**International Journal of Research in IT and Management (IJRIM)**
Available online at : http://euroasiapub.org/current.php?title=IJRIM
Vol. 6 Issue 7, July 2016, pp. 127~132
ISSN(o): 2231-4334, ISSN(p): 2349-6517| Impact Factor: 5.96 | Thomson Reuters ID: L-5236-2015

# Optimization of Data Retrieval Through Web Mining

**Dr. Jatinder Kumar**

***Keywords:***

Data Mining,
Extraction and Retrieval,
Optimization

***Abstract***

The evolution of the World Wide Web has brought enormous and ever growing amounts of data and information. It influences almost all aspects of people's lives. In addition, with the abundant data provided by the web, it has become an important resource for research. Furthermore the low cost of web data makes it more attractive to researches. Researchers can retrieve web data by browsing and keyword searching. In order to extract and manage the need based data, the tools and techniques of data mining are proving to be boon for the data analysts. The techniques of web mining are used to optimize the data extraction process. This paper throws light on the process through which the data retrieval from web mining can be optimized.



***Authors Correspondence***

***First Author***
**Assistant Prof. , A . S. College, Khanna**

The ever growing size of data bases is making it hard for the researchers and decision makers to extract the data as relevant for their analysis. 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 cannot 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 1 contains both relevant and irrelevant information, it should be able to analyze the date patterns of the collected data and help to build models and provide validity.

Wed usage mining is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web. The usage data records the user's behavior when the user browses or makes transactions on the web site. It is an activity that involves the automatic discovery of pattern from are or more Web servers. Organizations often generate and collect large volumes of data, most of this information is usually generated automatically by web servers and collected in server log. Analyzing such data can help these organizations to determine the value of particular customers, cross marketing strategies across products and the effectiveness of promotional campaigns, etc.

The first web analysis tools simply provided mechanisms to report user activity as recorded in the servers. Using such tools, it was possible to determine such information as the number of accesses to the server, the times or time intervals of visits as well as the domain names and the URLs of users of the Web server. However, in general, these tools provide little or no analysis of data relationships among the accessed files and directories within the Web space. Now more sophisticated techniques for discovery and analysis of patterns are emerging. These tools fall into two main categories, Pattern Discovery Tools and Pattern Analysis Tools.

The evolution of the World Wide Web has brought us enormous and ever growing Amounts of data and information. With the abundant data provided by the web, it has become an important resource for research. 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. However, traditional data extraction and mining techniques cannot be applied directly to the web due to its semi- structured or even unstructured nature.

This proposal describes the design and planned implementation of web mining research support system. This system is designed for identifying extracting, filtering and analyzing data from web resources. This system is composed of several stages. Information Retrieval (IR), information Extraction (IE), Generalization, and Analysis & Validation. The goal of this system is to provide a

general solution, which researchers can follow to utilize web resources, un their research. Some methods such as Natural Language Processing (NLP) and Artificial Neural Networks (AMM) will be applied to design new algorithms. Furthermore, data mining technologies such as clustering and association rules will also be explored for designing and implementing the web mining research support system. 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 extracted data and validate their correctness. The proposed system will be tested on a NSF sponsored Open Sources Software study. Our proposed work offers an integrated set of web mining tools that will help advance the state of the art in supporting researchers doing online research. The proposed research work will provide a general-purpose tools set which researchers can employ to utilize web resources in their research.

-- Web content mining

Web content mining is the process to discover useful information from the content of a web page. The type of the web content may consist of text, image, audio or video data in the web. Web content mining sometimes is called web text mining., because the text content is the most widely researched area. The technologies that are normally used in web content mining are NLP (natural Language Processing) and IR (Information Retrival).

--**Web Structure Mining-** Web structure mining is the process of using graph theory to analyse the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds.

The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyse and describe the HTML (Hyper Text Markup Language) or XML (eXtensible Markup Language) tags within the web page.

Web is a collection of inter- related files on one or more Web servers. Web mining is the application of data mining techniques to extract knowledge from Web data. Web data is

- Web content- text, images, records, etc.
- Web structure- hyperlink, tags, etc.
- Web usage- http logs, app server logs, etc

**Web Content Mining-** Web Content Mining is the process of extracting useful information from the contents of Web documents. Contents 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 a s lists and tables. Research activities in this field also involve using techniques from other disciplines such as information Retrieval (IR) and natural language processing (NLP)

--Pre processing Content

## Web Content Mining Applications

- Identify the topics represented by a Web Documents
- Categorize Web Documents
- Find Web Pages across different servers that are similar
- Queries- Enhance standard Query Relevance with User, Role, and/or Task Based Relevance.
- Recommendations- List of top "n" relevant documents in a collection or portion of a collection
- Filers- Show/Hide documents based on relevance score.

## What is Web Usage Mining?

A Web is a collection of inter- related files on one or more Web servers. But Web Usage Mining is Discovery of meaningful patterns from data enervated by client- server transactions on one or more Web localities.

Typical Sources of Data is automatically generated data stored in server access logs referrer logs, agent logs and client- side cookies, user profiles meta data page attributes content attributes usage data.

## Distributed Web Mining

Motivation : Data on the web is huge and distributed across various sites.

| | |
|---|---|
| Traditional Approach : | Integrate all data into one site and perform required analysis. |
| Problem : | Time consuming and not scalable. |
| Solution : | Analyze data locally at different locations and build an overall model. |
| Application | Personalization of Web Sites depending on user's life on the web (the users interests, locations and behavior across different sites). |

Web Services- what they provide

- Once a web service is deployed, other applications (and other web services) can discover and invoke the deployed services.
- Web services is also viewed as a important interoperability mechanism for the JEE and Microsoft's. NET worlds to come together.
- Services that follows from this
  - ✓ Messaging (e.g. SOAP, XML)
  - ✓ Description (e.g. WSDL, CML Schema)
  - ✓ Discovery (e.g. UDDI)
  - ✓ Security (e.g. TLS. SSL)

Web Data collected at the client and server level can help in better performance and providing better features for Web Services, Understanding of client- server interactions the data from the interactions can be mined for analyzing interesting patterns personalization of Web Services. The client level data can provide information to personalize Web services for the users. It also helps in fraud analysis.

## *References*

(01) C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah, Knowledge discovery from user web- page navigation. In Workshop on Research Issues in Data Engineering

(02) R. Kumar, P. Raghavan, S. Rajagopalan, and A Tomkins, Trawling the Web for emerging cyber- communities. Computer Networks
(03) S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large scale knowledge bases from the web.
(04) J. Srivastava Data preparation for mining World Wide Web browsing patterns. Knowledge and information Systems.
(05) J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data.
(06) S.K. Pal, V Talwar, and P. Mitra Web mining in soft computing framework Relevance.
(07) K. Snadhu, and M. Shih, Clustering of web users based on access patterns.
(08) S. Soderland. Learning information extraction rules for semi- structured and free text. Machine Learning.