A Survey on Web Content Mining Methods and Applications for Perfect Catch Responses

A. Richlin Selina Jebakumari[1] , Dr. Nancy Jasmine Goldena[2]

[1]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli
[2]Assistant Professor, Sarah Tucker College, Manonmaniam Sundaranar University, Tirunelveli

---

***Abstract* -** Web content mining, also known as text mining, is generally the second step in Web data mining. 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. Web content mining is very effective when used in relation to a content database dealing with specific topics. For example universities use a online library system to recall articles related to the general areas of study. This specific content database enables to pull only the information within those subjects, providing the most specific results of search queries in search engines. This method of allowing only the most relevant information being provided gives a higher quality of results. This increase of productivity is due todirect usage of content mining. 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 for perfect catching responses.

***Keywords*— Web Content Mining, Perfect Catching, Web data Mining, Text Mining, Web Search Queries.**

1. INTRODUCTION

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services[6].

There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.
- Web graph, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

At Scale Unlimited we focus on the last one – extracting value from web pages and other documents found on the web[2]. Note that there's no explicit reference to "search" in the above description. While search is the biggest web miner by far, and generates the most revenue, there are many other valuable end uses for web mining results[1]. A partial list includes:

- Business intelligence
- Competitive intelligence
- Pricing analysis
- Events
- Product data
- Popularity
- Reputation

When extracting Web content information using web mining, there are four typical steps.

- Collect – fetch the content from the Web
- Parse – extract usable data from formatted data (HTML, PDF, etc)
- Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
- Produce – turn the results of analysis into something useful (report, search index, etc)

When comparing web mining with traditional data mining, there are three main differences to consider:

- **Scale** – In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.

- **Access** – When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated (non-user) access to this data.

- **Structure** – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup[8].

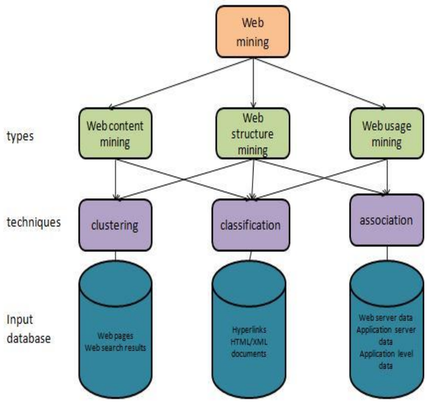

Fig-1: Web Mining Architecture

2. WEB CONTENT MINING METHODS

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

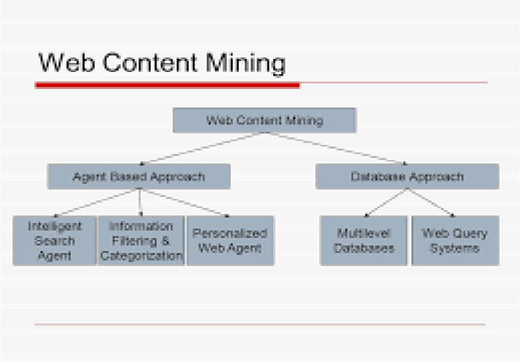

Fig-2: Progress of Web Content Mining

3. APPROACHES OF WEB CONTENT MINING

Web Content mining has following approaches to mine data: unstructured mining, structured mining, semi-structured mining and multimedia mining.[9]

3.1 Unstructured Data Mining

Text document is the form of unstructured data. Most of the data that is available on web is unstructured data. The research of applying data mining techniques to unstructured data is known as knowledge discovery in texts [4].

3.1.1 Information Extraction

Extracting the informations based on search query, if found display the results otherwise data not found error.

3.1.2 Topic Tracking

It is equilant to the key word searching and content identification markers.

3.1.3 Summarization

It acts as the total summary for the perfect catch responses.

3.1.4 Categorization

Different thematic content documents are grouped accordingly.

3.1.5 Clustering

Related content documents are grouped together.

3.1.6 Information Visualization

Heavy contentmaterials are represented as visual maps or graphs. It helps in visually analyzing the content. The user can interact by scaling, zooming and creating sub components.

3.2 Semi-Structured Data Mining

Semi-structured data arises when source does not impose rigid structure on data. If we want to extract data from web page and populate that data in database [3].

3.2.1 Object Exchange Model

The relevant information is extracted from semi-structured and is collected in a group of useful information and is then stored in Object Exchange Model (OEM).

3.2.2 Top down Extraction

This technique helps in extracting complex objects from a rich web sources and decompose them into less complex objects until atomic objects have been extracted.

3.2.3 Web Data Extraction Language

This technique helps in converting web data to structured data and then delivers this data to end users. The data is stored in the form of tables [9].

3.3 Structured Data Mining

The techniques are used to extract structured data from web pages [11]. Data in the form of list, tables and tree is structured data. The structured data is easy to extract as compared to unstructured data.

3.3.1 Web Crawler

Crawlers are computer programs which traverse the hypertext structure in web. Web crawlers can be used by anyone to collect information from the web. Search engines use crawlers frequently to collect information about what is available on public web pages [10]. There are two types of crawlers. They are internal and external web crawler. Internal web crawler crawls through internal pages of the website and the external crawler crawls through unknown sites [5].

3.3.2 Page Content Mining

Page content mining is a technique that is used to extract structured data which works on the pages that are ranked by the traditional search engines. The pages are classified by comparing the page content rank [11].

3.3.3 Wrapper Generation

The information is provided by the wrapper generator on the capability of sources. Web pages are ranked by traditional search engines. By using the page rank value the web pages are retrieved according to the query [4].

3.4 Multimedia Data Mining

Multimedia data mining is the process of finding interesting patterns from media data such as video, audio, text and images that are not accessible using queries [12].

3.4.1 SKICAT

SKICAT is a successful astronomical data analysis and cataloging system that produces digital catalog of sky object. It uses machine learning techniques to convert the objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set [7].

3.4.2 Color Histogram Matching

Color histogram matching consists of color histogram equalization and smoothing. Equalization tries to find the correlation between color components. The problem faced by equalization is the presence of unwanted artifacts in equalized images. The problem is solved by using smoothening [13].

3.4.3 Multimedia Miner

Multimedia miner consists of four major steps. Image excavator for extraction of images and videos, a preprocessor for extraction of image features and are stored in a database.[14]

4. LITERATURE SURVEY

Table-1: Literature survey data

Author	Approach	Implementation
Junker, et al. [15]	Inductive Logic Programming	Text categorization Learning rules
Kargupta, al. [16]	Supervised Hierarchical Clustering Decision trees Statistical Analysis	Text classification and Hierarchical Clustering
Nahm Mooney[17]	Decision trees	Predicting relationship
Nigam, et al.[18]	Maximum entropy	Text classification
Scott[19] And matwin	Rule based system	Text classification
Soderland [20]	Rule learning	Learning extraction rules
Weiss, et al. [21]	Boosted decision trees	Text categorization
Wiener, et al [22]	Neural Networks Logistic Regression	Text categorization
Witten, et al.[23]	Text Compression	Named entity classifier
Yang, et al.[24]	Clustering algorithms k-Nearest Neighbor Decision Tree	Event detection and tracking
Craven, et al. [25]	Modified Naive Bayes Inductive Logic	HypertextClassification learning Web page relation
Crimmins, et al. [26]	Unsupervised graphical supervised	Hierarchical and Classification Clustering

	classification algorithms	
F¨urnkranz [27]	Rule learning classification	Hypertext

5. ISSUES IN WEB CONTENT MINING

- ➢ Web information sets can be substantial; it takes ten too many terabytes to store on the database.
- ➢ It can't mine on a solitary server so it needs substantial number of server.
- ➢ Proper organization of software and hardware to mine multi-terabyte information sets.
- ➢ Limited customization, constrained scope, and restricted inquiry interface to individual clients.
- ➢ Automated information cleaning.
- ➢ Over fitting and under fitting of information.
- ➢ Over Sampling of information.
- ➢ Scaling up for high dimensional information.
- ➢ Difficulty in finding important data Removing new information from the web

6. IMPLEMENTATION

➢ Structured data extraction ‰

„ A large amount of information on the Web is contained in regularly structured data objects. ‰ Which are data records retrieved from databases. „ Such Web data records are important because ‰ they often present the essential information of their host pages, e.g., lists of products and services. „ Applications: integrated and value-added services, e.g., ‰ Comparative shopping, meta-search & query, etc.

➢ Classification Analysis

„ The Web has dramatically changed the way that consumers express their opinions. „ One can post reviews of products at merchant sites, Web forums, discussion groups, blogs „ Techniques are being developed to exploit these sources to help companies and individuals to gain market intelligence info. „ Benefits: ‰ Potential Customer: No need to read many reviews ‰ Product manufacturer: market intelligence, product benchmarking

➢ Information integration and matching

„ Many integration tasks, ‰ Integrating Web query interfaces (search forms) ‰ Integrating ontologies (taxonomy) ‰ Integrating extracted data ‰ Integrating textual information ‰ ... „ We only introduce integration of query interfaces. ‰ Many web sites provide forms to query deep web ‰ Applications: meta-search and meta-query

➢ Knowledge Synthesis ‰

„ Given a query, a few words ‰ a search engine returns a ranked list of pages. ‰ The user then browses and reads the pages to find what s/he wants. „ Sufficient ‰ if one is looking for a specific piece of information, e.g., homepage of a person, a paper. „ Not sufficient for ‰ open-ended research or exploration, for which more can be done.

➢ Template detection & page Segmentation

„ Most web sites, especially commercial sites, use well designed templates. ‰ A templatized page is one among a number of pages sharing a common look and feel. „ A templatized page typically contains many blocks: ‰ Main content blocks ‰ Navigation blocks ‰ Service blocks, ‰ Advertisements, etc. „ Each block is basically an (micro) information unit. „ Due to diverse information in the blocks, templatized pages affect search ranking.

7. TOOLS PERFORMANCE

The following table illustrates the Web content mining tools with its implementation scope for different types of web data.

Table-2: Tools Performance Analysis Report

Web Content Mining Tool	Structured data process	Semi Structured Data process	Unstructured Data process
R	Yes	No	No
Octoparse	Yes	Yes	No
OracleDM	Yes	Yes	Yes
Tableau	Yes	No	No
Scrapy	Yes	Yes	Yes
Weka	Yes	Yes	Yes
Orange	Yes	Yes	No

8. RECENT TRENDS

➢ *Data/information extraction:* Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.

➢ *Web information integration and schema matching:* Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.

➢ *Opinion extraction from online sources:* There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking. We will introduce a few tasks and techniques to mine such sources.

➢ *Knowledge synthesis:* Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.

➢ *Segmenting Web pages and detecting noise:* In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years.

9. APPLICATIONS

1. Get structured information from unstructured website content.

2. Discover implicit knowledge from explicit knowledge.

3. Text mining and Natural Language Processing features of language to discover knowledge Make conclusions.

4. Get interesting content of website such as navigation, ads, headers.

5.To find out, which parts repeat on other sites and which are unique.

6.Analysis of written text (content).

7.To study the structure of HTML andgeneralize it as much as possible.

8. Spam Mail Filtering, more than 50% of email traffic is spam, so we need efficient methods to filter spam but not throw away the needy content.

9. Cloud users need to extract the information from the cloud provided by web servers can utilize the web mining.

10. Online shopping systems use the web mining to extract the information of a product and its specification through web mining.

11. Opinion mining is the process of extracting reviews of a customer about the product and its specification using mining techniques.

12. 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.

13. Web wide tracking is effectively done using web mining methodologies.

14. 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].

10. CONCLUSION

Web content mining in information technology enhances the information provided on utility sites to be structured. This allows for a customer of the Web site to access specific information without having to search the entire site. With the use of this type of mining, data remains available through order of relativity to the query, thus providing productive marketing. The main purpose of web content mining is to gather, organize, categorize and provide the user with the best possible information that is available on World Wide Web [5]. This paper has discussed about the research issues in web mining and also provided detailed review about the basic concepts of web mining and web content mining. Several open research issues and drawbacks which are exists in the current techniques are also discussed. This study and review would be helpful for researchers those who are doing their research in the domain of web mining. The future scope of web content mining is to predict the user needs to improve the usability and scalability.

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.

[15] M. Junker, M. Sintek, and M. Rinck. Learning for text categorization and information extraction with ilp. In Proceedings of the Workshop on Learning Language in Logic, 1999.

[16] H. Kargupta, I. Hamzaoglu, and B. Stafford. Distributed data mining using an agent based architecture. In Proceedings of Knowledge Discovery And Data Mining, pages 211–214. AAAI Press, 1997.

[17] U. Y. Nahm and R. J. Mooney. A mutually beneficial integration of data mining and information extraction. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00), 2000.

[18] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61–67, 1999.

[19]S. Scott and S. Matwin. Feature engineering for text classification. In Proceedings of the 16th International Conference on Machine Learning ICML-99, 1999.

[20]S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1-3):233–272, 1999.

[21]S. M. Weiss, C. Apt´e, F. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. IEEE Intelligent Systems, 14(4):63–69, 1999.

[22]W. Wiener, J. Pedersen, and A. Weigend. A neural network approach to topic spotting. In Proceedings of the 4th Symposium on Document Analysis and Information Retrieval (SDAIR 95), pages 317–332, 1995.

[23]I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan. Text mining: A new frontier for lossless compression. In Data Compression Conference, pages 198–207, 1999.

[24] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14(4):32–43, 1999.
K. M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. In Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI98), pages 509–516, 1998.

[26]F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe. T´etrafusion: Information discovery on the internet. IEEE Intelligent Systems, 14(4):55–62, 1999.

[27] J. F¨urnkranz. Exploiting structural information for text classification on the www. In Advances in Intelligent Data Analysis, Third International Symposium, IDA-99, pages 487– 498, 1999.