonymity provided by the Web has led to a significant increase in attempted fraud, from unauthorized use of individual credit cards to hacking into credit card databases for blackmail purposes. Yet another example is auction fraud, which has been

(a) Change in Web Content of a document over time
(b) Change in Web Structure of a document over time
(c) Change in Web Usage of a document over time

(a)

(b)

(c)

Figure 3: Temporal Evolution for a single document in the World Wide Web

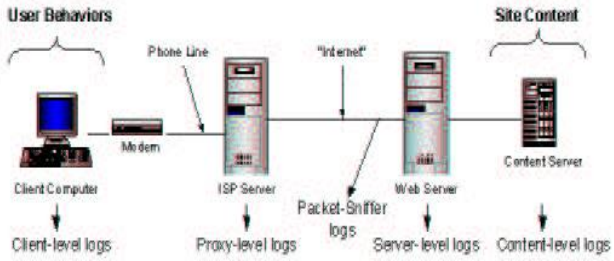

Figure 3.1: High Level Architecture of Different Web Services

increasing on popular sites like eBay. Since all these frauds are being perpetrated through the Internet, Web mining is the perfect analysis technique for detecting and preventing them. Research issues include developing techniques to recognize known frauds, and characterize and then recognize unknown or novel frauds, etc[7].The issues in cyber threat analysis and intrusion detection are quite similar in nature.

E.Web mining and privacy

While there are many benefits to be gained from Web mining, a clear drawback is the potential for severe violations of privacy[8][14]. Public attitude towards privacy seems to be almost schizophrenic - i.e. people say one thing and do quite the opposite. For example, famous case like [DG2000] and [DCLKa] seem to indicate that people value their privacy, while experience at major e-commerce portals shows that over 97 can be provided based on it. Spieker man et al [SGB2001] have demonstrated that people were willing to provide fairly personal information about themselves, which was completely irrelevant to the task at hand, if provided the right stimulus to do so. Furthermore, explicitly bringing attention information privacy policies had practically no effect[15]. One explanation of this seemingly contradictory attitude towards privacy may be that we have a bi-modal view of privacy, namely that ”I’d be willing to share information about myself as long as I get some (tangible or intangible) benefits from it, as long as there is an implicit guarantee that the information will not be abused”. The research issue generated by this attitude is the need to develop approaches, methodologies and tools that can be used to verify and validate that a Web service is indeed using an end-user’s information in a manner consistent with its stated policies.

VII.CONCLUSIONS

The Web content mining tools are primordial to scanning the many HTML documents, images, and text provided on Web pages. The result is provided to the search engines, in order of relevance giving more productive results of each search. As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge

from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we have briefly described the key contributions made by the IT field, the prominent successful applications, and outlined some promising areas of future research. Our hope is that this overview of content mining tools may provides a better idea for fruitful discussion.

REFERENCES

[1] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava , *"Web Mining: information and Pattern Discovery on the WWW"*

[2] Mary Garvin , *"Data Mining and the Web: What They Can Do Together*"

[3] Han J Kamber M, "Data Mining : concepts and Techniques" , Second Edition Morgan Kaufmann publishers .2006

[4]Lieu, B., *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data* (Springer-Verlag, Berlin, Heidelberg 2007).

[5] M. Zdravko, T.L. Daniel,, *Data mining the Web : Uncovering patterns in Web content, structure & usage* (WileyInterscience Publication, 2007).

[6] J. Srivastava , P. Desikan , V. Kumar, "Web Mining – Concepts, Applications and Research Directions", *Studies in Fuzziness and Soft*

Computing, Volume 180, pp. 275–307, (2005).

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

Edmonton, CA, 2002.

[8] M. Spiliopoulou, "Data Mining for the Web", *Proceedings of the Symposium on Principles of Knowledge Discovery in Databases* (PKDD), 1999.

[9] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.

[10] Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.

[11] Mozenda, http://www.mozenda.com/web-mining-software Viewed 18 February 2013.

[12] Web Content Extractor help. WCE, http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013.

[13] Raymond Kosala, Hendrik Blockee, "Web Mining Research : A Survey", ACM Sigkdd Explorations Newsletter, June 2000, Volume 2.

[14] Magdalini Eirinaki "Web Mining : A Roadmap" Http : //WWW.engr.sjsu.edu/meirinaki/papers/NEIS.pdf

[15] Qingyu Zhang & Richard S. Segall, "Web Mining: A Survey of Current Research", *Information Technology and Decision Making*, **7(4)**, 683-720, 2008.

[16] Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, *International Journal of Information Technology & Decision Making. Vol.7*, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

[17] Pol, K., Patil, N., Patankar, S. and Das, C. 2008. A Survey on Web Content Mining and extraction of Structured and Semi structured Data. IEEE First International Conference on Emerging.

[18] Nimgaonkar, S. and Duppala, S. 2012. A Survey on Web ContentMining and extraction of Structured and Semi structured data, IJCA Journal

[19] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations; Vol. 2, 1-15.

[20] Faustina Johnson and Santosh Kumar Gupta Web Content Mining Techniques: A Survey. International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012