ISSN: 2250–3676

[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-6, 1653 – 1656

AN APPROACH OF DEEP WEB MINING FOR DATA EXTRACTION

Shaheen Parveen[1], Ajay Kushwaha[2]

[1]*Research Scholar, CSE Department, RCET Bhilai, C.G, India,* ***shaheenparveen6@gmail.com***
[2]*Head of Department, CSE Department, RCET Bhilai, C.G, India,* ***kushwaha.bhilai@gmail.com***

Abstract

The Web mining extracts useful information from the web pages. Web mining techniques seek to extract knowledge from Web data, including web documents, hyperlinks between documents, and usage logs of web sites. Web usage mining mines knowledge from diverse websites. Extracting appropriate data from deep web pages is an exigent dilemma due to the overflow of data in to the web. Web servers generates a huge amount of information on web users browsing activities. These are called click stream or web access log data. The click stream data can be enriched with information about the content of visited pages.

The aim of this research paper is to develop a practically implemented search engine in which it extracts only the relevant links by analyzing user behaviour. With the continuous growth of Web services the user data collected by Web based organizations has reached enormous capacity. The paper will undertake a review of the existing literature available on this arena and develop an empirical model of search engine showing real time data flow after retrieval of significant information from data warehouse.

Index Terms: *Web Mining, Web Usage Mining, Click stream, Web Server Logs, Web Data Extraction*

----------*****----------

1. INTRODUCTION

The Web is the universal information space that can be accessed by companies, governments, universities, students, teachers, businessmen and some users. Web mining is the region which uses Data mining technique. Web sites has a lot of consistent web pages that are developed and maintained by an organization. This is the most direct link a company has its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. Web mining is the process of discovering and analyzing the useful information from the World Wide Web. The Web can be viewed as the largest unstructured data source available, although the data on the Web sites, which composed them, is structured. This presents an exigent task for effective design and access to Web pages. Web mining is a term used for applying data mining techniques to Web access logs. It is useful in personalized web pages, for web search, user tracking, understanding of user behaviour, decision making etc.

Based on the different emphasis and different ways to obtain information, web mining can be divided into three categories: Web content mining, Web structure mining and Web usage mining.

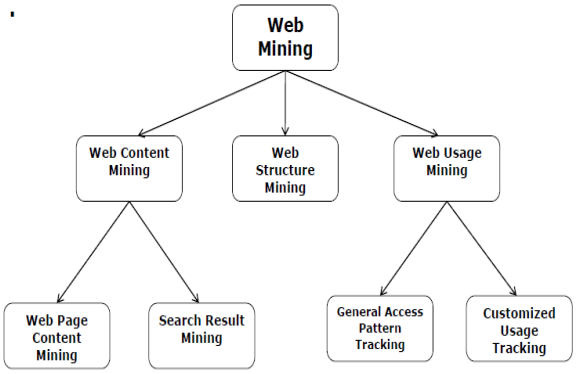


Fig. 1

Web Contents Mining is the process of extracting knowledge from documents and content description through search engines. For example:-Crawl a web site containing data such as product prices and descriptions that you want to mine. Web structure mining is the process of obtaining knowledge from the organization of the Web and the links between Web pages. For example:-Google's Page rank. Web Usage Mining is the process of discovering and analyzing of user access patterns, through the mining of log files and associated data from a

particular Web site. For example: - Number of visitors, Popularity like music, products, movies etc.

2. LITERATURE REVIEW

Since the long time Web usage mining extract knowledge from different web site. Web usage mining extract knowledge from the web log file but the problem is these log files are private and the companies doesn't share their private information to public. If log files aim publically everyone get equal opportunity to get identical commercial value in their businesses. Furthermore, many companies providing Application Program Interface (API) for different services so Pardeshi and Patel (2012) undertook a study to propose an API which store private log information in central web mining server and from there general people can access private log information. They proposed an API which gathers user behaviour information and is helpful for many companies like Google, Amazon, eBay, PayPal. When used in the context of web development on API is defined as set of HTTP request message, along with a definition of the structure of response messages, which is an Extensible XML or Java Script Object Notation (JSON) format, this service enable direct access from client side. The main challenge of the project is to collect user behaviour information from all web site to the central database and from their information will be available to all people. They developed Google Chrome Add-ons for the browser in order to gather the information as easy as possible. Each time when user click a link in web page or write an URL, the chrome extension send the information to the central database. This log information will be helpful for business which improve their website and to get equal commercial value like other top rated web sites.

Naveena Devi and Sreevani (2010) studied the prediction of user's browsing behaviour which is useful for personalization like building proper web site, improving marketing strategy, forecasting market trends, and increasing the competitive strength of enterprises etc. Devi and Sreevani wanted to implement a system which is used for open web resources practical e-business data sets within a short period of time to achieve dynamical and statistical analysis result to the enterprise. In this paper, the main area of research is pre-processing and classification of useful patterns from web data using mining techniques. It is divided in to three dependent tasks:- Pre-processing, pattern discovery and pattern analysis.

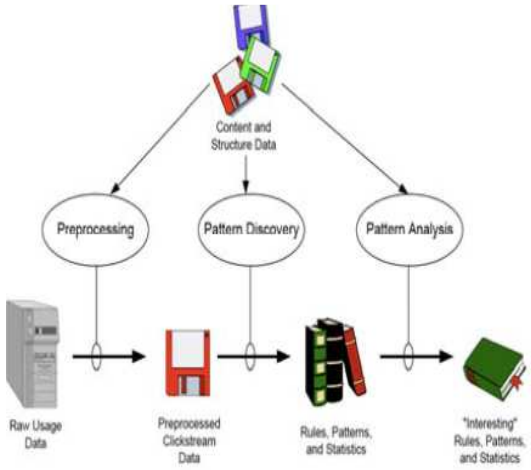


Fig. 2

The users suffer from huge information on internet, they have to filter irrelevant information by themselves. Data used for mining is collected from web servers all of which generate noisy data so data cleaning methods are necessary. After practised with APIs they design their value added business and then integrate web APIs into their websites to perform selected business operations.

Nagone, Kapse and Bhagwat (2011) aim to build an ecommerce web application using APIs. Ecommerce advice system directly interacts with users to find their own needed goods and complete the purchase process. This paper, analyses the commonly-used methods and then proposes an improved Apriori-Based Personal Recommendation Algorithm for E-commerce. It provides system features are secure registration for users, secure Login for a user, central control over the information for administrator, maintaining database. The software they are designing is helpful for vendors and customers.

Chen, Chau and Tseng (2009) studied that it is vital for students to acquire knowledge and hands-on experience in Web mining. In this paper, design a Web mining application using the open Web APIs provided by Google, Amazon, and eBay. The concept of Web API enables direct access to modules from the client side. They suggested four examples Wish sky, Tucson Book Exchange, Cell phone Intelligent Auctioning (CIA), and Scribble. Wish sky enables customers to monitor the news and ongoing auctions related to the products. Tucson Book Exchange is an online book exchange system. CIA is a cell phone auction history analysis system which provides value-added service to eBay cell phone buyers

ISSN: 2250–3676

[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-6, 1653 – 1656

and sellers. Scribble is a science fiction book portal which helps customers find their books of interest. They create interesting business models and integrate necessary system components to implement them.

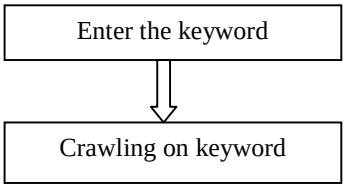
Claudia Elena (2011) used data mining techniques and researches that click stream or web access log data better understand and characterize web users. The goal of their project is to analyse user behaviour by mining enriched web access log data. The focus of this paper is how to use frequent pattern techniques for discovering different types of patterns in a Web log database. Click□stream means a sequence of Web pages viewed by a user. Analysis of clicks is the process of extracting knowledge from web logs. This analysis involves data pre-processing and then applying data mining techniques. Log files contain information about: domains, sub domains and host names; resources requested by the user, time of request, protocol used. Analysis is based on accurate information and quality data so pre-processing plays an important role. One of the popular data mining techniques is the association rules or frequent item sets mining algorithm. Items that occur often together can be associated to each other and these together occurring items form a frequent item set. It is based on two main steps: candidate generation and pruning. In this paper we have seen the techniques of Association and sequence mining can be applied to the web usage mining task. This information can be helpful for site developer for customer satisfaction. So, the web site designers can determine the web pages and bring them to the right position and in right time.

3. PROPOSED METHODOLOGY

The work is having three major parts first is web surfer search engine second is training of dataset last is extracting relevant data after mining from large data warehouse.

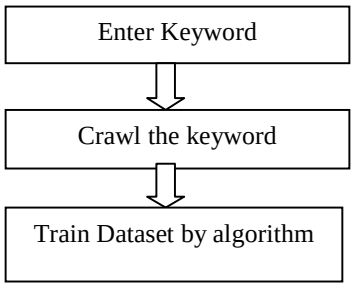
3.1 Web surfer search engine

In this proposed method, search engine will provide only relevant link to the customers. At the time of crawling, the word gets divided and the sentence will be split into tokens and each token is assigned with consecutive numerical value. Meanwhile, each word is matched up in web and then frequency check will decide whether that particular keyword exist in web or not



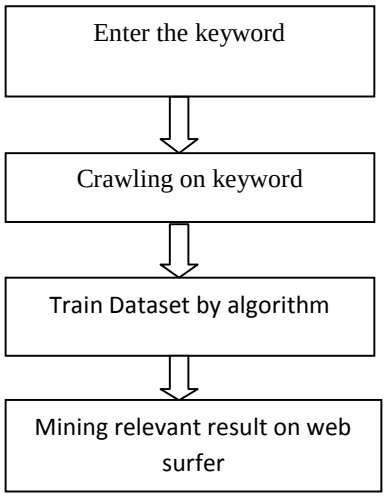
3.2 Training Data Set

In this stage the set of data will train and find out all the related links and then store it into a large data warehouse. The training data set is the major part of this stage. Training data set has a vital role in the identification of useful data.



3.3 Extracting Relevant Data after Mining

This is the final and most important step of our work in which the relevant data is extracted from a huge set of data stored in data warehouse. The focus will be on mining the data from data warehouse and fetch the relevant links in to the search engine.



4. PROPOSED OUTCOME

At this stage the proposed outcome is keyword which user will enter the keyword in search engine that keyword goes to the web, fetch all the related links and stored in to the large data warehouse. Then from there the essential data will be extracted with the help of data mining techniques and user will get the useful data as output.

REFERENCES

[1] Claudia Elena Dinuca, Association and Sequence Mining in Web Usage, Economics and Applied Informatics, 2011.
[2] Hsinchun Chen, Xin Li, Michael Chau, Yi-Jen Ho, Chunju Tseng, Using Open Web APIs in Teaching Web Mining, ACM, 2009.
[3] Sachin Pardeshi, Ujwala Patil, Central web mining services–public and free access log files, WJST, 2012.
[4] B.Naveena Devi, O.Sreevani, Dynamic Modelling Approach for Web Usage Mining Using Open Web Resources, IJEST, 2010.
[5] Sanket Nagone, Bharat Kapse, Mayur Bhagwat, E-Commerce Application using Web API and Apriori Algorithm of Data Mining, IJCA, 2011.
[6] Claudia Elena Dinuca, The process of data pre-processing for Web Usage Data Mining through a complete example, Annals of the "Ovidius" 2011.