ISSN (Online) : 2249 - 054X

International Journal of Computing and Corporate Research

Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I2M1-032013

VOLUME 3 ISSUE 2 March 2013

AN INAUGURAL STUDY OF WEB SEARCHING AND MINING

Tamanna Jain

M.Tech. (CSE) Student

Maharishi Markandeshwar University, Mullana, Haryana, India

ABSTRACT

In the present stratagem, web mining is the most progressive area where the exploration is going on promptly. Excavating from the web is the right solution for knowledge exploration on the web. The **W**orld **W**ide **W**eb is a capacious and diversified source of information and therefore fetching information is today's one of the most critical venture. Most of the people use Internet for retrieving the information, but many times it happens that they get lots of insignificant and irrelevant content even after navigating several links that makes the surfer frustrated. To avoid this frustration of the user, one of the best solutions for eliciting information from the web is the web mining techniques that are used to extract the useful and relevant data as per the requirements of the needy user. This manuscript expounds the endowments of web, web page, search engines together with the needs and the problems of web mining. Web mining is an application of data mining where the information retrieval is based on the relevancy of the content, hyperlinks and the usage web records.

KEYWORDS: Data mining, Search Engine, Web, Web Mining, Web Page

1. INTRODUCTION

The World Wide Web or www or W3 commonly known as web is *a collection of interlinked hypertext documents attainable via Internet*. One can get each and every kind of information in this mesh of knowledge. Users always want to get the appropriate and updated information within a short span of time but because of the bulky nature of the web, it becomes quite annoying for the user to search, fetch, filter and validate the relevant information. The following characteristics of the web spawn the searching more difficult:

1) **Bulk:** the amount of the information available on the web is very huge.

2) **Diverse:** the coverage of content is very wide and diverse.

3) **Growth:** size of the web is growing exponentially. It is estimated that each day 1-7 millions of new pages are added up on the web.

4) **Duplication:** availability of the redundant data that is composed of 30% of the data over web.

5) **Queries:**generally search engine queries are short, quite cryptic due to synonyms and homonyms.

6) **Hyperlinks:** a normal text document contains the connotations but the Web documents contain hypertext links to the other web documents. It has been approximated that about half of the traffic on the web is users navigating using links.

7) **Index Pages:** some results from the search engines return the index pages from various sides providing little content but many links.

8) **Dynamic:** the Web changes considerably with time. The link structure of the Web itself changes rapidly as the new links are entrenched and the existing ones extirpated.

9) **Demanding Users:** users are very anxious, demanding and intolerant. They want the results in few seconds otherwise they will make a move to some other search engines having all kinds of information including the latest news and technology.

1.1 WEB TERMINOLOGY

a) *Web:* set of nodes connected via hypertext links.

b) *Link:* depicts the relationship between two or more resources. A link is ingrained within a document using anchor tags.

c) *Web Page:* collection of information, consisting of one or more web resources, diagnosed by a single URL.

d) *Web site:* collection of inter linked web pages, including a home page, residing at the same network location

e) *URL*: an identifier for a physical resource, for instance, an index or file path. URL's are location dependent.

f) *Web server*: serves web pages using http to client machines so that a browser can display them.

g) *Client:* user or an application retrieving a Web resource.

h) *Proxy: an* agent which acts as a server as well as a client for the purpose of retrieving information on behalf of other clients.

i) *DNS:* a distributed database of the name to address mappings.

j) *Cookie:* data sent by the web server to a web client, to be stored locally and sent back to the server on subsequent requests.

k) *Information Pull:* obtaining information from the search engine.

l) *Information Push:* sending information to the users.

2. SEARCH ENGINE

A search engine is basically a software code that is designed to search the information on web. It is really a general class of programs that search documents for specified keywords and returns a list of all documents where the keywords were found. To find the specific information on the vast expanse of the World Wide Web, search engines are one of the best key. There are basically three types of search engines:

i) Powered by robots

ii) Powered by human submissions

iii) Hybrid of the above two

Search engines are huge databases of web pages as well as software packages for indexing and retrieving the pages that enable users to find the information of interest to them.

Querying a search engine involves the user specifying a (few) number of keywords that are used by the search engine to search its indexes to find relevant information.

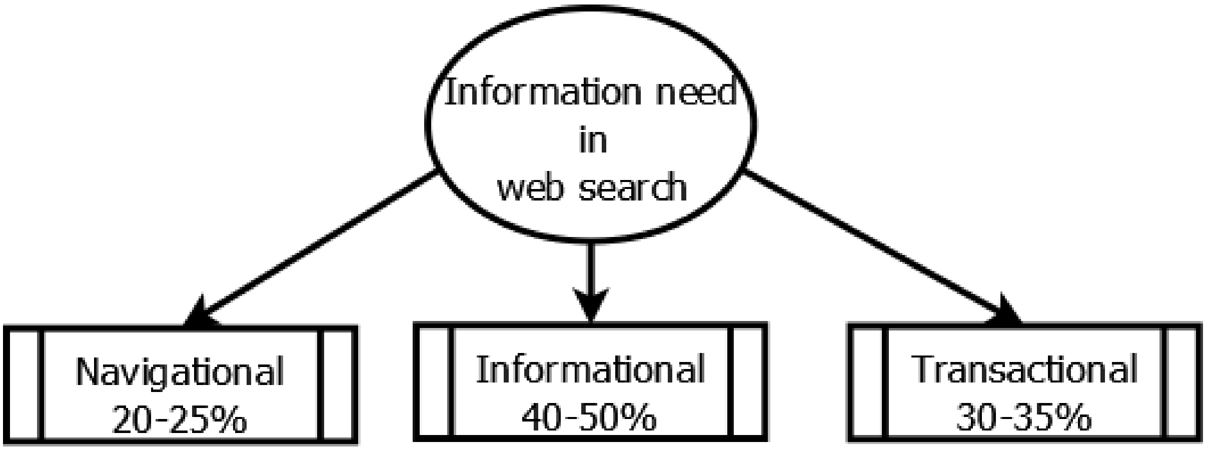

Fig 1.1 Estimation of information need

Navigational: the primary information used in these queries *to reach a web site* that the user has in mind.

Informational: the primary information need in such queries is *to find a web site* that provides useful information about a topic of interest.

Transactional: the primary need in such queries is *to perform some kind of transaction*.

The quality of search results from a search engine ideally should satisfy the following requirements:

i) Precision

ii) Recall

iii) Ranking

iv) First screen

v) Speed

2.1 SEARCH ENGINE TERMINOLOGY

- **Spider trap:** a condition of dynamic web sites in which a search engine's spider becomes trapped in an endless loop of code.
- **Meta tag:** a special HTML tag that provides information about a web page.
- **Deep link:** a hyperlink either on the web page or in the results of a search engine query to a page on a web site other than the site's home page.
- **Robot:** a program that runs automatically without human intervention.

2.2 SEARCH ENGINE FUNCTIONALITY

A search engine is a collection of software modules that is designed to carry out a variety of tasks. Some of those tasks are listed as:

- ***Collecting information***
- ***Evaluating and categorizing information***
- ***Creating a database and indexes***
- ***Computing ranks of web documents***
- ***Checking queries and executing them***
- ***Presenting results***
- ***Profiling users***

3. WEB MINING

Web mining is an application of ***data mining*** *techniques to find charismatic and potentially advantageous information from* ***web data****.*

Data Mining is the process of analysing data from different panorama and epitomizing it into useful information.

- *Data:* facts, numbers or texts that can be processed by a computer.
- *Information:* patterns, associations, or relationships among all the data that can provide information.
- *Knowledge:* information can be converted into knowledge about historical patterns and future trends.

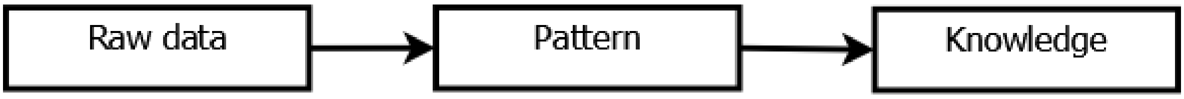

Fig 3.1 data mining

Web Data is a wide term that is used for managing data online. The categorization of web data is as follows:

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

Web mining is a *multidisciplinary* field:

1. Data mining
2. Machine learning
3. Natural language processing
4. Statistics
5. Databases
6. Information retrieval
7. Multimedia

4. PROBLEMS IN WEB MINING

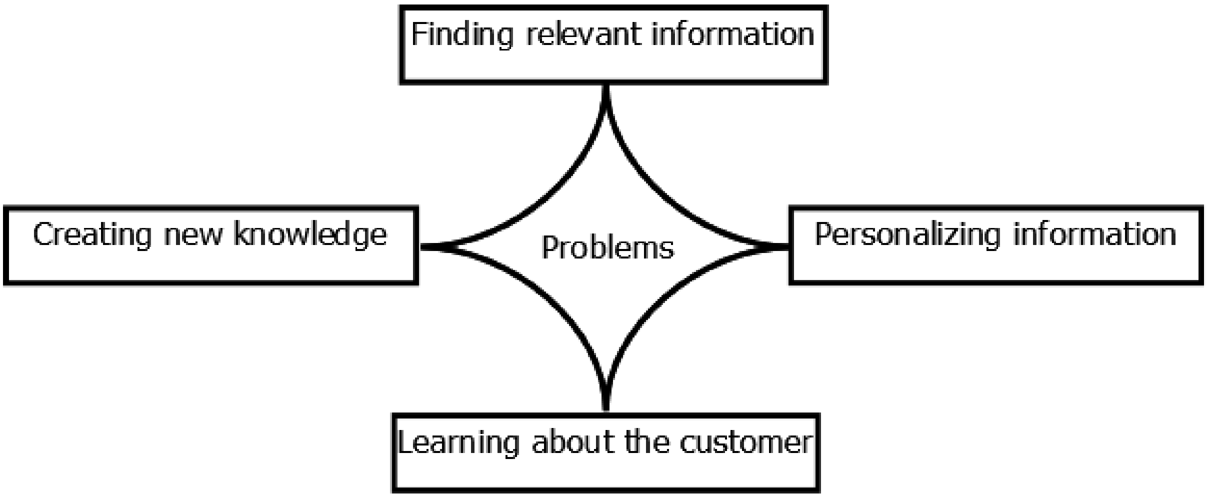
Finding relevant information
Creating new knowledge
Problems
Personalizing information
Learning about the customer

Fig 4.1: problems in web mining

- Finding relevant information: Low precision and un-indexed information.
- Creating new knowledge: building new information out of the available information on web.
- Personalizing information: catering to personal preference in content and presentation.
- Learning about the customer: to know more about the customers as what does the customer wants to do? And how to use web data to effectively market products and/or services.

5. WEB MINING PROCESS

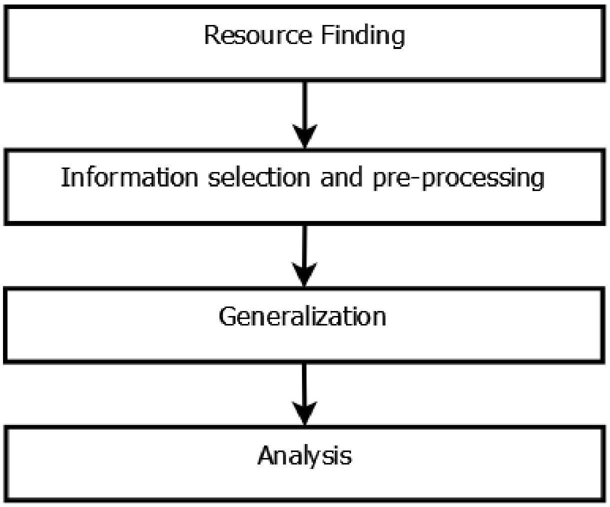

Fig 5.1 web mining process

Web mining process can be disintegrated into four subtasks:

1. **Resource finding:** the task of finding the data that is either online or offline from the text sources available on the web such as electronic newsletters, electronic newswires and also all the manual selection of web resources.

2. **Information selection and pre-processing:** selecting and pre-processing specific information from selected documents of web resources.

   Challenges in pre-processing:

Problem	Solution
Who are the users	IP vs Real people
How long did users stay	Measuring session time
Where did users go	Server side vs client side
What did the users view	Content processing

Fig 5.2 challenges in pre-processing

3. **Generalization:** automatically recognizing general patterns at individual web sites and across multiple sites. Techniques to be used in generalization:

a. Data mining

b. Machine learning

c. Natural language processing

d. Neural network approach

e. Statistics

4. **Analysis:** affirmation and assimilation of the mined patterns.

6. WEB MINING TECHNIQUES

Depending on the classification of web data, web mining can be classified as shown in the fig 6.1.

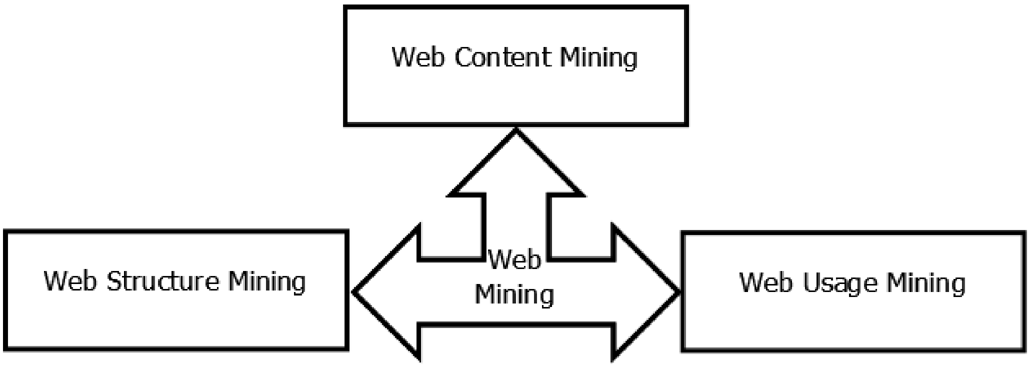

Fig 6.1 web mining techniques

1. ***Web content mining:***mining, extraction and integration of useful data, information and knowledge from web pages content.

2. ***Web structure mining:*** discovery of useful information from the structure of hyperlinks.

3. ***Web usage mining:*** discovery of user access patterns from web usage log records.

Web mining usually have two approaches namely:

- Content based approach: the system searches for the items that match based on the analysis of content using the user preference.
- Collaboration approach: the system tries to find users with similar results and the endorsements based on what dissimilar users did.

REFERENCES

1. URL: http://en.wikipedia.org/wiki/Web_mining
2. URL: http://www-users.cs.umn.edu/~desikan/publications/wmo.pdf
3. URL: https://cs.uwaterloo.ca/~tozsu/courses/cs748t/surveys/wang.pdf
4. URL: http://www.cs.uic.edu/~liub/WebContentMining.html
5. URL: http://iosrjournals.org/iosr-jce/papers/Vol5-Issue4/F0543136.pdf