ISSN:0975-9646

A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites

Ananthi.J
Department of Computer Science and Engineering,
Hindusthan College of Engineering and Technology

***Abstract*—Web mining provides high performance system to the users to search for the product and obtains information of a particular product by searching through the servers that contains the sources. Web content mining used to extract the features of a product and labels the attributes in the result. Labeling is the process of identifying and naming the attributes after the information retrieval process. After the extraction and labeling process the information gained can be used for the analysis of the product and explorations. Web content mining is simply an integration of data from various sources by analyzing customers' view. This paper also presents a survey on web content mining methods used for mining and application of web content mining. The paper shows some of the emerging techniques used for extraction of data from online shopping sites.**

***Keywords*— Web Content Mining, Information Extraction, web document types, Mining techniques and Attribute Extraction and Online shopping sites Introduction**

I. INTRODUCTION

Web is taking an important place in human's life and day by day it increases the number of information based on the expectations of the customers using it. Daily Updation is needed to fulfil the needs of the users.

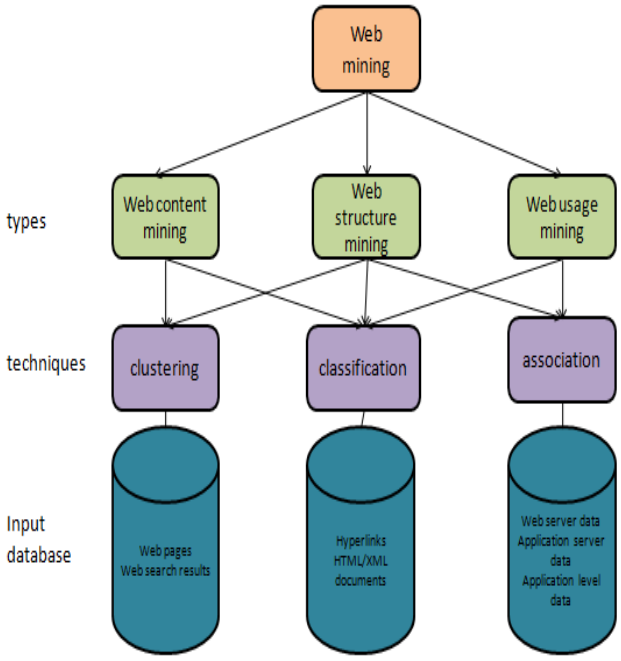

Fig 1 Categories of Web Mining

Web mining is used to extract the web information that is needed by the users so that the necessary details can be fetched and utilized. Automation is everywhere and in every field to avoid the human work in creation of anything. Web mining utilizes the automatic way of information extraction from the World Wide Web according to the preferences [2]. The three categories used for mining the web are mentioned below in the figure 1.

A. Web Content Mining

Web content mining is the process of extracting useful information from the web documents. It contains the generation of wrappers. Wrapper is a set of extraction rules to extract the data from the web pages, this can done either manually of automatically. The collection of data to be integrated may contain images, texts, audios or videos etc… this web content mining involves document tree extraction, data classification, and data clustering and finally labeling the attributes for results. Research activities are going on in information retrieval methods, natural language processing and computer vision.

B. Web Structure Mining

The process of discovering structures information from the web documents are called as web structure mining. This mining can be performed either document level or hyperlink level. The hyperlinks provide clear navigation and point to the pages. This is used to retrieve the useful information in the form of structure. Hyperlink analysis can be done based on knowledge models, scope and properties of analysis and types of algorithms. The methods that are done in the web usage mining are Data cleaning, Transaction identification, Data integration, Transformation, Pattern Discovery, Pattern Analysis

C. Web Usage Mining

Web usage mining is used to discover the interesting usage patterns form the usage data. This includes server data (IP address), Application server data (web logic), and Application level data (events). This is otherwise a Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. The source database is access logs, referrer logs, agent logs, and client-side cookies

II. METHODS OF WEB CONTENT MINING

The figure 2 shows the web content mining process and the information retrieved in the structured format.

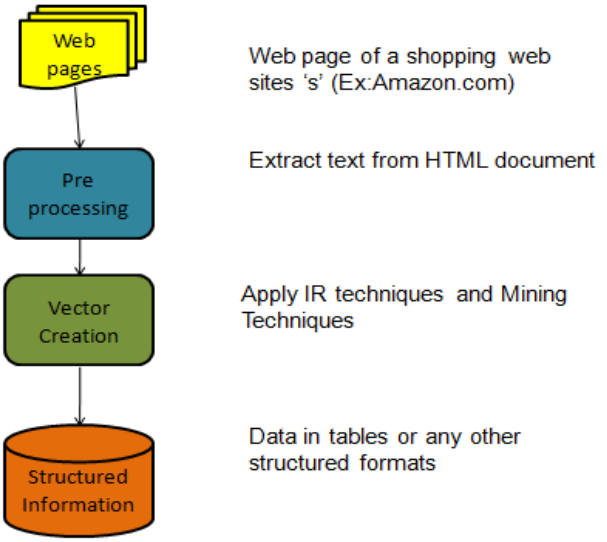

Fig 2 the Progress of Web Content Mining

Based on the documents in the web the traditional methods are partitioned into four parts [3] [7]. The techniques that are used for the four types of web documents are listed below in the table 1. Table 1 techniques of web content mining for various web documents.

TABLE I
TECHNIQUES FOR WEB CONTENT MINING

Document	Techniques	Process
Unstructured	Information Extraction	Extracting information from unstructured data and converts into structured data. Pattern matching and transformation are used
	Topic Tracking	Tracks the topics searched by the user and predicts the documents and produce to the user that of interest. Prediction techniques are used
	Summarization	Reduce the length of documents by minimizing the length of the documents. Analyzing the semantics and interprets the meaning of words
	Categorization	Documents are placed into a predefined group.
	Clustering	Used to group the similar documents Grouping based on the properties are identified
	Information Visualization	To build a graphical representation to the user Feature extraction, indexing techniques are used
Structured	Web crawlers	Traverse the hypertext structure of the web. Internal crawlers go through internal web pages of sites. External web crawlers go to the unknown links or sites.
	Wrapper Generation	Set of information extraction rule to extract the useful data from web pages. Provides a lot of meta information Page ranking is used
	Page Content Mining	Extracts the content of a page. Page ranking is used to display the results according to the rank
	Using OEM	Object Exchange Model. To understand the information structure of the web. Self describing structure of the data is produced
Semi Structured	Top Down Extraction	Complex objects of rich resources are converted into less complex objects.
	Web Data Extraction Language	Converts web data to structured data and delivers to end users.
Multimedia	SKICAT	Based on astronomical data analysis and cataloging system
	Color Histogram Matching	Find the correlation between the color components. Unwanted artifacts are removed using smoothing techniques
	Multimedia Miner	Extraction of images. Videos for the feature extraction, and feature comparison for matching queries
	Shot Boundary Detection	Automatic detection of boundaries

III. Emerging Techniques Of Web Content Mining For Online Shopping Sites

Online shopping systems information extraction helps to find the product specification and its features from the vast amount of products and its views. In earlier days the techniques used for the information extraction from web documents are based on the HTML documents. A tree structure is formed based on the HTML document of a web page. From that the information is retrieved through the search methodologies of a tree. The leaf node must be a text node which is to be extracted from the product. Then Hidden Markov Model parses and classifies the information needed and extraction is performed. This model was used to learn the attributes automatically.

Jun Zhu (2005) introduces 2D conditional Random Field to extract the object information automatically. This paper analyzed web documents of online shopping site as a 2D grid that consisting of object blocks. From the object blocks the needed blocks are extracted and modeled. The modeled data was labeled to identify the attributes of the particular product of user's specification [10].

Gengxin Miao (2009) focuses on the list of objects that appears repeatedly based on the tag paths in the DOM tree of the respective web documents. Then based on the comparison of the occurrence patterns of the tag paths the visually appearing signals are identified and clustering is performed based on the similarity measures of tag paths. This method had higher accuracy when comparing to previous methods [11].

Wei Liu (2010) presents an approach that extracts the products and its specifications from the online shopping web sites based on the visual features. All the visual features are considered such as content feature, format feature etc.. of the text document and clustered based on the similarity measures. This implementation also takes the DOM tree for data records extraction. From that extracted record the data items which are the product information can be retrieved [12].

AliGhobadi (2011) presents an improved web information extraction which is based on ontology. To extract the attributes that is of semantic meaning the ontology method of label identification for attributes are used. These processes make use of assumptions on information and fully understand the semantics of the HTML documents and extract the information automatically [13].

Xiaoqing Zheng (2012) introduced structural semantic entropy used for locating the data of interest in a web page, based on the measurement of the density of occurrence of the relevant information. Due to the difficulty of writing and, maintaining the wrappers and blocks identification in the vision based extractors this method has been introduced. Entropy measure is calculated to identify the density of the product specified and labeled [14]

IV. Applications Of Web Content Mining

Web content mining is used in various fields of large information maintenance. Cloud users need to extract the information from the cloud provided by web servers can utilize the web mining. Online shopping systems use the web mining to extract the information of a product and its specification through web mining. Opinion mining is the process of extracting reviews of a customer about the product and its specification using mining techniques. Web search makes the user to search over 2 billion data. It maintains the ranks among the pages and advertisement ordering and publish based on the user query. Web wide tracking is effectively done using web mining methodologies. Web communities can be maintained such as facebook. That is the users of same field of interest can be grouped and they can communicate through the network analyzed. Using web mining the customers'' behavior can be understood. Web page personalization now a days are very important to maintain the confidential information. Web mining is used for maintaining personalized data. Digital library performs automated citation indexing using web mining techniques. e-services include e-banking, search engines, on-line auctions, on-line knowledge management, social networking, e-learning, blog analysis, and personalization and recommendation systems. This can be analyzed for the customers and enable provision to the customers based on their recommendations [8].

V. Conclusions

Data mining techniques used for web information extraction are incredible system and recommended for the maintenance of highly confidential data. This is affluent, most intelligent resource extractor, and useful to maintain the historical data. Vast amount of data is maintained by the web sources and can be clearly extracted by the web mining techniques when the techniques are used accurately based on the requirements of the users.

References

[1] T.V.Mahendra,N.Deepika,N.Kesaca Rao," Data Mining for High Performance Data Cloud using Association Rule Mining",International Journal of Advanced Research in Computer Science and Software Engineering ,Vol2,Issue 1,January 2012.

[2] T.Sunil Kumar,Dr.K.Suvarchala, "A Study: Web Data Mining Challeneges and Application for Information Extraction",IOSR Journal of Computer Engineering (IOSRJCE), Vol 7,Issue3,Nov-Dec 2012,pp 24-29.

[3] Faustina Johnson, Santosh Kumar Gupta," Web Content Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 888),Volume 47– No.11, June 2012,pp.44-50

[4] S.Balan,P.Ponmuthuramalingam,"Astudy of Various Techniques of Web Content Mining Research Issues and Tools, International Journal of Innovative Research and Studies, Vol 2 Issues 5,May 2013

[5] Darshna Navadiya, Roshni Patel," Web Content Mining Techniques-A Comprehensive Survey", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, December- 2012,pp.1-6

[6] Basavaraj S. Anami, Ramesh S. Wadawadagi, Veerappa B. Pagi," Machine Learning Techniques in Web Content Mining: A Comparative Analysis", Journal of Information & Knowledge Management, Volume 13, Issue 01, March 2014

[7] Govind Murari Upadhyay, Kanika Dhingra,"Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013,pp.610-613

[8] Kohavi, R., Mason, L., Parekh, R., Zheng, Z. (2004) "Lessons and Challenges from Mining Retail E-commerce Data" Machine Learning, Vol. 57 No. 1-2, pp. 83-113

[9] Sandhya,Mala Chaturvedi,Anita Shrotriya,"Graph Theoratic Techniques for Web Content Mining", The International Journal Of Engineering And Science (IJES), Vol 2,Issue 7 July 2013,pp.35-41

[10] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, Wei-Ying Ma," 2D Conditional Random Fields for Web Information Extraction ",Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[11] Gengxin Miao,Junichi Tatemura, Wang-pin Hsiung, Arsany Sawires, Louise E.Moser,"Extracting Data Records from the Web Using Tag Path Clustering",International World Wide Web conference Committee (IW3C2),April,2009,pp.981-990.

[12] Wei Liu, Xiaofeng Meng , Weiyi Meng ,"ViDE: A Vision-based Approach for Deep Web Data Extraction", IEEE Transactions on Knowledge and Data Engineering, Volume:22 , Issue: 3, March 2010,pp. 447 – 460

[13] Ali Ghobadi,Maseud Rahgozar,"An ontology based Semantic Extraction Approach for B2C eCommerce",The International Arab Journal of Information Technology Vol.8, No. 2,April 2011,pp.163-170

[14] Xiaoqing Zheng,Yiling Gu,Yinsheng Li,"Data Extraction from Web Pages Based on Structural Semantic Entropy", International World Wide Web conference Committee (IW3C2),April 2012,pp.93-102