International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 8, August 2014

Survey on Web Content Mining and Its Tools

T. Shanmugapriya[1], P. Kiruthika[2]

[1]M.E. Scholar, Department of Computer Science, SNS College of Engineering, Coimbatore, Tamil Nadu 641035, India
[2]Assistant Professor, Department of Computer Science, SNS College of Engineering, Coimbatore, Tamil Nadu 641035, India

Abstract: ***The task of searching, downloading, filtering related document is said to be web content mining. It is the discovery of useful information from the web content, including text, images, audio, video, etc,. Web content mining becomes complicated when it has to mine structured, semi-structured and unstructured and multimedia data. Unstructured Data Mining Techniques are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering, and Information Visualization. Structured Data Mining Techniques are Web Crawler, Wrapper Generation, Page Content Mining. Semi-Structured Data Mining Techniques are Object Exchange Model (OEM), Top down Extraction, Web Data Extraction Language.***

Keywords: Web content mining, Unstructured Data Mining, Structured Data Mining, Semi-Structured Data Mining

1. Introduction

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Us-age data. a brief overview of the three categories.

1.1. Web Content Mining: Web Content Mining is the process of extracting 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, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.

1.2. Web Structure Mining: The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used are Hyperlinks and Document structure.

1.3.Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to under-stand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered are Web Server Data, Application Server Data, and Application Level Data.

2. Web Content Mining

The Web content data consist of unstructured data such as free texts, semi-structured data such as HTML documents, and a more structured data such as data in the tables or database generated HTML pages[1].Web content extraction is concerned with extracting the relevant text from Web pages by removing unrelated textual noise like advertisements, navigational elements, contact and copyright notes. Web crawling involves searching a very large solution space which requires a lot of time, hard disk space and lot of usage of resources. The research done in Web content mining from two different points of view: IR and DB views. IR view is mainly to assist or to improve the information finding and filtering the information to the users usually based on either inferred or solicited user profiles. DB view mainly tries to model the data on the Web and to integrate them so that more sophisticated queries other than the keywords based search could be performed. The three types of agents are Intelligent search agents, Information filtering/Categorizing agent, Personalized web agents. 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. In Database approach it consists of well formed database containing schemas and attributes with defined domains. The algorithm proposed is called Dual Iterative Pattern Relation Extraction for finding the relevant information used by search engines. The content of web page includes no machine readable semantic information. Search engines, subject directories, intelligent agent, cluster analysis and portals are employed to find what a user must look for.

2.1. Web Document Clustering: Web Document Clustering is another approach to finding relevant document on atopic or about query keywords. The user could apply clustering to a set of documents returned by a search engine in response to a query with the aim of finding semantically meaningful clusters, rather than a list of ranked documents, that are easier to interpret. K-means and agglomerative method are used for web document cluster analysis [1].Suffix Tree Clustering uses a phrase-based clustering approach rather than using single word frequency.STC algorithm which consists of document cleaning, identifying base clusters and combining base clusters.

2.2. Finding similar web pages: It has been found that almost 30% of all web pages are very similar to other pages and about 22% are virtually identical to other pages. Similarity between Web pages usually means content-based similarity. It is also possible to consider link-based similarity and usage-based similarity. Link based similarity is related to concept of co-citation and is primarily used for discovering a core set of web pages on a topic. Usage-based similarity is useful in grouping pages or users into meaningful groups. We define two concepts to finding similar web pages are Resemblance and Containment.

2.3. Finger printing: This is one approach to finding similar web pages. In one recent approach it has been suggested that not only the content be compared but the expression also be compared [2].An approach for comparing a large number of documents is based on the idea of finger printing documents. A document may be divided into all possible substring of length L. These substrings are called Shingles. The web is very large and this algorithm [38] requires enormous storage to store the shingles and very long processing time to finish pair wise comparison for say even 100 million documents. Sometimes this approach is called full finger printing.

2.1.1. Unstructured Data Mining Techniques:

1. Information Extraction: pattern matching is used to extract information from unstructured data[3]. 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.

2. Topic Tracking: 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 applied in the case of mining unstructured data.

3. Categorization: This technique counts the number of words in a document 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.

4. Clustering: Clustering is a technique used to group similar documents. Same documents can appear in different group. Clustering helps the user to easily select the topic of interest.

2.1.2. Structured Data Mining Techniques:

1. Web Crawler: Crawlers are computer programs that traverse the hypertext structure in the web.

2. Wrapper Generation: In Wrapper Generation, it provides information on the capability of sources. 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.

2.1.3. Semi-Structured Data Mining Techniques:

1. Object Exchange Model (OEM): A main feature of object exchange model is self describing; there is no need to describe in advance the structure of an object.

2. 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.

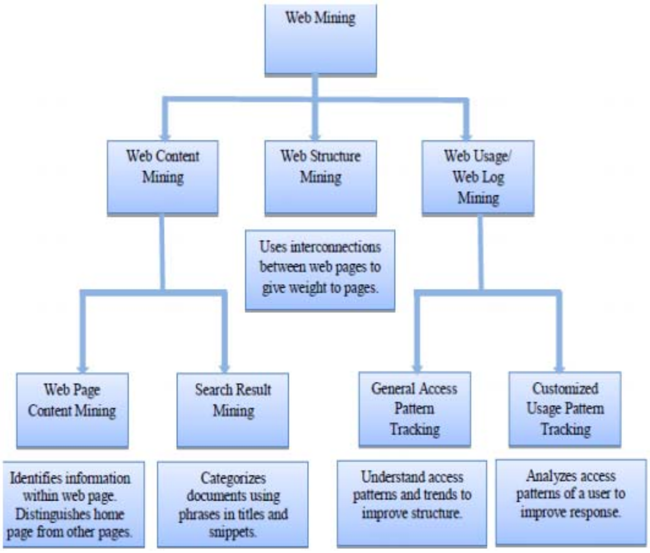
Web Mining
Web Content Mining
Web Structure Mining
Web Usage/ Web Log Mining
Uses interconnections between web pages to give weight to pages.
Web Page Content Mining
Search Result Mining
General Access Pattern Tracking
Customized Usage Pattern Tracking
Identifies information within web page. Distinguishes home page from other pages.
Categorizes documents using phrases in titles and snippets.
Understand access patterns and trends to improve structure.
Analyzes access patterns of a user to improve response.

Figure 1: Web Mining Taxonomy

3. Web Content Mining Tools

Web content mining tools helps to download the essential information. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and Web Content Extractor, Rapid Miner.

1. Rapid Miner: Rapid Miner is open source software and it is a tool for extracting information from web, Contains inbuilt algorithm. It can generate algorithm by itself.

Features:

- Easy to use.
- Reduce time.
- Open source software.

2. Screen-scaper: Screen-scraping is a tool for extracting/mining information from web sites [12]. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper. Features: Screen-scraper present a graphical interface allowing the user to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data. Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked. This also facilitates scraping of information at periodic intervals. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time, after which the results are displayed in a single interface.

3. Automation Anywhere: It is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining [13].

Features:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly. Web record and Web data extraction.

4. Web Info Extractor: This is a tool for data mining, extracting Web content, and Web content analysis. It can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

Features:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file
- Can deal with Web page in all language
- Running multi-task at the same time
- Support recursive task definition.

5. Mozenda: This tool enables users to extract and manage Web data [4]. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, the data to be used in other applications or as intelligence. There are two parts of Mozenda's scraper tool:

i.Mozenda Web Console: It is a Web application that allows user to run agents, view & organize results, and export publish data extracted.
ii.Agent Builder: It is a Windows application used to build data extraction project.

Features:

- Easy to use.
- Platform independency. However, Mozenda Agent Builder only runs on Windows.
- Working place independence.

6. Web Content Extractor: It is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet[14].It offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner. This tool allows users to extract data from various websites such as online stores, online actions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source.

Features:

- Helps to extract/collect the market figures, product pricing data, or real estate data.
- Helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- Assists users in automate extraction of auction information from auction sites.
- Assists to Journalists extract news and articles from news sites.
- Helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- Extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

4. Conclusion and Future Work

For the survey, it focus on the opportunity to analyze Web data and extract all manner of useful knowledge from it. The Web Data Mining tools are primordial to scanning the many HTML documents, images, and text provided on Web pages. Web mining is promising as well as challenging, and this field will help produce applications that can more effectively and efficiently utilize the Web of knowledge. The different types of web content mining tools are discussed and features are discussed. Mining using Rapid Miner is in undergone for study. All the tools except rapid miner are examined based on usability. Hence the future work will be to examine the usability of rapid miner.

References

[1] Overview of Web Content Mining Tools, The International Journal of Engineering And Science (IJES), Volume 2, Issue 6, 2013, ISSN: 2319 – 1813 ISBN: 2319 – 1805.
[2] Web Mining - Concepts, Applications & Research Directions, Jaideep Srivastava, Prasanna Desikan,

Vipin Kumar. University of Minnesota, Minneapolis, MN 55455, USA

[3] Web Content Mining, The 14th International World Wide Web Conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[4] Mozenda, http://www.mozenda.com/web-mining-software Viewed 18 February 2013.

[5] Web Content Extractor help. WCE, http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013

[6] Bharanipriya, V., Prasad, V.K., Web Content Mining Tools: A comparative Study, International Journal of Information Technology and Knowledge Management. Vol. 4, No 1, pp. 211-215 (2011).

[7] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, “Web Mining Concepts, Applications and Research Directions”.

[8] P. Boldi and S. Vigna, “The WebGraph framework I: Compression techniques,” in Proc. 13th Int’l. World Wide Web Conference. Manhattan, USA: ACM Press, 2004, pp. 595–601.

[9] Martin Ester, Hans-Peter Kriegel, Mattis Schubert “Accurate and efficient Crawling for relevant Websites”

[10] Bing Liu, Kevin Chen Chuan Chang,”Editorial Issue on Web Content Mining”, issue2, 2004.

[11] Jaraslav Porkorny “Page content Rank: An approach to web content mining”.

[12] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.

[13] Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013

[14] Web Info Extractor Manual. WIE, http://webinfoextractor.com/wiedoc.htm Viewed 19 February 2013

Author Profile

T. Shanmugapriya completed Bachelor of Engineering in Sri Subramanya College of Engineering and Technology, Palani, India. Pursuing Master of Engineering in SNS College of Engineering. Area of Interest is Web Mining and Cloud Computing. Project and research work will be based on Web Mining.