International Journal of Research Instinct
(Peer-reviewed Journal)

ISSN ONLINE 2348-2095
Research Article

A Survey on Web Content Extraction Techniques

[1]D. Saravanan, [2]N. Sugavaneswaran

[1&2]*Asst. Professor in Computer Science,Srimad Andavan Arts and Science College,Trichy-5*

[1]*dmsaro@gmail.com*, [2]*sugavaneswara@rediffmail.com*

ABSTRACT

The World Wide Web has rich source of voluminous and heterogeneous information which continues to expand in size and complexity. It contains an enormous and valuable content of textual or multimedia form. Many Web pages are unstructured and semi-structured, so it consists of noisy information like advertisement, links, headers, footers, etc. Web pages may contain text, images and videos. This noisy information makes extraction of web content tedious. There are many techniques proposed to extract whereas data mining cannot be applied directly. For effective retrieval of Web information, Web mining is used. Many applications get benefits from the extracted content such as crawlers, indexers, document classification, and Information retrieval. This survey aims at providing a comprehensive overview of many approaches that are constructed for extracting Webpage content.

Keywords:

Web Mining, Web Content Extraction, Unstructured text Extract , Structured Extract, Semi-Structured text Extract and Multimedia Extract.

1. INTRODUCTION

Today, there are billions of HTML documents, images and other media files on the Internet. Taking into consideration the wide variety of Web, the extraction of interesting content has become a necessity. A solution to this problem is to use data mining techniques. The term Data Mining literally means, “Drilling data “. As with any drilling, its purpose is to extract an element of knowledge. The World Wide Web has grown explicitly which provides access to all people at any place and at any time. It facilitates any one to upload or download relevant data and the valuable content in the Web site can be used in all fields . The data in the Web are unstructured and semi-

structured, lots of insignificant and irrelevant document are obtained as a result after navigating several links. For effective retrieval of Web information, Web mining is used.

2. WEB MINING

Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

2.1 Web usage mining - Web usage mining is process of discovery of user access profile patterns from web server logs, which maintain history of each user when browsing the internet. In sever maintain logs containing information about the each user profile, list of pages accessed, time of accessing pages, user interested pages etc. This kind of information is used to maintain web site in an effective and efficient manner. Also finding the path from (Uniform Resource Locator) URL to last URL, associated list of web sites visited.

2.2 Web structure mining - Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Web structure mining is a tool used to identify the relationship between Web pages linked by information or direct link connection. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Techniques of web structure mining:

1. PageRank: this algorithm is used by Google to rank search results. The name of this algorithm is given by Google-founder Larry Page. The rank of a page is decided by the number of links pointing to the target node.

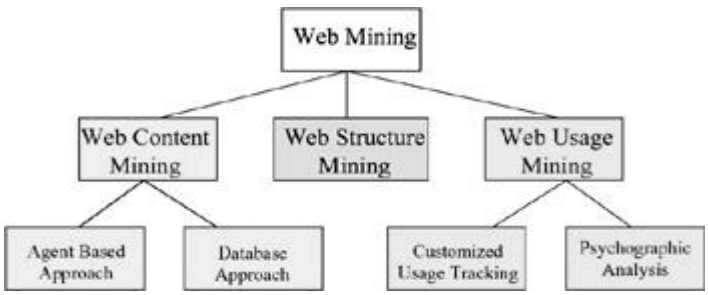

Fig. 1 Types of Web Mining

2.3 Web Content Extraction

Web content extraction is the mining, extraction and integration of useful data, information and knowledge from Web page content. It is permitting much of information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web. There are two types of approaches used in web content mining such as Agent based approach and Database approach.

i) **Information agent** – In Information agent agents searches the information for information according to a particular query using domain characteristics and user profiles. It used number of techniques to filter data according to the predefine information.

ii) **Database approach** - In Database approach consists of well formed database and it containing schemas and attributes with defined domains.

Web content Extraction has the following approaches to extract or mine data (1) Unstructured text extract, (2) structured extract, (3) Semi- structured text extract, and (4) Multimedia extract.

2.3.1 Unstructured Text extract – In most of the web content data is of unstructured text data format. Content mining requires application of data mining and text mining techniques. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT). Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.

a) Information Extraction – In Information extraction software identifies key phrases and relationships within text. It works for predefined sequences in text, a process called pattern matching. This technology can be used in large volumes of text.

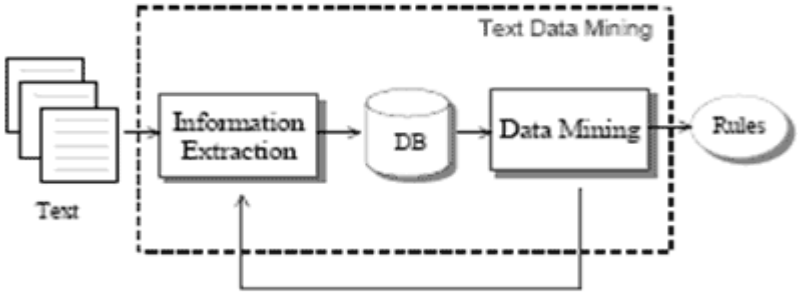

Fig. 2 Text mining

b) TopicTracking - Topic tracking is one of the technologies. It can be used in the text mining process. The main purpose of topic tracking is to identify and follow events presented in multiple news sources, radio and TV broadcasts. It tracks the information together and makes it easy for user to get a general understanding.

c) Summarization - Text summarization is one of the another technology. It can be used to summarize the lengthly document. Text summarization software works with summarizes the text and it take the user to read the first paragraph.

d) Categorization - Text categorization is also known as text classification, or topic spotting. Text categorization is the task of automatically sorting a set of documents into categories from a predefined set.

e) Clustering - Clustering is a technique. It is used to group similar documents. That means the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses are to be minimized.

f) Information visualization - Information visualization or information visualisation is the study of visual representations of abstract data to reinforce human cognition. Information visualization mappings the large amounts of text, allowing the user to visually analyze the content. The user can interact with the document by creating sub-maps, zooming and scaling, Information visualization is useful when a user needs to narrow down a broad documents.

2.3.2 Structured Extract - Structured data is very easier to extract when compared to unstructured text document. The following techniques are used to extract the from unstructured to structured text document such as Web Crawler, Wrapper Generation, Page content Mining.

1. Web Crawler - A Web crawler is software.It is also be called a Web spider. The main function of this software is to update their web content or indexes of others sites' web content.

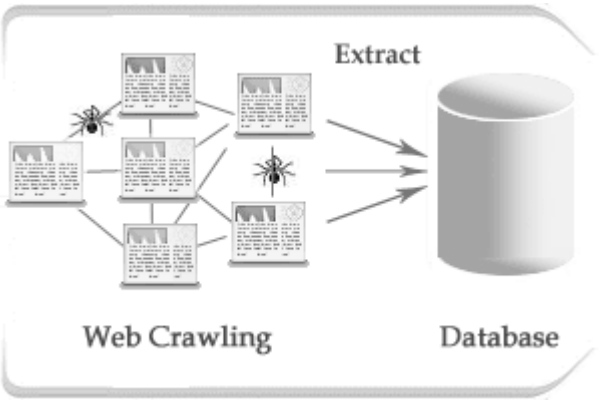

Fig.3 Web Crawler

Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

2. Wrapper Generation – Wrapper in data mining is a program that extracts content of a particular information source and translates it into a relational form. Many webpages presents structured data and it typically descriptions of objects retrieved from underlying databases and displayed in Web pages following some fixed templates. There are two main approaches to wrapper generation such as wrapper induction and automated data extraction. Wrapper induction uses supervised learning to learn data extraction rules from manual trained examples and automated extraction is possible because most Web data objects follow fixed templates.

3. Page content Mining - Page Content Mining is structured data mining technique. It works on the pages ranked by traditional search engines. That means it classifies the pages.

2.3. 3 Semi- Structured Text Extract - Semi-structured text extract data from structured relational tables with numbers and strings and to restore some kind information structure that has been lost through publication such as:

i) Table extraction: To find and extract tables from documents.

ii) Comments extraction : To Extract comments from actual content of article in order to restore the link between each sentence.

4. MULTIMEDIA EXTRACT

The techniques of Multimedia data extract are;

a) SKICAT b) Color Histogram Matching c) Multimedia Miner and d) Shot Boundary Detection.

5.CONCLUSION

World wide web has become one of the most significant resources nowadays. It brings the information mainly in the form of web pages. It may contain informative contents as well as non-informative contents. The non-informative contents like advertisements, header, footer, copyright statements, etc are called noisy parts. It has been proved that almost 40-50% contents are these types of noisy data. Web mining is an application of data mining technique to extract informative contents from non-informative contents. The advantage of eliminating non-informative content will saving in storage and indexing. This paper describes various methods for extracting web information from the huge volume of data present in world wide web. As future work, research is to

be continued on alternative method for extracting core contents from web pages

REFRENCES:

[1] S.Mahesha, Dr. M S Shashidhara, and Dr. M. Giri, "An Implementation of Web Content Extraction Using Mining Techniques" IFRSA International Journal of Data Warehousing & Mining |Vol 2|issue4|November 2012.

[2] Erdinç Uzun, Hayri Volkan AgunTarık Yerlikaya," A hybrid approach for extracting informative content from web pages". E. Uzun and al. / Information Processing and Management 49 (2013) 928–944 .

[3] Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona, "Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it" O. Arbelaitz and al. / Expert Systems with Applications 40 (2013) 7478–7491

[4]S.S. Bhamare, Dr.B.V., "Survey on Web Page Noise Cleaning for Web Mining", International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013.