International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 5, September-October 2014 **ISSN 2278-6856**

WEB MINING: DAY-TODAY

[1]K. Mohammad Mujahid, [2]Mr. I.S.Raghuram, [3]M. Niranjan Kumar, [5]K.V. Chaitanya krishna [6]T. Mohaneshwar

Dept. Of Computer Science and Engineering
G.Pullaiah college of Engineering and Technology, Kurnool

Abstract
In this paper we study and present facts about how to extract the useful information on the web and also give the superficial knowledge and comparisons about data mining. 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. This paper describes the current, past and future of web mining.

Keyword:- *Mining Categories,* Web mining subtasks, Mining Text databases, Keyboard-based retrieval, Invested index, Example explanation of Google Search, how it works

1.Data Mining
Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, dataware houses, or other information repositories. Data mining is often defined as finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning. Data mining involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis.

2.WEB MINING
Web Mining is based on knowledge discovery from web. It is extract the knowledge framework represents in a proper way. Web min g is like a graph & all pages are node & each connects with hyperlinks. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. By using web mining easily extract all features and information about multimedia before this web mining difficult to extract information in proper way from web. We search the any topic from web difficult to get accurate topic information but Now's day it is easy to get the proper information about any things. Web mining is based on data mining technique by using data mining technique discover the hidden data in web log. Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research. Based on the aforesaid four subtasks from above figure, web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. Here, evaluation includes both "generalization" and "analysis.

Fig: Web Mining Processing

Web Mining Categories:
Web mining can be categorized in to three area of interest based on which part of the web to mine:
a) Web Content Mining
b) Web Structure Mining
c) Web Usage Mining

a) Web Content Mining
Web mining is basically extract the information on the web. Which process is happen to access the information on the web. It is web content mining. Many pages are open to access the information on the web. These pages are content of web. Searching the information and open search pages is also content of web. Last accurate result is defined the result pages content mining.

b) Web Structure Mining
We can define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's shown the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages/webs. It is shown the link one web page to another web page.

c) Web Usage Mining
It is discovery of meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of inter related files on one or more web servers. It is automatically generated the data stored in server access logs, refers logs, agent logs, client sides cookies, user profile, meta data, page attribute, page content & site structure. Web mining usage aims at utilize data mining techniques to discover the usage patterns from

web based application. It is technique to predict user behavior when it is interact with the web.

Web usage mining is categories in three phases:-

Preprocessing- According to client, server and proxy server it is first approach to retrieves the raw data from web resources and processed the data .it is automatically transformed the original raw data.

Pattern Discovery- According the data preprocessing discovered the knowledge and implements the techniques to discover the knowledge like as machine learning and data mining procedures are carried out at this stage.

Pattern Analysis- pattern analysis is the process after pattern discovery. Its check the pattern is correct on the web and how to implement on web to extract the information on your web search / extract knowledge from the web.

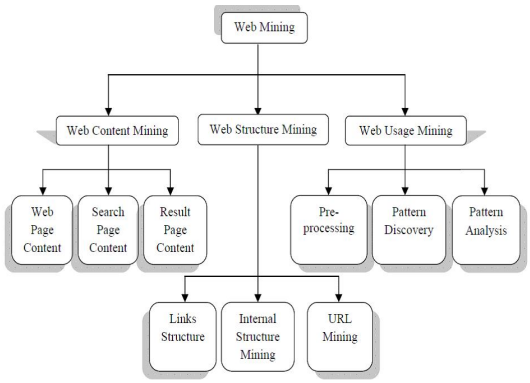


Fig: Web Mining Structure

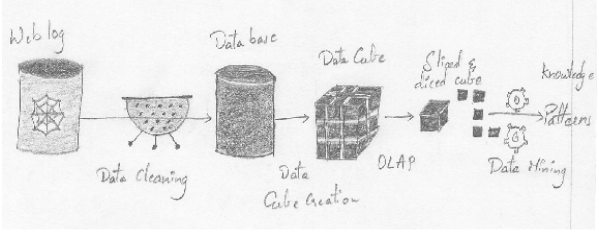
Web mining subtasks

1. Resource finding
2. Task of retrieving intended web documentation
3. Information selection and pre process
4. Automatic selection and preprocessing specification from retrieving various resources
5. Generalization
6. Automatic discovery of patterns in web sites
7. Analysis
8. Validation and interpretation of mined patterns

3.Mining Text databases

Text databases:

Large collection of documents from various sources: news articles, research papers, books digital libraries, e- mail messages &web pages, library database, etc. Data stored is usually semi structured. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data.



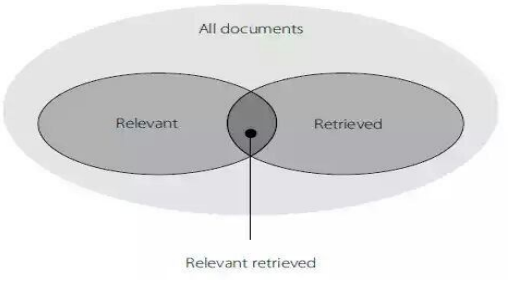
4.Information retrieval

A field developed in parallel with database systems. Information is organized into documents; Information retrieval consists of a Typical IR systems & Online library catalogs & document management systems. Information retrieval v/s database systems Some DB problems are not present in IR, e.g. update transaction management, complex objects. Some IR problems are not addressed well in DBMS.

Information retrieval problem:

Locating relevant document based on user input, such as keywords or example documents.

Example: Unstructured documents, approximate search using keywords & relevance.



Precision the percentage of retrieved documents that are in fact relevant to the query

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall the percentage of documents that are relevant to the query & were, in fact retrieved

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Keyboard-based retrieval:

A document is represented by a string. Which can be identified by a set of keywords .Queries may use expression of keywords. E.g. car & repair shop, tea or coffee. DBMS but not Oracle Queries & retrieval should consider synonyms, e.g. repair & maintenance

Method

Create a term frequency matrix, frequency- matrix SVD construction; compute the singular valued decomposition of freq- matrix by splitting it into 3 matrices, U,S,V

Vector identification: For each document d replace its original document vector by a new excluding the eliminated terms.

Index creation: store the set of all vector indexed by one of a number of techniques other text retrieval indexing techniques

Invested index

Maintains two hash : or B+-tree indexed.

Document _table : a set of document records

<doc id, postings _list>

Term_ table a set of document term records

<term, posting _ list>

Answer query: find all doc associated with one or a set of terms

Advantage: easy to implement

Disadvantage: Do not handle well synonymy & polysemi, & posting lists could be too long.

Signature file:

Associate a signature with each document .A signature is a

representation of an ordered list of terms that describe the document. Order is obtained by frequency analysis, stemming & stop lists.

5.WEB SEARCH ENGINE

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

6.Web Search-Google

Google is one of the most popular and widely used search engines. It provides users access to information from over 2 billion web pages that it has indexed on its server. The quality and quickness of the search facility makes it the most successful search engine. Earlier search engines concentrated on web content alone to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining information from the web. Page Rank, which measures the importance of a page, is the underlying technology in all Google search products, and uses structural information of the web graph to return high quality results. The Google toolbar is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. The full version of the toolbar, if installed, also sends the click-stream information of the user to Google. The usage statistics thus obtained are used by Google to enhance the quality of its results. Google also provides advanced search capabilities to search images and find pages that have been updated within a specific date range. Built on top of Netscape's Open Directory project, Google's web directory provides a fast and easy way to search within a certain topic or related topics.

7.Page Rank Algorithm

Page Rank is a numeric value that represents the importance of a page present on the web. When one page links to another page, it is effectively casting a vote for the other page. More votes implies more importance. Importance of the page that is casting the vote determines the importance of the vote. Google calculates a page's importance from the votes cast for it. Importance of each vote is taken into account when a page's Page Rank is calculated. Page Rank is Google's way of deciding a page's importance. Page Rank Notation is "PR" The original Page Rank algorithm which was described by Larry Page and Sergey Brin is given by

PR(A)=(1-d + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Where, PR(A) – Page Rank of page A PR(Ti) – Page Rank of pages Ti which link to page A C(Ti) - number of outbound links on page Ti d - Damping factor which can be set between 0 and 1 A simple way of representing the formula is, (d=0.85) *Page Rank (PR) = 0.15 + 0.85 * (a share of the Page Rank of every page that links to it)*

The amount of Page Rank that a page has to vote will be its own value * 0.85. This value is shared equally among all the pages that it links to. Page with PR4 and 5 outbound links > Page with PR8 and 100 outbound links. The calculations do not work if they are performed just once. Accurate values are obtained through much iteration. Suppose we have 2 pages, A and B, which link to each other and neither have any other links of any kind. Page Rank of A depends on Page Rank value of B and Page Rank of B depends on Page Rank value of A. We can't work out A's Page Rank until we know B's Page Rank, and we can't work out B's Page Rank until we know A's Page Rank. But performing more iterations can bring the values to such a stage where the Page Rank values do not change. Therefore more iterations are necessary while calculating Page Ranks.

Text based algorithms

The text based algorithm consist of many ways of approaching data through mining. The following are few ways of approaches

1. A vision-based page segmentation algorithm
2. Discovering conceptual relations from text
3. Spectral voice conversion for text-to speech synthesis
4. A new benchmark collection for text categorization research
5. Ontology learning for text
6. Word sence disambiguation

References

[1] Web Mining: Today and Tomorrow - 3rd International Conference on Electronics Computer Technology (ICECT 2011)
[2] Web Mining: An Overview - CVPR Unit, Indian Statistical Institute, Kolkata.
[3] Web Mining: Accomplishments & Future Directions - Jaideep Srivastava, University of Minnesota, USA
[4] Web Mining - Concepts, Applications, and Research Directions by Jaideep Srivastava, Prasanna Desikan, Vipin Kumar

Authors

Mr. I.S. Raghuram received his M.Tech degree in Computer Science and Engineering from JNTU Hyderabad (A.P), India. He was a Lecture, Assistant Professor with Department of CSE, GPCET, Kurnool. His research interests include Data Mining, Networks and Mobile Computing.

K.MD. Mujahid received his B.Tech degree in Information Technology from JNTU Anatapur (A.P), India. He studied in GPCET, Kurnool.

M. Niranjan Kumar received his B.Tech degree in Information Technology from JNTU Anatapur (A.P), India. He studied in GPCET, Kurnool.

K.V. Chaitanya krishna received his B.Tech degree in Information Technology from JNTU Anatapur (A.P), India. He studied in GPCET, Kurnool.

T. Mohaneshwar received his B.Tech degree in Information Technology from JNTU Anatapur (A.P), India. He studied in GPCET, Kurnool.