Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Content Mining Techniques - A Review

N. Sathyapriya
Assistant Professor, Department of Information Technology,
Sri G.V.G Visalakshi College for Women (Autonomous),
Udumalpet, Tirupur (dt). Tamil Nadu, India

Abstract---The World Wide Web serves as a huge, widely distributed, global information service centre for news, advertisements, financial management, education, government, e-commerce and many other information services. The web also contains a rich and dynamic collection of hyperlink information and web page access and usage. With more information on WWW it has become necessary to apply some strategy so that valuable knowledge can be extracted and consequently returned to the user. Data mining concepts and techniques when applied to WWW with its existing technologies are known as web mining. Web mining is mining of data related to WWW. The paper brings in light the value of Web Content Mining. The paper gives an insight into web content mining and its techniques as a review. Web content mining is the scanning and mining of text, pictures and graphs of web page to determine relevance of content to the search query.

Keywords---Web Content, Web Mining, Structured, Unstructured, Semi structured, Multimedia.

I. INTRODUCTION

The World Wide Web has lot of information and continues to increase in size and complexity. It is very herculean task to search relevant information from huge amount of data. Web is becoming a challenge. Web data is updated at every second. When any specific keyword or any web page is searched, number of links or result is displayed. But all the data which is displayed on the web is not relevant. So efficient and effective techniques are required to retrieve relevant data. The relevant data can be retrieved by techniques like,

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

A. Web Content Mining

It is related to text mining because much of the web contents are text based. Text mining focuses on unstructured texts. It is the mining, extraction and integration of useful data, information and knowledge from web page content. Web content mining is semi - structured nature of the web. It describes the discovery of useful information from the web documents. In web content mining, the content may be text, image, audio, video, metadata and hyperlinks etc. Web content mining also distinguishes personal home pages with other web pages.

B. Web Structure Mining

This kind of mining emphasizes on the data which describes the structure of the content. It tries to discover useful knowledge from the structure and hyperlinks.

Web structure mining can be viewed as creating a model of the web organization or a portion thereof. This can be used to classify web pages or to create similarity measures between documents.

C. Web Usage Mining

Web usage mining performs mining on web usage data, or web logs. It refers to the discovery of user access patterns from the web usage logs. A web log is a listing of page reference data. Sometimes it is referred to as click stream data because each entry corresponds to a mouse click.

II. WEB CONTENT MINING TECHNIQUES

Web content mining can be thought of as extending the work performed bybasic search engines. There are many different techniques that can be used to search the internet.

- Unstructured Data Mining Technique
- Structured Data Mining Technique
- Semi Structured Data Mining Technique
- Multimedia Data Mining Technique

A. Unstructured Data Mining Technique

The web page is in the form of text. in this technique, the data is searched and retrieved. The data may be unknown information. It is related to text mining because much of the web contents are text based. Text mining focuses on unstructured texts. Retrieval of information from HTML web pages is a challenging task, since HTML web pages have multiple tags and the web pages are highly unstructured.

1) Topic Tracking:

It is a technique by which a registered user can track the topic of his/her interest. It checks the document viewed by the user and tries to locate other related documents. Topic tracking is generally used by registered site like yahoo.

The ads that are displayed after one logins are related to the subject of mails that are received. Any advancement done anywhere is notified to the registered user. A drawback in this technique is sometimes the user is not provided with desired information.

2) Summarization:

It is used to reduce the length of the document or multiple documents into a short set of words or paragraph that conveys the meaning of the text. It helps the user to decide whether the topic is of his/her interest or not.

Two methods:

Extractive: It works by selecting a subset of existing words, phrases or sentences in the original text to form the summary.

Abstractive: It builds an internal semantic representation and then use natural language generation technique to create a summary that is closer to what a human might generate.

3) Categorization:

It is the technique of categorizing the document. It counts the number of words in a document then it decides the main topic from the count. it ranks the document according to the topic. Document having majority content on a particular topic are ranked first.

4) Clustering:

The documents are categorized using categorization technique. Same document can appear in different groups. The problem of finding best grouping can be handled by clustering. There are various clustering algorithms used to select the topic of interest from the best relevant grouping.

B. Structured Data Mining Technique

Structured data are the data records retrieved from underlying database and displayed in the web pages. It can be displayed either as tables or forms. Data can be extracted from these sources using structured data extraction technique.

1) Web Crawler:

It is an automated program or script that scans or crawls through internet pages to create an index of the data it is looking for. Search engines frequently use web crawler to collect information about what is available on public web page. Their primary purpose is to collect data so that when internet surfers enter a search term on the site, it quickly provides the surfer with relevant web sites.

2) Wrapper Generators:

To facilitate effective search on the world wide web. Several meta search engines have been formed which do not do the search themselves but take help of the available search engines to find the required information. Meta search engines are connected to search engine by the means of wrappers.

C. Semi Structured Data Mining Technique

Semi structured data arises when the source or environment does not impose a rigid structure on the data when data is combined from several heterogeneous sources. For example Bibliographic data.

1) Top Down strategy:

Complex objects are extracted by decomposing them into less complex objects until atomic objects have been extracted. Through this technique, a couple of examples are sufficient for extracting hundreds of objects on a new webpage. it works by traversing the structure of example object in preorder form visiting all its components and concatenate them to form new resultant object.

2) Wrapper:

It uses the extractor to retrieve the relevant data in OEM (Object Exchange Model) format, and then executes the query at the wrapper.

D. Multimedia Data Mining Technique

It can be defined as the process of finding interesting patterns from media data such as audio, video, image and text that are not accessible using queries. It is to use the discovered patterns to improve decision making.

1) Image Mining:

It focuses on detecting abnormal patterns as well as retrieving images. it discovers meaningful information or image patterns from a huge collection of images. It includes digital image processing, image understanding, database, Artificial intelligence.

2) Video Mining:

It is to find the interesting patterns from large amount of video data. Multimedia data is video data such as text, image and metadata, visual and audio The processing are indexing, automatic segmentation, Content based retrieval etc.

3) Audio Mining

It is a technique by which the content of an audio signal can be automatically searched, analyzed and rotten with wavelet transformation.

III. CONCLUSION

The paper concludes that different technique are required as data available on web is not homogeneous. Using various techniques the structured, unstructured, semi structured data can be searched and the relevant data can be retrieved. Apart from being unstructured, structured or semi structured the data present on web can be in any form. It may be in text form, audio, image or in video form. All these mining techniques are in infancy stage. By these techniques searching of contents over the web is faster and exact.

REFERENCES

[1] Faustina Johnson and Santosh kumar gupta, *Web Content Mining Techniques:A survey*, International Journal of Computer Applications, june 2012.
[2] Govind Murari, Upadhyay, Kanika Dhingra,*Web Content Mining: Techniques and uses*, Nov 2013, IJARCSSE.
[3] Darshna Navadiya, Roshni patel, *Web Content Mining Techniques - A comprehensive Survey*, IJERT,Dec 2012.
[4] Monika yadav, Mr. Pradeep mittal,*Web Mining : An introduction*,IJARCSSE, March 2013.
[5] Margaret H.dunham, *Data Mining introductory and advanced topics*, Pearson Education,2009.
[6] Jiawei Han and Micheline kamber ,*Data mining Concepts and Techniques*, Morgan Kaufmann publishers, edition -2.
[7] Arun K Pujari, *Data Mining Techniques*, University press, Edition 2001.

