Web Content Mining Techniques: A Survey

Faustina Johnson

Department of Computer Science & Engineering
Krishna Institute of Engineering & Technology,
Ghaziabad-201206, India

Santosh Kumar Gupta

Department of Computer Science & Engineering
Krishna Institute of Engineering & Technology,
Ghaziabad-201206, India

ABSTRACT

The Quest for knowledge has led to new discoveries and inventions. With the emergence of World Wide Web, it became a hub for all these discoveries and inventions. Web browsers became a tool to make the information available at our finger tips. As years passed World Wide Web became overloaded with information and it became hard to retrieve data according to the need. Web mining came as a rescue for the above problem. Web content mining is a subdivision under web mining. This paper deals with a study of different techniques and pattern of content mining and the areas which has been influenced by content mining. The web contains structured, unstructured, semi structured and multimedia data. This survey focuses on how to apply content mining on the above data. It also points out how web content mining can be utilized in web usage mining.

Keywords

Web Content mining, Web Usage Mining, Structured Data, Unstructured Data, Semi-structured Data, Multimedia Data.

1. INTRODUCTION

The advancement in technology paved the way for faster communication. The previous decade experienced a dramatic development in computer technology, such that with the press of a finger the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of web increased due to enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet. And this technique was named as web mining. Web mining is further classified into three: They are Web content mining, Web Structure mining, Web Usage mining [15]. Using the objects like text, pictures, multimedia etc. content mining is done in the web. In Web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users. And the study of this navigational pattern will trace out the interest of the users [15].

The structure of the paper is as follows: Section 2 presents theoverview of web mining, web content mining, techniques and tools of web content mining. Section 3 deals with literature work. Section 4 shows comparative study of web mining tools and Section 5 and 6 are conclusion and future scope respectively.

2. WEB MINING

2.1 Overview

Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data [15]. Web mining is used to capture relevant information, creating new knowledge out of the relevant data, personalization of the information, learning about Consumers or individual users and several others. Web mining uses data mining techniques to automatically discover and extract information from World Wide Web [2].

Several other techniques like Information retrieval, Information extraction and machine learning have been used in the past to discover the new knowledge from the huge amount of data available in the web. These techniques have been compared with web mining [15]. Information retrieval works by indexing text and then selects useful information [28]. Information Extraction focuses on extracting relevant facts whereas information retrieval selects relevant document. Web mining is now a part of Information retrieval system and Information Extraction system. IEhelps in preprocessing phase before web mining. It also helps in indexing which further helps in retrieval. Machine learning is not related to web mining directly but it supports web mining because it improves text classification process better than traditional Information Retrieval process [17]. Web mining is classified into the major three categories described in Fig. 1 [3].Web Content mining mines the content like text, image, audio, video, metadata, hyperlinks and extracts useful information. Since Web content mining examines the content of the web as well as the result of the search. Web Content mining mines the content like text, image, audio, video, metadata, hyperlinks and extracts useful information.

Web mining helps to understand customer behavior, helps to evaluate the performance of a web site and the research done in web content mining indirectly helps to boost business. Web content mining examines the search result of search engine. Manually doing things consumes a lot of time. When the data to be analyzed is in large quantities, then it is hard to find out the relevant data. Since now in every field of life manual work is replaced by technology. Same happened in the case of internet. As people already admit that internet is really a magic of technology. Web Mining became a boon to this magic. In the early stages Web contained few amount of data. So there was no need of web mining tools. As years passed Web got accumulated with large amount of data. Then retrieval of data according to users need became hard task. Web mining came as a rescue for this problem.

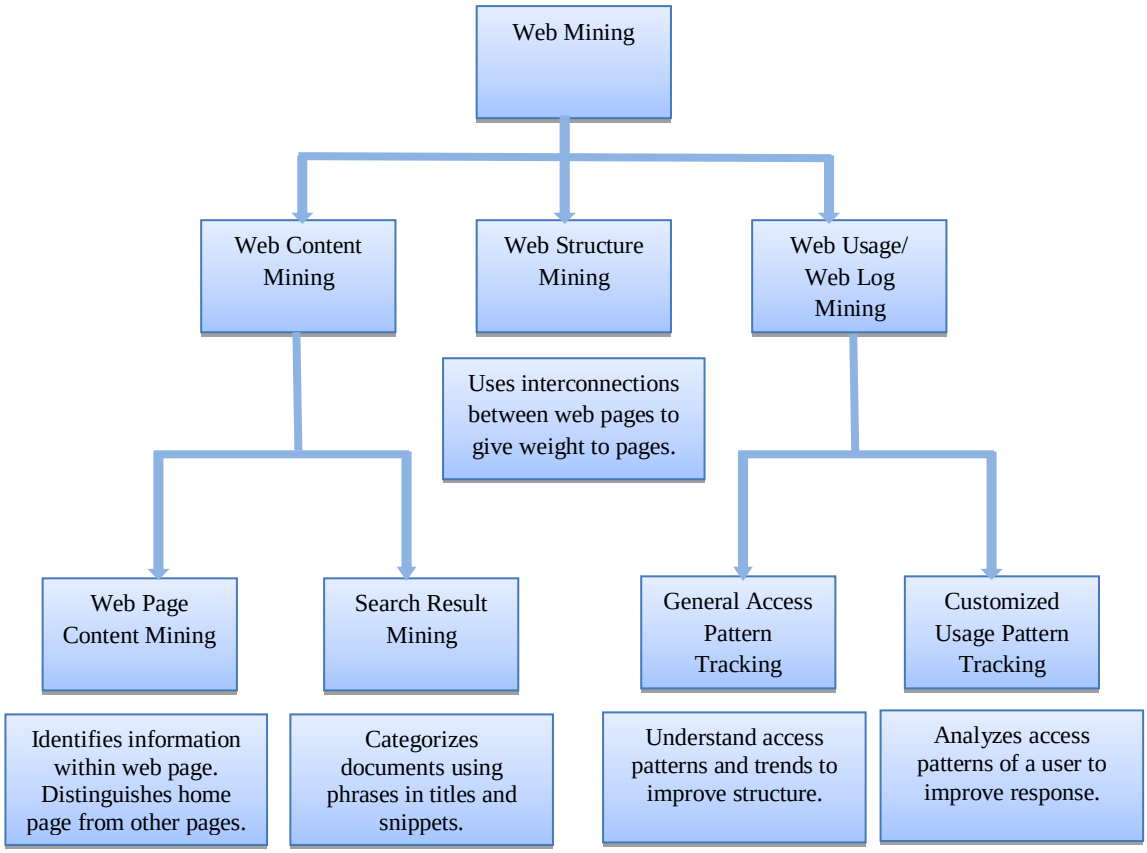

Fig1: Web Mining Taxonomy

Since Web content mining examines the content of the web as well as the result of the search. It can be further classified into Web page content mining and Search result mining. Web page Content mining is a traditional search of web page via content while search result mining is a further search of pages found from previous search [6]. Web Structure mining mines the structures like HTML or XML tags and gets information from the actual organization of the page. It uses interconnections between web pages to give weight to the page [6].Web usage mining is the application of data mining techniques to understand the web usage patterns. It mines data from logs, user profiles, user sessions, cookies, user queries, bookmarks, mouse click, scrolls etc. The three phases of web usage mining are preprocessing, pattern discovery, and pattern analysis [25]. Web usage mining is classified into two, they are General access pattern tracking and customized usage tracking. The mining using the history of the web page visited by the user is known as General access pattern tracking. It understands access patterns and trends to improve structure. When it is targeted on a specific user it becomes Customized usage tracking. It analyzes access patterns of a user to improve response [6].

2.2 Web Content Mining

Traditional technique of searching the web was via contents. Web Content mining is the extended work performed by search engines [6].Web Content mining refers to the discovery of useful information from web content such as text, images videos etc. [15, 13]. Two approaches used in web content mining are Agent based approach and database approach [13, 6].

The three types of agents are Intelligent search agents, Information filtering/Categorizing agent, Personalized web agents [13]. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles [13, 6]. In Database approach it consists of well formed database containing schemas and attributes with defined domains. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data. Fig 2 explains the web content mining techniques.

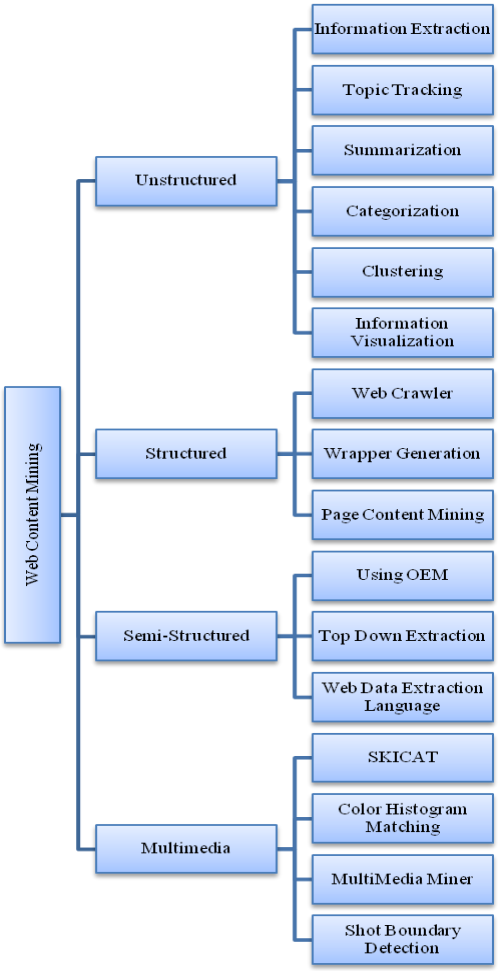

Fig 2: Web Content Mining Techniques

2.2.1 Unstructured Data Mining Techniques

Content mining can be done on unstructured data such as text.Mining of unstructured data give unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques [21]. Basic Content Mining is a type of text mining [6]. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.

Information Extraction

To extract information from unstructured data, pattern matching is used. It traces out the keyword and phrases and then finds out the connection of the keywords within the text. This technique is very useful when there is large volume of text. IE is the basis of many other techniques used for unstructured mining [8]. Information extraction can be provided to KDD module because information extraction has to transform unstructured text to more structured data. First the information is mined from the extracted data and then using different types of rules, the missed out information are found out. IE that makes incorrect predictions on data are discarded [12].

Topic Tracking

Topic Tracking is a technique in which it checks the documents viewed by the user and studies the user profiles. According to each user it predicts the other documents related to users interest. In Topic Tracking applied by yahoo, user can give a keyword and if anything related to the keyword pops up then it will be informed to the user. Same can be appliedin the case of mining unstructured data. An example for topic tracking is that if we select the competitors name then if at anytime their name will come up in the news then this information will be passed to the company. Topic tracking can be applied in many fields. Two such areas are medical field and education field. In medical field doctors can easily come to know latest treatments. In education field topic tracking can be used to find out the latest reference for research related work. Topic tracking helps to track all subsequent stories in the news stream. Disadvantage of topic tracking is that when we search for topics we may be provided with information which is not related to our interest. For example if user sets an alert for 'web mining' it can provide us with topics related to mineral mining etc. which are not useful for user [12].

Summarization

Summarization is used to reduce the length of the document by maintaining the main points. It helps the user to decide whether they should read this topic or not. The time taken by the technique to summarize the document is less than the time taken by the user to read the first paragraph. The challenge in summarization is to teach software to analyze semantics and to interpret the meaning. This software statistically weighs the sentence and then extracts important sentences from the document. To understand the key points summarization tool search for headings and sub headings to find out the important points of that document. This tool also give the freedom to the user to select how much percentage of the total text they want extracted as summary. It can work along with other tools such as Topic tracking and categorization to summarize the document. An example for text Summarization is Microsoft word's AutoSummarize [8].

Categorization

Categorization is the technique of identifying main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document. It does not process the actual information. It decides the main topic from the counts. It ranks the document according to the topics. Documents having majority content on a particular topic are ranked first. Categorization can be used in business and industries to provide customer support [8].

Clustering

Clustering is a technique used to group similar documents. Here in clustering grouping is not done based on predefined topic. It is done based on fly. Same documents can appear in different group. As a result useful documents will not be omitted from the search results. Clustering helps the user to easily select the topic of interest. Clustering technology is useful in management information system [8].

Information Visualization

Visualization utilizes feature extraction and key term indexing to build a graphical representation. Through visualization, documents having similarity are found out [12]. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed. It helps the user to visually

analyze the contents. User can interact with the graph by zooming, creating sub maps and scaling. This technique is useful to find out related topic from a very large amount of documents [8].

2.2.2 Structured Data Mining Techniques

The techniques used for mining structured data are Web Crawler, Wrapper Generation, Page content Mining.

Web Crawler

There are two types of Web Crawler which are called as External and Internal Web crawler. Crawlers are computer programs that traverse the hypertext structure in the web. External Crawler crawls through unknown website. Internal crawler crawls through internal pages of the website which are returned by external crawler.

Wrapper Generation

In Wrapper Generation, it provides information on the capability of sources. Web pages are already ranked by traditionalsearch engines. According to the query web pages are retrieved by using the value of page rank. 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.

Page Content Mining

Page Content Mining is structured data extraction technique which works on the pages ranked by traditional search engines. By comparing page Content rank it classifies the pages [18].

2.2.3 Semi-Structured Data Mining Techniques

The techniques used for semi structured data mining are Object Exchange Model (OEM), Top Down Extraction, and Web Data Extraction language.

Object Exchange Model (OEM)

Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange model (OEM). It helps the user to understand the information structure on the web more accurately. It is best suited for heterogeneous and dynamic environment. A main feature of object exchange model is self describing, there is no need to describe in advance the structure of an object.

Top down Extraction

In top down extraction, it extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted.

Web Data Extraction Language

In Web data extraction language it converts web data to structured data and delivers to end users. It stores data in the form of tables [18].

2.2.4 Multimedia Data Mining Techniques

Some of the Multimedia Data Mining Techniques are SKICAT, color Histogram Matching, Multimedia Miner and Shot Boundary Detection.

SKICAT

SKICAT is a successful astronomical data analysis and cataloging system which produces digital catalog of sky object. It uses machine learning technique to convert these objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set [19].

Color Histogram Matching

Color Histogram matching consists of Color histogram equalization and Smoothing. Equalization tries to find out correlation between color components. The problem faced by equalization is sparse data problem which is the presence of unwanted artifacts in equalized images. This problem is solved by using smoothening [3].

Multimedia Miner

MultiMedia Miner Comprises of four major steps., Image excavator for extraction of image and Video's, a preprocessor for extraction of image features and they are stored in a database, A search kernel is used for matching queries with image and video available in the database. The discovery module performs image information mining routines to trace out the patterns in images [30].

Shot Boundary Detection

It is a technique in which automatically the boundaries are detected between shots in video [5, 24].

2.2.5 Web Content Mining Tools

Web Content Mining tools are software that helps to download the essential information for users. It collects appropriate and perfectly fitting information. Some of them are Web Info Extractor, Mozenda, Screen-Scraper, Web Content Extractor, and Automation Anywhere 5.5 [4].

Web Info Extractor

This tool is helpful for mining and extracting content and monitoring content update**.**

Mozenda

Users can set up agents that regularly extract, store and circulate data to several destination.

Screen-Scraper

It searches a database, SQL server or SQL database, which interfaces with the software to achieve content mining requirements.

Web Content Extractor

It is a powerful and easy tool for web scraping, data mining and data retrieval

Automation Anywhere 5.5

It retrieves web data effortlessly, screen scrape from web pages or use it for web mining.

3. LITERATURE WORK

Web Content mining can be done by retrieving information from unstructured document such as free text and semi-structured document such as hypertext documents. In unstructured documents mining can be done by using word positions in the documents, text classification, event detection and tracking, finding extraction patterns in the text

documents. The method used for semi-structured documents are hypertext classification and clustering, learning relations between web documents, learning extraction pattern or rules, and finding patterns in semi-structured data [15]. Web content mining is being used in various different areas like In [27] web content mining is used for mining online news sites. Beyond analyzing the news, they focused on current society interest and measured the social importance of ongoing events. Dynamic Crawler was used for resource finding. For trend analysis they used domain independent statistical analysis. Four stages of dynamic news analysis are Resource Identification, preprocessing, Generalization and Analysis. In Resource Identification phase, dynamic web crawler downloads page from current URL and then filters the downloaded pages and analyze the identified news report. Steps are repeated till the queue of URL's are empty. In preprocessing stage the news are converted into structured format. Interesting trends among new topics are found in Generalization stage. In analysis phase user analyzes thepattern and the process is repeated until interesting news is found. According to this system crawler downloaded 350 web pages each day and only 130 were selected for further analysis during the period of two weeks. It has been shown by experimentation that enough differences on news topics after a period of two weeks have been found.

Another area where web content mining has been proved very useful is a web content suggestion system for distance learning and is described in [29]. Two ways of Suggestions are collaborative filtering and content based filtering. Collaborative suggestion clusters students into groups with similar behavior .Content based filtering provide web pages to the students who have navigation records. Web page navigation behavior is stored in personal records. Students who are new to attend the course will be having less navigation record so they are asked to poll the interest. Content Suggestion system works with the help of six components such as Student Assistant Agents, Student Identification Component, suggestion Generation Component, Suggestion Delivery Component, Data Warehouse.

A new algorithm for Web Content mining succeeded in extracting the information from query interfaces and then matches correlated attribute. First it mines the content of query interfaces and using clustering techniques information is extracted and they are placed in special domains. Query interfaces residing in the domain are matched by the system with the user query and finally query interface nearly similar to the user query are selected. Jaccard measure is used by this algorithm to distinguish positive and negative correlated attribute [2].

Problems faced by Web Content mining such as extracting information from heterogeneous environment, the redundancy, the linked nature of the web, the dynamic and noisy nature of the web were highlighted. Solutions for some of the above problems were also discussed [16, 23]. In [23] Content based Image retrieval was discussed in detail.

Web usage mining result can be improved by analyzing web content. The system integrates web page clustering into log file association mining and cluster labels are used as web page content indicators. The Web page clustering was done using K-means algorithm. The clusters obtained from the web log file and integrated data file were manually summarized. Then Apriori association rule mining algorithm was applied. This system utilized Web content mining for web usage mining [11].

Integration of web content mining into web usage mining is also possible [14, 26]. In [26] the textual content of the web pages are extracted through frequent word sequence. Then they are combined with web server logs tostudy association rule of user's behavior. The result of the proposed system helps in better recommendation, web personalization, web construction and web user profiling.

Connection between Web Content Mining and Web Structure mining was discussed in [10]. In this approach the web page content is compared with the information defined by the structure of the web site. Each web page is described with a set of keyword. This information iscombined with the link structure which generates context based description. This comparison helps in finding out semantic information of a web page and its neighborhood.

Page Content algorithm was created and the aim of the project was to create a better algorithm than Page Rank algorithm. The importance of page determines the importance of termwhich the page contains. The importance of the term is calculated based on a given query. For inner classification, page Content Rank uses neural network [20].

A system was proposed which provides irrelevant data along with the useful data thereby increasing the result of web content mining [22]. A review was done for implementable technique of web content mining and it explained how it can be applied to business field benefitting both the customer and the producer [1].

4. COMPARATIVE STUDY OF WEB CONTENT MINING TOOLS

Table 1 shows the web content mining tools and the tasks these tools perform [4].

Table 1. Tools and their Respective Tasks

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

5. CONCLUSION

This paper discusses the techniques of web content mining. Web content mining has been proved very useful in the business world. The survey also discusses the techniques used for extracting information from different types of data available in the internet and how this extracted data can be used for mining purposes. Users feel difficulty in finding desired information and deciding which information is

relevant to them from general purpose search engines. Web content mining solves this problem and helps the users to fulfill their needs. Topic Tracking is useful in predicting the web content related to users interest. Summarization helps the user to decide whether they should read a particular topic or not. Categorization can be used in business and industries to provide customer support. Clustering and information visualization are the techniques frequently being used for mining. Web content mining can also be applied to business application like mining online news site and developing a suggestion system for distance learning. Content Mining helps to establish better relationship with customer by providing exactly what they need. At the end paper discusses about different tools that can be used in web content mining.

6. FUTURE SCOPE

The future work involves developing of autonomous agents that analyze the discovered rules to provide meaningful courses of action or suggestions to users. Future scope of Web Content Mining includes predicting user needs in order to improve the usability, scalability, user retention, and framing an efficient framework for Web Personalization through efficient use Web Log file.

Semantic web is a future vision in which web content can be manipulated by automated systems for analysis and synthesis. In the internet information are mainly human readable. It is hard for the browser to understand the content; it can only interpret HTML mark-up to visualize its content. The three difficult tasks in the internet are Content Interpretation, Selection and management. Presently these three tasks are managed by humans. The semantic web will correct the balance between machine and human by reducing the three difficult tasks and make it automatic.

7. REFERENCES

[1] Ahmed, S. S., Halim, Z., Blaig, R. and Bashir, S. 2008. Web Content Mining: A Solution to Consumers Product Hunt. International Journal of Social and Human Sciences 2, 6-11.

[2] Ajoudanian, S. and Jazi, M. D. 2009. Deep Web Content Mining. World Academy of Science, Engineering and Technology 49.

[3] Bassiou, N. and Kotropoulos, C. 2006. Color Histogram Equalization using Probability Smoothening. Proceedings of XIV European Signal Processing Conference

[4] Bharanipriya, V. and Prasad, K. 2011. Web content Mining Tools: A Comparative study. International Journal of Information Technology and Knowledge Management. Vol. 4. No 1,211- 215.

[5] Cooper, M., Foote, J., Adcock, J. and Casi, S. 2003. Shot Boundary Detection via Similarity Analysis. In Proceedings of TRECVID 2003 workshop.

[6] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.

[7] Etzioni, O. 1996. The World Wide Web: quagmire or gold mine?. Communications of the ACM. Vol. 39. Issue 11. pp. 65-68.

[8] 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.

[9] Fayyad, U. M. 1995. SKICAT: Sky Image Cataloging and Analysis Tool. ACM Proceedings of the 14th International joint Conference on Artificial Intelligence. Vol. 2.

[10] Gedov, V., Stolz, C., Neuneir, R., Skubacz, M. and Siepel, D. 2004. Matching Web Site Structure andContent. ACM. Proceedings of the 13th International World Wide Web Conference on Alternate track papers and posters.

[11] Guo, J., Keselj, V. and Gao, Q. 2005. Integrating Web Content Clustering into Web Log Association Rule Mining. Springer Verlag. Vol. 3501 LNAI, 182-193.

[12] Gupta, V. and Lehal, G. S. 2009. A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. Vol. 1 .pp. 60-76.

[13] Inamdar, S. A. and shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.

[14] Kazienko, P. and Kiewra, M. 2003. Link Recommendation Method Based on Web Content and Usage Mining. New Trends in Intelligent Information Processing and Web Mining Proc. of the International IIS: IIPWM '03 Conference. Advances in soft Computing, Springer Verlag. 529-534.

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

[16] Liu, B. and Chiang K. C. 2004. Editorial Special Issue on Web Content Mining. ACM. Journal of Machine Learning Research 4, 177-210.

[17] MitChell, T. 1997. Machine Learning. McGraw Hill.

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

[19] Oh, J. and Bandi, B. 2002.Multimedia Data Mining Framework for Raw video sequences. ACM. Third International Workshop on Multimedia Data Mining. Pp. 1-10.

[20] Pokorny, J. and Smigansky, J. 2005. Page Content Rank: An Approach to the Web Content Mining. In proceedings of IADIS International Conference Applied Computing. Algarve, Portugal.

[21] 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 Trends in Engineering and Technology. pp.543-546.

[22] Poonkuzhali, G., Thiagarajan, K., Sarukesi, K. and Uma G. V. 2009. Signed Approach for Mining Web Content Outliers. World Academy of Science, Engineering and Technology 56.

[23] Singh, B. and Singh, H. K. 2010. Web Data Mining Research: A Survey. Computational Intelligence and Computing Research (ICCIC).IEEE International Conference, 1-10.

[24] Smeaton, A. F., Over, P. and Doherty, A. R. 2010. Video Shot Boundary Detection: Seven years of TRECVID Activity. Elsevier, Computer Vision and Image Understanding. Vol. 114, Issue 4. Pp. 411-418.

[25] Srivastava, J., Cooley, R., Deshpande, M., Tan, P. N. 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data.

[26] Taherizadeh, S. and Moghadam, N. 2009. Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting User's Behaviors. International Journal of Information Science and Management. Vol. 7, No. 1.

[27] Torreblanca, A. M., Gomez, M. M. and Lopez, A. L. 2002. A Trend Discovery System for Dynamic Web Content Mining. Proceedings of the 11th International Conference on Computing.

[28] Van. C. J. 1979. Information Retrieval. Butterworths.

[29] Yang, C. Y., Hsu, H. H. and Hung, J. C. 2006. A Web Content Suggestion System for Distance Learning. Tamkang Journal of Science and Engineering. Vol. 9, No. 3, 243-254.

[30] Zhang, J., Hsu, W. and Lee, M. L. 2001. Image Mining: Issues, FrameWorks and Techniques. In Proceedings of the 2nd International Workshop Multimedia Data Mining. pp. 13-20.

.