IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. VI (Nov.-Dec. 2016), PP 57-60
www.iosrjournals.org

Web Content Mining: Tool, Technique & Concept

Anil Kumar Sinha[1], Nidhi Raj[2], Shameemul Haque[3], Alimul Haque [4] and N.K.Singh[5]

[1]*Department of MCA, V.K.S.University, Ara, Bihar.*
[2]*B.Tech (VI-Sem)., Department of Computer Science & Engineering, S.B.M. Jain College of Engineering, Jain University, Bangalore-560004, India.*
[3,4,5]*University Department of Physics, V.K.S.University, Ara, Bihar. India.*

Abstract: *Today the evolution of the World Wide Web has brought us enormous and ever growing amounts of data and information.. The web is full of structured or unstructured information which is directly or indirectly influencing society or people in every day life. Due to huge data available with the web it has become a challenge for people with interest in utilizing this information. For mining the huge available data on web in the form of frequently used patterns and log files for meaning use is considered in this paper. The paper presents a preliminary review of Web content mining contributions in the field of web mining and the prominent successful algorithms and some tools.*

Keywords: *Social Web Mining, Web Content Mining, Web Mining, and Application of Web content Mining.*

I. Introduction

The continuous growth of World Wide Web and Internet produces bulk amount of data and information. The low cost abundant data provided by the web, Influences almost all aspects of people's lives. Hence this web data is more attractive to researchers [1]. Researchers can retrieve web data by browsing and keyword searching [6]. However, there are several limitations to these techniques. It is hard for researchers to retrieve data by browsing because there are many following links contained in a web page. Keyword searching will return a large amount of irrelevant data. On the other hand, traditional data extraction and mining techniques can not be applied directly to the web due to its semi-structured or even unstructured nature. Web pages are Hypertext documents, which contain both text and hyperlinks to other documents. Furthermore, other data sources also exist, such as mailing lists, newsgroups, forums, etc. Thus, design and implementation of a web mining research support system has become a challenge for people with interest in utilizing information from the web for their research. A web mining research support system should be able to identify web sources according to research needs, including identifying availability, relevance and importance of web sites; it should be able to select data to be extracted, because a web site can be viewed.

II. Retrieval and Mining techniques

There are four type of mining technique as shown in table: 1. Data mining, Text mining, Web mining and Spatial mining. Data mining is a technique to discover and analyze the new pattern or knowledge previously unknown from any data. Text mining is a technique to discover and analyze the new pattern or knowledge previously unknown from TEXT data. Web mining is a technique to discover and analyze the new pattern or knowledge previously unknown from Web related data. Spatial mining is a technique to discover and analyze the new pattern or knowledge previously unknown.

Table 1: Retrieval and mining techniques

S. No	Purpose	Data Information Sources			
		Any data	**Textual Data**	**Web Related data**	**Geographical data**
1	Retrieving known data or document efficiently and effectively	Data Retrieval	Information Retrieval	Web information Retrieval	Geographical or spatial information
2	Finding new pattern or knowledge previously unknown	Data mining	Text Mining	Web mining	Spatial Mining

Web Mining

Web mining is a technique (Table: 2) to discover and analyze the useful information from the Web related data and services [12]. According to Etzioni [2], web mining can be divided into four subtasks:**(i) Information Retrieval/Resource Discovery (IR):** Find all relevant documents on the web. The goal of IR is to automatically find all relevant documents, while at the same time filter out the non relevant ones. Search engines are a major tool people use to find web information.

(ii) Information Extraction (IE): automatically extract specific fragments of a document from web resources retrieved from the IR step. Building a uniform IE system is difficult because the web content is dynamic and diverse. Most IE systems use the "wrapper" [3] technique to extract specific information for a particular site. Machine learning techniques are also used to learn the extraction rules.
(iii) Generalization: discover information patterns at retrieved web sites. The purpose of this task is to study users' behaviour and interest. Data mining techniques such as clustering and association rules are utilized here
(iv) Analysis/Validation: analyze, interpret and validate the potential information from the information patterns. The objective of this task is to discover knowledge from the information provided by former tasks. Based on web data, we can build models to simulate and validate web information.

Web Mining Tasks

Web mining tasks can be divided into several classes. Table: 2 shows different categories of Web mining tasks. In web content mining as per IR view task based on unstructured and semi structured data where as the main data is text document and hypertext document. The method used in this process is machine learning, variants and statistical. The application based on categorization, clustering, finding extraction rules, finding patterns in text and various user modeling. In web structure mining task based on link structure data where as the main data is also linked structure. The method used in this process is proprietary algorithms and application based on categorization and clustering. In web uses mining data collected from user interaction where as main data collected from server log and browser log. The method used in this process based on machine learning, statistical and association rules.

Table: 2 Web mining: view, data, method and application

	Web Mining			
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR view	**DB view**		
View of data	Unstructured , Semi-structured	Semi-structured Web site as DB	Links structure	Interactivity
Main Data	Text Document , Hypertext document	Hypertext document	Links structure	Server Log, browser Log
Representation	Bag of words, n-grams, Terms, Phrases, Concepts or ontology, relational	Edge- labeled graph (OEM), Relational	Graph	Relational table, Graph
Method	TFIDF and variants, Machine learning, Statistical	Proprietary algorithms, ILP, association rules	Proprietary algorithms [1]	Machine learning, Statistical, association rules
Application Categories	Categorization, clustering, Finding extraction rules, finding patterns in text, user modeling	Finding frequent sub structures, web site schema discovery	Categorization, Clustering	Site construction, adaption, and management, Marketing, user modeling

Web Content Mining

Web Content Mining [7] is the process to describes the discovery of 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, metadata, hyperlinks or structured records such as lists and tables [4]. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages [5].

The web content mining is differentiated from two different points of view [8]: Information Retrieval View and Database View.R.kosla and H.Blockeel [9] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structures between the documents for document representation. As for 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 of to transform a web site to become a database. Multimedia data mining is part of the content mining, which is engaged to mine the high- level information and knowledge from large online multimedia sources.

Web Content Mining methods and technique

Web page consists of text, images, audio, video, metadata, hyperlinks or structured records and tables. It may be categorize as Unstructured Data, Structured Data, Semi-Structured Data and Multimedia Data. The mining method used to discover knowledge are Unstructured text mining [11], Structured Data Mining, Semi-Structured Data Mining and Multimedia Data Mining [10] as given in table: 3.

Table: 3 Different approaches for web content mining

Web Content Data Type	Mining Method	Techniques
Unstructured Data	Unstructured text mining	Information Extraction
		Topic Tracking
		Summarization
		Categorization
		Clustering
		Information Visualization [10]
Structured Data	Structured Data Mining	Web Crawler
		Wrapper Generation
		Page content Mining [10]
Semi-Structured Data	Semi-Structured Data Mining	Object Exchange Model (OEM),
		Top Down Extraction
		Web Data Extraction language [10]
Multimedia Data	Multimedia Data Mining	SKICAT
		Color Histogram Matching
		Multimedia Miner
		Shot Boundary Detection

Web content Mining Tool

Some of the widely used web content mining tools are Web Content Extractor, Web info Extractor, Automation Anywhere, Screen Scraper and Mozenda. The tool names their properties and the people to whom it helps are given in table: 4.

Table: 4 Web content Mining Tool

Tool Name	Tool Properties	To Whom Help
Web Content Extractor	Extract and collect market figure, Product pricing data, real estate data	Business man
	Extract the information about books, titles, authors ISBN, images, price from online book sellers	Book lovers
	Extract news , Articles from book sites	Journalist
	Extract information about Vacation and holiday places, their name, addresses, descriptions, images , price from web sites	Tourist
Web info Extractor	Easy to define extraction task, no need to learn template rule	People involve in Data mining
	Retrieve unstructured data as well as tabular data to file, database	People involve in Data mining
	Monitor web page and retrieve new content	People involve in Data mining
	Unicode support can process web page in all language	Any one
	Running multiple task at a time	People involve in Data mining
Automation Anywhere	Intelligent automation is used	Business and IT Task
	Unique SMART Automation Technology	
	Creating automation tasks few minutes	Record mouse and keyboard strokes
	Distribute tasks to multiple computers easily	
	It is used to automate scripts in disparate formats	
Screen Scraper	GUI interface is provided	To build sitemaps
	Item have been created from external languages such as .NET, java, PHP and ASP	
	Programming language can be used to access screen scraper	
	Mine data on products and download them to a spreadsheet	extract multiple types of data - text, tables, images, links and more
Mozenda	Mined data can be accessed online, exported , as well as used throughout an API	
	It perform your scraper within the clouds	

III. Conclusion

In this paper, first we have mainly focused on Retrieval and Mining techniques. After that, we have introduced the web mining types - Web content mining, web structure mining and web usage mining. The web mining task- IR will identify web sources by predefined categories with automatic classification. IE will use a hybrid extraction way to select portions from a web page and put data into databases. Generalization will clean data and use database techniques to analyze collected data. Simulation and Validation will build models based on those data and validate their correctness. After that, we have introduced the web mining techniques in the area of the Web Content Mining. After that Web Content Mining Approaches to mine Web Content Data and Web Content Mining Tool. The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. Thus various Data mining techniques and web content mining are used

to extract useful information or knowledge from web page contents. By these techniques we can make our search of contents over the web faster and exact

References

[1]. Q. Han, X. Gao, and W. Wu, Study on Web Mining *Algorithm Based on Usage Mining*, 2010.
[2]. O. Etzioni. The world-wide web: Quagmire or gold mine? *Communications of the ACM*, 39(11), 1996, 65-68,.
[3]. L. Eikvil. Information extraction from world wide web - a survey. *Technical Report 945, Norweigan Computing Center*, 1999.
[4]. A. J. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques," *Journal of and applied information technology*, 2005.
[5]. W. Bin and L. Zhijing, "Web Mining Research," in Proceedings of the *fifth International Conference on Intelligenceand Multimedia Applications* (ICCIMA'03), 2003.
[6]. A. Laender, B. Ribeiro-Neto, A. Silva, and J. Teixeira. A brief survey of web data extraction tools. *In SIGMOD Record*, 31, June 2002.
[7]. Abdel hakim Herrouz, Chabane Khentout, and Mahieddine Djoudi,'' Overview of Web Content Mining Tools", *The International Journal of Engineering And Scienc*e (IJES), Vol.2, Issue 6, 2013, pp. 106-110
[8]. R.Cooley, B.Mobasher, and J.Srivastava.Web mining: Information and pattern discovery on the World Wide Web. In proceedings of the 9th IEEE *International Conference on Tools with Artificial Intelligence* (ICTAI'97), 1997.
[9]. R.kosala and H.Blockeel. Web mining Research: A survey. *SIGKDD*, Volume 2, Issue 1 , 2000.,pp 1- 15.
[10]. F.Johnson, and S.K.Gupta,., Web Content Minings Techniques: A Survey, *International Journal of Computer Application.* 47(.11), 2012, 44.
[11]. Darshna Navadiya and Roshni, Web Content Mining Techniques-A comprehensive Survey, *International Journal of Engineering Research & Technology (IJERT)* , Vol. 1 Issue 10, , 2012 , ISSN: 2278-0181
[12]. Han J, and Kamber M, "Data Mining: Concepts and Techniques", *Second edition, Morgan Kaufmann Publishers*, 628-648. 2006, [Accessed on Feb. 18, 2013].