International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

Information Extraction from Online Shopping Sites using Web Content Mining Methods and Techniques

Gadamsetty Vasavi[1]
Research Scholar,
Department Of Computer Science,
SPMVV, Tirupati.

Dr. T. Sudha[2]
Professor,
Department of Computer Science,
SPMVV, Tirupati.

Abstract— ***At beginning, the extraction of valuable knowledge from the Web has been difficult. Web mining is an application of Data Mining techniques to extract knowledge from Web Content, Web Structure and Web usage. All these are collection of technologies to satisfy the potential sources. Web content mining is used to extract the features of a product and labels the attributes. The process of identifying and naming the attributes after the information retrieval is called Labeling. After the extraction and labeling the resultant information can be used for the analysis of the product and explorations. Web content mining is simply an integration of data from various sources analyzed from the customer's point of view. This paper represents an analysis on web content mining methods and techniques. Also this paper presents some of the emerging techniques used to extracting the data from online shopping sites and also shows some applications of Web content mining.***

Keywords— Web Content Mining, Information Extraction, Web Mining techniques, Applications of Web mining, Attribute Extraction and Introduction of Online shopping sites.

I.INTRODUCTION

Web is taking an important role in human's life and day by day it increases the information based on the expectations of the customers using it. Updated information is necessary to fulfill the needs of the users.

Web mining is the application of Data Mining to automatically fetch and evaluate information from the web services and documents. 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]. Web mining The three categories used for mining the web are mentioned below in the figure 1.

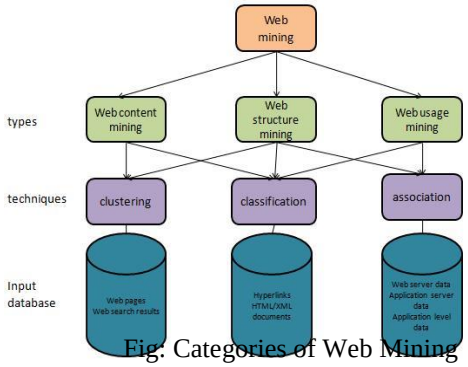

Fig. Categories of Web Mining

Web Content Mining:

Web content mining is the mining extraction and integration of useful data, information and knowledge from web page contents. Content data is the collections of facts a web page. It may consist of text, images, audio, video, or structured records such as lists and tables. 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. 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. It contains

the generation of wrappers. A set of extraction rules to extract the data from the web pages can do either manually or automatically is called Wrapper. Applications of text mining to web content have been most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. 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 the Information Retrieval (IR) and Natural Language Processing (NLP). A significant work had done in extracting knowledge from images in the area of image processing and computer vision. But applications of these techniques to web content mining have been limited.

A. Web Structure Mining:

Web Structure Mining is a tool used to identify the relationship between web pages linked by information or direct link connection. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related web pages. This connection allows a search engine to get data related to the search query from the web page of related web site. This can be divided into two types based structure information.

Hyperlinks:

A hyperlink is a structural unit that connects a location in a web page to a different location that can either same or different web page. A hyperlink that connects to a different part of the same web page is called intra-document hyperlink and hyperlink that connect two or more different pages is called an inter-document hyperlink.

Document Structure:

The content within a web page can also be organized in tree-structured format based on the HTML and XML tags within the page. Mining efforts focused on automatically extracting Document Object Model(DOM) structures out of documents (Wang and Liu 1998; Moh, Lim, and Ng 2000).

C. Web Usage Mining:

Web Usage Mining is the process of extracting useful information from server logs i.e. users history. Web usage mining is the application of data mining techniques to discover usage patterns from the web usage data. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining can classify into three types based on the type of usage data considered.

Web Server Data

User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data

Commercial applications like Weblogic, StoryServer have some significant features that enable E-commerce applications. Key feature of this is the ability to track various types of business events and log them in application server logs.

Application Level Data

A very new kind of events that can be defined in an application, and logging them for generating histories of these events.

II.METHODS OF WEB CONTENT MINING

The Web Content Mining is differentiated from two different points of view: information Retrieval View and Database View.

Research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use list of words, which are based on the statistics. Single words in isolation that represent unstructured text and take single word found in the training corpus as features. For semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents.

In the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site to transform a web site to become a database. This type of mining uses the ideas and principles of data mining and knowledge discovery to screen more specific data.

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

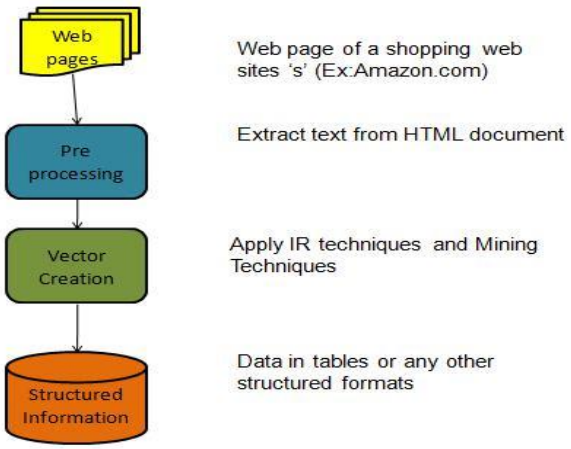

Fig 2: the Progress of Web Content Mining

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.

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:

Information extraction from Online shopping systems helps to find the product specification and its features from the huge amount of products and its views. In earlier days the techniques used for the information extraction from web documents depends on the HTML documents. A tree structure is formed based on the HTML document of a web page. From that information is retrieved through the search methods of a tree. The leaf node must be a text node which is extracted from the product. Then Hidden Markov Model parses and classifies the needed information and extraction is performed. This model was used to know 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 consist 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 are visually appeared as signals are identified. 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 as content features and format features, 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 that based on the measurement of the density of occurrence of the relevant information. This method has been introduced due to the difficulty of writing and maintaining of the wrappers and blocks identification in the vision based extractors. 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.

1. Cloud users need to extract the information from the cloud provided by web servers can utilize the web mining.
2. Online shopping systems use the web mining to extract the information of a product and its specification through web mining.
3. Opinion mining is the process of extracting reviews of a customer about the product and its specification using mining techniques.
4. Web search makes the user to search over two billion data. It maintains the ranks among the pages and advertisement ordering and publish based on the user query.
5. Web wide tracking is effectively done using web mining methodologies.
6. Web communities can be maintained such as face book. That is the users of same field of interest can be grouped and they can communicate through the network analyzed.
7. 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. CONCLUSION

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] Mining Techniques-A Comprehensive Survey", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, December- 2012,pp.1-6

[7] 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

[8] 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

[9] 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

[10] 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

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

[12] 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.

[13] Wei Liu, Xiaofeng Meng , Weiyi Meng , "ViDE: A Vision-based Approach for Deep Web Data Extraction", IEEE 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 ference Committee (IW3C2),April 2012,pp.93-102