ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Enhanced Indexing and Scraping for Educational Search Engine using Web Usage Mining

[1] Ramkrishna R. Gaikwad,[2] Mansi Bhonsle,
[1] PG Scholar, [2] Assistant Prof
[1][2] Computer Engineering Department, G H Raisoni COEM, Pune Maharashtra, India.

Abstract – **Nowadays the growth of World Wide Web has better a lot with more assumption. Large amount of text, multimedia files, images website documents were present in the web and it is still increasing in its forms. Education Search engine has become an important daily network application tool to search information. Data mining is the form of extracting data present in the internet. We propose an Education Search Engine in two-stage technique, namely Smart Crawler, for efficient gathering deep web interfaces. To achieve more accurate results for a focused crawl, Smart Crawler ranks websites links to prioritize highly relevant result in websites link rankings. In the second stage, Web Usages mining in web scraping is a method for extracting textual characters from screens so that they could be analyzed. Web scraping is the process of collecting information from the World Wide Web. The results showed that the smart crawler and scarper can realize the high-efficient and flexible data collection function, and laid the foundation for Web data mining. This efficiently retrieves web data mining interface from large-scale sites and achieves higher.**

Keywords- Information extraction, web crawler, web usage mining, web scraping

I. INTRODUCTION

Search engine optimization is a means to obtain a better ranking of the search results via a variety of approaches to manage the website to make it more compliance with the principles of search engine ranking. The goal is to increase the exposure of website and high flow rate. When obtaining a higher ranking in the result of search engine, the webpage can have a significant incremental click through rate. A search engine is an information retrieval software program that discovers crawls, transforms and stores information for retrieval and presentation in response to user queries. The intelligent dynamic crawlers, changed the traditional crawler will be data extraction rules in the process of curing the drawbacks, making the rule settings dynamic, at the same time, TF-IDF method is used to calculate the Web document correlation, and the automatic acquisition of data extraction rules is realized, which reduces the development cost and maintenance cost and improves the development efficiency of the crawler [1]. A search engine normally consists of four components that are search interface, crawler, indexer, and database. A search interface is a single entry point providing access to the content of an index. It allows users to send queries as well as display, sort and save results. In organize the institutions into three clusters, and the preferred K-means clustering algorithm successfully did meet our goals. We can see this from the Figures 3, 4 and 5 respectively. Therefore, this paper sheds some light on how the limitation of web 2.0 featuring unknown knowledge discovery may be reduced. Moreover, showing irrelevant search results can also be decreased using our proposed technique [2]. Online search engines store images, link data and metadata for the document as well. Indexers enable objects to be indexed in a similar manner to arrays. A get access or returns a value. A set access or assigns a value. The keyword is used to define the indexer. The value keyword is used to define the value being assigned by the set indexer. Databases are collections of links, keyword and descriptions e.g. schemas, tables, queries, reports, views, and other elements.

'Web search' might be simply explained as a search for information on the World Wide Web, which is a collection of interconnected web pages. The pieces of information are gathered with searching tools such as directories and web search engines. They search through a large number of websites and then provide relevant results within few seconds. Search systems are based on single box-and-button search screens, which guarantee simple usage. However, a particular piece of information is not necessarily easily found in all cases. This depends on the kind of information, quality of the web search engine and searchers' abilities to formulate their queries. Searching tools are divided into human-powered directories and crawler-based search engines. They are not used separately, while search engines also provide directories and vice versa. However, one of the tools is generally preferred. For example Google is mainly a web search

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

engine, but Google Directory might be used as a subsidiary Google product. On the other hand, the directories are mostly compiled as portals, i.e. gateway sites, i.e. they are not 'pure' directories. Portals are designed to offer not only directories but also a variety of tools such as search engine, news etc.

Type of Search Engine

A. Human-powered directories

The Its also provides a characterization of a directory as "a classified listing of web sites, in which brief records for sites are placed within an appropriate hierarchical taxonomy". In other words, the web sites are classified into categories and further subcategories by human editors. People search the Web, review and categorize the sites by their topics. The titles and the web site descriptions are taken into consideration when choosing the right category too.

At the same time, human selection might be paradoxically considered as both an advantage and a disadvantage of directories. People might evaluate the web site more precisely and then reduce the number of discrepancies such as duplications of listings, broken links and out-of-date information. On the other hand, human-powered categorizing does not enable to include the whole web site but only one main page and therefore the database is thousand times smaller than the ones created by web search engines. Because of the size of the database, directories are suitable and also used for the 'local' Internet only.

Directories are labeled 'libraries', where people do not know the 'title of a book' or 'subject guides', because their content is divided into the subject areas and subdivisions. Users browse the category names and simultaneously specify their search. Thus directories are used when the searchers have a notion of what they are looking for but they do not know what exactly they should enter into the search box or when they have only a general query.

B.Crawler-based search engines

The crawlers, or spiders, scan the Internet and identify new pages. After the identification, the page is indexed under virtually every word on the page, the URL, metatags, the URLs of links etc. The indexing program then retrieves related pages in the database according to the user's query. It also determines the order of the results while following a relevance-ranking algorithm. Although the collections of data are compiled by crawlers twenty-four hours a day, the results might not be up-to-date, while robots visit the one particular web site approximately once a month. New pages are identified from newly registered domains and from the links on already existing web pages. Nevertheless, the spiders do not 'visit' every web page. The crawling systems ignore hidden, small and difficult-to-access web pages. The connectivity to the web page and channel overload plays an important role as well. If the connectivity is low and the channel is overloaded, the pages will not be crawled. In addition, worldwide servers concentrate more on pages rather written in English than in other languages. However, the statistics of ignored web sites do not exist. In comparison to directories, web search engines are more suitable for detailed search and their database contents billions of pages. Web search engines search not only the titles and the web site descriptions but also 'full text' of the site. The websites are indexed by robots not by humans and thus there may appear some of the above mentioned discrepancies.

C. Meta Web Search

Meta web search engine provides collective searches. It is a searching tool that does not have its own database and uses other web search engines and sends them the query at the same time. The results are then compiled on a single web site or in multiple frames or windows. The result duplicities should be deleted. The number of web search engines that are used by Meta search varies between a few and hundreds of engines. It might seem that Meta web search could settle the debate over using several web search engines, e.g. Iskar recommends gathering and comparing results from number of the search engines. Never the less, Meta engines are used less than regular web search engines because of the prevailing number of their negatives. First, the major web search engines do not support Meta searches.

Web Mining

Web mining is an application of data mining techniques to find information patterns from the web data. Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents. The contents of data mined from the Web may be a collection of facts that Web pages are meant to contain, and these may consist of text, structured data such as lists and tables, and even images, video and audio.

A. Web Content Mining

Web content mining can be used for mining of useful data, information and knowledge from web page content.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

It is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. Techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.

B. Web Usage Mining

Web usage mining is used for mining the web log records that is access information of web pages and helps to discover the user access patterns of web pages. Analysis of similarities in web log records can be useful to identify the potential customers for e-commerce companies.

C. Web Structure Mining

The web structure mining can be used to discover the link structure of hyperlink. It is used to identify that the web pages are either linked by information or direct link connection. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected.

II. RELATED WORK

In past few years, there has been an increasing amount of literature on websites recommendation systems, which focuses on providing accurate recommendations for website to be used by a users or customers.

The mining of web data still be present as a challenging research problem in the future. Because the web documents possess numerous file formats along with its knowledge discovery process. There are many concepts available in Web Mining but this paper tried to expose the Web content mining strategy and explore some of the techniques, tools in Web Content mining [3].

The extraction of web information and web crawler technology are studied. According to the characteristics of the management of books and documents, an analysis system is designed to parse article for the selected document search website. Through setting the uri of the document or a specified journal, the system can automatically extract information in the corresponding page. The results show that the method of information extraction by HtmlParser is accurate and efficient and the outputs meet the requirements of archive [4].

A method based on Process Mining and Model Driven Engineering has been proposed in order to improve the security of Web information systems. Besides, the proposed method has been applied to the SID BiD case study, and some promising preliminary results have been obtained. Despite the fact that the experimental results have been obtained by considering a particular technology to build Web information systems (in particular, JEE), the proposed approach is technology independent [5].

Web scraping is related to web indexing, whose task is to index information on the web with the help of a bot or web crawler. Here the legal aspect, both positive and negative sides are taken into view. Due to a vast community and library support for Python and the beauty of coding style of python language, it is most suitable for Scraping data from Websites. As Scraper opens up another world of retrieving information without the use of API, and mostly it is anonymously accessed [6].

Web Structure Mining based on the analysis of patterns from hyperlink structure in the web. Like as Data Mining, Web Mining has four stages i.e. Data Collection, Preprocessing, Knowledge Discovery and Knowledge Analysis. This paper based on the first two stages Data collection and Preprocessing. Data collection is to collect the data required for analysis. Data preprocessing is considered as an important stage of Web Structure mining because of data available on web is unstructured, heterogeneous and noisy. Achieved the Information system avoid the affection of redundant data and reserved the original structure of hyperlinks, the Information System can be widely used to the web structure analysis and achieve high performance [7].

Web query classification and web page classification are leading major role to classify the web documents in IR using classification algorithms, based on the user query. The classified documents are indexed using ranking algorithm. Solving the complicated problems in relevancy, accuracy and ranking the web documents and can design a better ranking algorithm could lead this research work in future [8].

The various preprocessing algorithms and its heuristics are applied and examined by implemented using programming languages. Data preprocessing algorithms are used to parse the raw log files that involve splitting of the log files and then cleansed to obtain superior quality of data. Based on this data, the unique users are identified which in turn helps to identify user sessions. At present sentiment analysis has become a very promising field that uses web log files for sentiment analysis and opinion mining. So preprocessing of data for any conclusion making progression that is based on web log is required. So handling of data mining methods and knowledge discovery on the web is now on the spotlight of boosting a number of research communities [9].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

Web scraping is related to web indexing, whose task is to index information on the web with the help of a bot or web crawler. Here the legal aspect, both positive and negative sides are taken into view. Due to a vast community and library support for Python and the beauty of coding style of python language, it is most suitable for Scraping data from Websites. As Scraper opens up another world of retrieving information without the use of API, and mostly it is anonymously accessed [10].

Search engine optimization is a means to obtain a better ranking of the search results via a variety of approaches to manage the website to make it more compliance with the principles of search engine ranking. It can effectively retrieve combination of key words to benefit the website, where websites for small and medium enterprises or emerging website operator can take the best effective means through search engine optimization to increase the website's level of exposure in order to increase the chance of successful operation of website [11].

To extract maximum value from search queries, search engines must develop efficient approaches for generating more precise and multi-functional clusters of similar queries. However, most prevalent clustering methods all suffer from certain limitations on clustering this highly diverse set of queries. Instead of implicitly considering the content and click-through data, we can personalize the method to provide recommendations by considering the relations between users [12].

The engineering education community increasingly faces the need to assess outcomes of instructional practices or to assess impacts of student characteristics, institutional culture, or educational innovations on student learning and retention. Importantly, institutional and program accreditation require assessment of student outcomes, and many grant proposals require assessment of student learning outcomes. However, the typical engineering faculty member lacks training in assessment of educational outcomes and is unaware of instruments available for assessing these outcomes [13].

The correlation that exists between frequencies of use of one hundred most commonly used words in three corpora of Croatian language and number of web pages which contain those words returned by Google Search engine. Google Search engine contains an extremely large number of indexed documents, which represent a sample of the Croatian language [14].

Fast growing of massive digital content in the Internet makes it difficult for learners to find educational contents that suit their learning competences and preferences. In this research, we aim to augment the existing Google search engine with a dynamic learner profiling feature and a recommendation mechanism to enable the delivery of personalized search results to learners based on academic performance and searching behavior. The search results are further enhanced using relevant information from other learners with similar profiles [15].

III. PROPOSED SYSTEM

The To leverage the large volume information buried in web mining, previous work has proposed a number of techniques and tools, including web mining understanding and integration, hidden web Crawlers and web scraping. For all these approaches, the ability to smart crawl web and web scraping are a key challenge. Olston and Najork systematically present that crawling web has three steps: locating deep web content sources, selecting relevant sources and extracting underlying Content Following their statement, we discuss the two steps closely related to our work as below. Web Crawler was originally a separate search engine with its own database, and displayed correct results. More recently it has been repositioned as a metasearch engine, providing a composite of separately identified sponsored and non-sponsored search results from most of the popular search engines.

Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online change monitoring and comparison, website review scraping, gathering real web pages, website data monitoring.

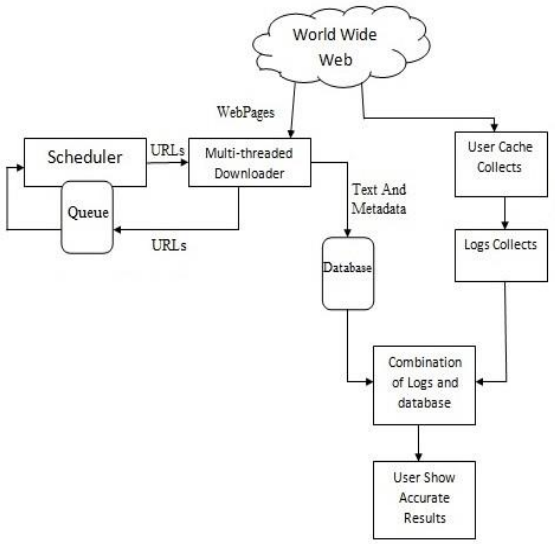

Fig 1: System Architecture

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

Generic crawlers are mainly developed for characterizing deep web and directory construction of deep web resources that do not limit search on a specific topic, but attempt to fetch all searchable forms. The Database Crawler in the Meta Querier is designed for automatically discovering query interfaces. Database Crawler first finds root pages by an IP-based sampling, and then performs shallow crawling to crawl pages within a web server starting from a given root page. The IP based sampling ignores the fact that one IP address may have several virtual hosts, thus missing many websites. To overcome the drawback of IP based Sampling in the Database Crawler; Denis et al. propose a stratified random sampling of hosts to Characterize national deep web, using the Host graph provided by the Russian search engine Index. I-Crawler combines pre-query and post-query approaches for classification of searchable forms.

One hundred most commonly used websites, their word forms and frequencies are manually entered into a database. Website data collection we visit every website and fetch the links. We create database in 40 thousand links collects. We working of database links for preprocessing or removing that links like website image link or not present keywords etc. The data analysis in user which website refers and many time to spent website we collects that and create average time for rank change.

Education search engine contain various websites present in world proximate 8 cr and above. The search engine survey all website not visits user and that not possible. In Google survey 99.6% people only click on first page and not going next page. The remaining website is loss for user. In this project we change rank with the basis of User Visits Count and spent time count that is how many people Visit website and spent time in website.Example, The College Principle post with the basis of Experience, Performance, Education, Knowledge and Research. Principle is proper guide to student. And our education search engine proper guide ranking with the basis of content of website, structure, definition of topics and good reference of papers.

IV. MODULES

Educational website visiting and extracting internal links that is fetch stored links, keywords and description. We selected 100 educational websites that isapproximately 50000 links presented. For example,

Educational Websites:-

www.admissioninengineering.co.in
www.analyticsvidhya.com
www.arxiv.org
beginnersbook.com
books.google.co.in
btechsmartclass.com
codelabs.developers.google.com
codingstreet.com
cs-fundamentals.com
cumminscollege.org
freecomputerbook.com
ieeexplore.ieee.org
indiabix.com
intellipaat.com
lecturenotes.com
link.springer.com
me.stanford.edu
mitcoe.ac.in
ocw.mit.edu
OpenOffice.org
raisoni.net
tutorialspoint.com
unipune.ac.in

Fig 2: Selected Websites

In this fig. 100 educational websites present list present.

	Address	keyword	Description	Title
2	http://www.indiabix.com/	Aptitude Questions and Answers - IndiaBIX	Learn and practice Aptitude questions and ansv	Welcome to IndiaBIX.com !
3	http://www.indiabix.com/online-test/categories/	Online Tests - Online aptitude tests for interview, com	Online aptitude tests for competitive examina	Why Online Test?
4	http://www.indiabix.com/c-programming/questions-and-answers/	C Programming Questions and Answers	C Programming questions and answers with ex	C Programming Interview Questions anc
5	http://www.indiabix.com/online-test/data-interpretation-test/	Online Data Interpretation Test - Online tests for intervi	Practice Online Data Interpretation Test and fir	Why Online Data Interpretation Test?
6	http://www.indiabix.com/java-programming/questions-and-answers/	Java Programming Questions and Answers	Java Programming questions and answers with	Java Programming Interview Questions
7	http://www.indiabix.com/online-test/digital-electronics-test/	Online Digital Electronics Test - Online tests for intervie	Practice Online Digital Electronics Test and finc	Why Online Digital Electronics Test?
8	http://www.indiabix.com/c-programming/c-preprocessor/	C Preprocessor - C Programming Questions and Answer	This is the c programming questions and answe	Why C Programming C Preprocessor?
9	http://www.indiabix.com/c-programming/strings/	Strings - C Programming Questions and Answers	This is the c programming questions and answe	Why C Programming Strings?
10	http://www.indiabix.com/java-programming/operators-and-assignmer	Operators and Assignments - Java Programming Questi	This is the java programming questions and ans	Why Java Programming Operators and A
11	http://www.indiabix.com/_files/css/css-res-min-1.4.css			
12	http://www.indiabix.com/c-programming/c-preprocessor/036001	C Preprocessor Yes / No Questions - C Programming Qu	This is the c programming questions and answe	
13	http://www.indiabix.com/technical/software-testing/	Software Testing Interview Questions and Answers	Software Testing interview questions and ansv	Why Software Testing?
14	http://www.indiabix.com/biochemical-engineering/questions-and-ans	Biochemical Engineering Questions and Answers	Biochemical Engineering questions and answe	Biochemical Engineering Interview Que
15	http://www.indiabix.com/online-test/verbal-reasoning-test/	Online Verbal Reasoning Test - Online tests for intervie	Practice Online Verbal Reasoning Test and find	Why Online Verbal Reasoning Test?
16	http://www.indiabix.com/c-programming/c-preprocessor/discussion-2	C Preprocessor - C Programming Questions and Answer	This is the c programming questions and answe	
17	http://www.indiabix.com/c-programming/c-preprocessor/discussion-2	C Preprocessor - C Programming Questions and Answer	This is the c programming questions and answe	
18	http://www.indiabix.com/c-programming/c-preprocessor/discussion-2	C Preprocessor - C Programming Questions and Answer	This is the c programming questions and answe	
19	http://www.indiabix.com/online-test/verbal-reasoning-test/latest	Online Verbal Reasoning Test - Online tests for intervie	Practice Online Verbal Reasoning Test (from La	Why Latest Online Verbal Reasoning Tes
20	http://www.indiabix.com/technical/software-testing/cmmi/	CMMI - Software Testing Interview Questions and Ansv	Software Testing interview questions and ansv	Why "Software Testing - CMMI" Intervie
21	http://www.indiabix.com/c-programming/c-preprocessor/discussion-2	C Preprocessor - C Programming Questions and Answer	This is the c programming questions and answe	
22	http://www.indiabix.com/biochemical-engineering/water-treatment/	Water Treatment - Biochemical Engineering Questions	This is the biochemical engineering questions	Why Biochemical Engineering Water Tre
23	http://www.indiabix.com/c-programming/c-preprocessor/036002	C Preprocessor Yes / No Questions - C Programming Qu	This is the c programming questions and answe	
24	http://www.indiabix.com/online-test/mechanical-engineering-test/	Online Mechanical Engineering Test - Online tests for ir	Practice Online Mechanical Engineering Test ar	Why Online Mechanical Engineering Tes

Fig 3: Websites internal Data

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

In this fig. We store data in excel file, data like websites internal links, keywords, title and description.

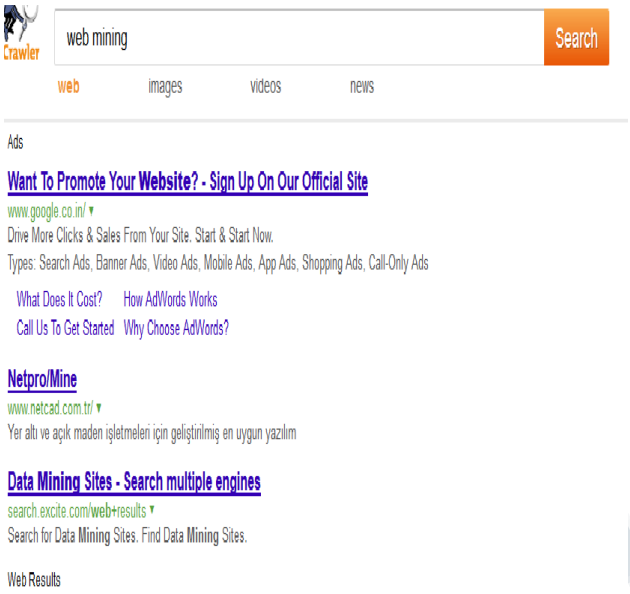

Fig 4: User Search View

In this fig. User search any educational related information in this search engine and get accurates results.

V. ALGORITHMS

A. TF-IDS

Selection This paper illustrates the method of automatic data acquisition by means of a basic TF-IDF model. In order to ensure that the data collected by the crawler is related to the subject of interest, this paper first defines a topic keyword dictionary, at the same time, the Web document is represented by n weight words, And then use TF-IDF method to calculate keywords (keywords, W)and Web documents(web page, P)

(1)Term Frequency (Term Frequency, TF):
The number of keywords in a Web document in the number of normalized representation is the word frequency, used to characterize the relevance of the subject keyword to the document.

$$TF(p,w) = \frac{fre(p,w)}{\sum_{w_i \in W} fre(p,w_i)} \qquad (1)$$

fre(p,w) Indicates the number of occurrences of the keyword w in the Web document p.

(2)Inverse Document frequency (IDF):
Inverse Document frequency is reflected by the degree of scarcity of the keyword and its relevance to the document.

$$IDF(p,w) = \frac{|P|}{\sum_{p_i \in P} dfre(p_i,w)} \qquad (2)$$

dfre(p,w) indicates the number of Web documents p that appear for the keyword w, |P| indicates the total number of documents. The total number of documents divided by the number of changes in the number of documents. Define the correlation coefficient between 1 keyword and Web document. For the keyword w with the Web document |p| their correlation coefficients are calculated as follows:

$$R(p,w) = TF(p,w) \times IDF(p,w) \qquad (3)$$

The correlation coefficient reflects the close relationship between the two, the higher the value, the higher the similarity of the subject. Based on the TF-IDF method of data extraction rules Automatic access technology cannot achieve strict field requirements Web data structured storage, the method calculates the correlation coefficient between the Web document and the interest keyword, and then completes the extracted stored procedure directly from the result, You can achieve automated storage.

B. Maximum Entropy

To apply maximum entropy method to a domain, we need to select a set of features to use for setting the constraints. For text classification with maximum entropy algorithm, we use word counts as our features. In this paper for each word-class combination we express a feature as:

$$f_{w,c'}(d,c) = \begin{cases} 0 & \text{if } c \neq c' \\ \frac{N(d,w)}{N(d)} & \text{Otherwise,} \end{cases} \qquad (4)$$

Where,
N(d,w) is the number of times word w occurs in document d,
N(d) is the number of words in d.
If a word occurs in one class, we would expect the load for that word-class pair to be higher than for the word paired with other classes.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

With this representation, if a word occurs often in one class, we would expect the load for that word-class pair to be higher than for the word paired with other classes. In text classification, we assume that features accounting for the number of times a word occurs should improve classification When constraints are estimated in this fashion, it is sure that a unique distribution exists that has maximum entropy. Moreover, it can be shown that the classification is always of the exponential form:

$$P(c|d) = \frac{1}{Z(d)} \exp(\sum_i \lambda_i f_i(d,c)), \quad (5)$$

Where each
fi(d,c) is a feature,
ʎ i is a parameter to be estimated, Z(d) is simply the normalizing factor to ensure a proper probability:

$$Z(d) = \sum_c \exp(\sum_i \lambda_i f_i(d,c)) \quad (6)$$

The invariant measure function is actually the prior density function encoding 'lack of relevant information'. It cannot be determined by the principle of maximum entropy, and must be determined by some other logical method, such as the principle of transformation groups or marginalization theory.

VI. CONCULSION

In this work, a method based on Web Structure Mining and Web Usage Mining has been proposed in order to improve the result of Web data mining systems. The importance of web mining continues to increase due to the increasing tendency of web documents. Traditional keyword based searching returns too many websites during web search that causes information overloading. This information overloading is a common problem nowadays. The results show that the method of information extraction by smart crawler and scraping are accurate and efficient and the outputs meet the requirements of archive. This search engine creates purpose of Education website ranking accurately. Due to saving a lot of manpower and time costs, it obviously improves the efficiency and automation websites ranking.

REFERENCES

[1] ZHENG Guojun2, JIA Wenchao1, SHI Jihui2, SHI Fan1, ZHU Hao2, LIU Jiang "Design and Application of Intelligent Dynamic Crawler for Web Data Mining,"2017 Ninth IEEE International Conference on e-Business Engineering.

[2] Syed Md. Galib, Ajay Shah, Md. Motiur Rahman, Maitri Debnath " Clustered and Smarter Web mining using Semantic Web," 2015 Ninth IEEE International Conference on e-Business Engineering.

[3] Simona Bernardi, Ra´ul Pirac´es Alastuey, Raquel Trillo-Lado "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study" International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013

[4] Jingtao Shang, Jianjun Lin, Van Qin, Bo Li, MengmengWu, "Design of Analysis System for Documents Based on Web Crawler" 2016 2nd IEEE International Conference on Computer and Communications

[5] Simona Bernardi, Ra´ul Pirac´es Alastuey, Raquel Trillo-Lado, "Using Process Mining and Model-driven Engineering to Enhance Security of Web Information Systems" 2017 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)

[6] Deepak Kumar Mahto, Lisha Singh, "A Dive into Web Scraper World" 2016 International Conference on Computing for Sustainable Global Development (INDIACom)

[7] Suvarn Sharma, Amit Bhagat, "Data Preprocessing Algorithm for Web Structure Mining" 2016 Fifth International Conference on Eco-Friendly Computing and Communication Systems (ICECCS-2016).

[8] Srinaganya.G., Dr.J.G.R.Sathiaseelan, "A Technical Study on Information Retrieval using Web Mining Techniques" IEEE Sponsored 2nd International Conference on Innovations in

Information, Embedded and Communication systems (ICIIECS) 2015.

[9] P. Sukumar, L. Robert, S. Yuvaraj, "Review on Modern Data Preprocessing Techniques in Web Usage Mining (WUM)" 2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions Review.

[10] Neha Verma, Prof. (Dr.) Jatinder Singh, "Improved Web Mining for E-Commerce Website Restructuring" 2015 IEEE International Conference on Computational Intelligence & Communication Technology.

[11] Tsung-Fu Lin, Yan-Ping Chi, "Application of Webpage Optimization for Clustering System on Search Engine –Google Study" 2014 International Symposium on Computer, Consumer and Control.

[12] Yuan Hong, Jaideep Vaidya and Haibing Lu, "Search Engine Query Clustering using Top-k Search Results" 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[13] Denny Davis, Sarah Brooks, Shane Brown, Howard Davis, Jennifer LeBeau, Brian French, Michael Trevisan, "Search Engine for Engineering Education Assessment Instruments" 2013 IEEE.

[14] Krešimir Pavlina, "Using Google Search Engine for Word Frequency Analysis" 2012 International Conferences on Technology Interface.

[15] Mohammad Mustaneer Rahman, Nor Aniza Abdullah, Fnu Aurangozeb, "A Framework for Designing a Personalised Web-based Search Assistant Tool for eLearning" 2017 Fifth International Conference on Information and Communication Technology (ICoICT).

[16] Andres Baravalle, Mauro Sanchez Lopez, Sin Wee Lee, "Mining the Dark Web Drugs and fake ids" 2016 IEEE 16th International Conference on Data Mining Workshops.

[17] Mohamed El Asikri, Jalal Laassiri, "Contribution To Ontologies Building Using the Semantic Web and Web Mining" 2016 IEEE 16th International Conference on Data Mining Workshops.

[18] C. Ramesh, K.V. Chalapati Rao, A. Govardhan, "Ontology Based Web Usage Mining Model" International Conference on Inventive Communication and Computational Technologies (ICICCT 2017).

[19] Jay Young, Lars Kunze, Valerio Basile, Elena Cabrio, Nick Hawes1 and Barbara Caputo, "Semantic Web-Mining and Deep Vision for Lifelong Object Discovery" 2017 IEEE International Conference on Robotics and Automation (ICRA) Singapore.

[20] Alejandro Corbellini, Daniela Godoy, Cristian Mateos, Alejandro Zunino, Ignacio Lizarralde, "Mining Social Web Service repositories for social relationships to aid service discovery" 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR).