
In this paper, the use of Naive Bayesian Classification algorithm is proposed for classifying the interested users and also we present a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users. The performance of this algorithm is measured for weblog data with session based timing, page visits, repeated user profiling, and page depth to the site length. Experimental conclusion states that the performance metric i.e., time taken and memory to classify the web log files are more efficient when compared to existing C4.5 algorithm.

Aarti Singh (2012)

This research paper emphasis on proving agent-based framework for mining semantic web contents employing clustering techniques. Clustering will help provide user with query relevant cluster of web contents, which will better satisfy user requirement and will provide optimal utilization of web surfing time. This work has proposed agent based solution for mining semantic web contents, with the aim to provide context based knowledge oriented results to the user. With the working association of web mining techniques with agent technology it has been observed that the proposed framework will lead to improved performance, reduced network traffic, and better results.

Alexender Shraer, Maxim Gurevich, Marcus Fontoura, Vanja Josifovski (2013)

With the social prospective, a very popular social blogging website “Twitter” is considered for finding of several tweets relevant to particular news all over world. The categorization is carried out on tweets (comments/reviews) in relationship with recent news and placed along with the news for people to get an idea about new comments. The proposed algorithm “top-k pub-sub” is suitable because the subscription (news) are majorly placed in the cache on the basis of frequency classification which increases the performance as the published items (tweets) arrives, they are indexed first & then matched with the subscription very quickly by access the cache but the issue of ranking quality is not mainly focused, as it includes popular ranking functions like cosine similarity & BM25.
Amit Chauhan, Himanshu Uniyal, Dr. Bhasker Pant (2013)

This paper describes a proposed framework consists of two modules: web page cleaning and main text extraction and text classification. It removes noisy data from the web pages and then retrieves the most relevant information from the same web page and finally the classification is performed by using Support Vector Machine which gives accurate and efficient result. For removing noise and relevant text extraction, the researchers have taken following steps: (1) Block Extraction (2) Block Importance Retrieval (3) Cleaned text files generation (4) Preparation Feature Vector. The resultant classification shows the average accuracy of three classes is 73.3333%, hence the method used for classification proves to be the efficient one.

Anjali Singh (2012)

For web content extraction the methodology is to use the DOM structures of web documents to efficiently implement a technique to remove irrelevant data, to optimize the WEB mining process. In this approach, firstly build the Semantic Tree to partition the web page into the content elements based on the web page tags. The main objective is to develop a technique that keep common navigation structure as it is, but removes images, advertisement and improve surfing efficiency. We are able to improve the network bandwidth requirements using this application at an intranet level .As the results from the feature selection for the segment training, the features such as InnerHTML, OuterHTML and Tag ID’s provides an effective way to implement highly efficient Web Content Extraction.

Badr Hssina et al (2013)

In this paper, a proposed method for web data extraction is deployed that uses hand-crafted rules developed in Java platform. The main objective of the proposed system is to extract patterns based on user’s interested domains. The Architecture of proposed describes the extraction by Connecting to any website and getting the data from that website, It discards comments, delete Meta tags, removes Scripts, eliminates ads and then proceed to content extraction based on the user's interests (includes list of links, images, media) and then categorize the content into the user
interest areas, which finally helps the website owner to identify the users choice and provide them with exact required contents.

D.V.N. Siva Kumar, Sabyasachi Patra (2013)

This paper discusses an algorithm for automatically categorizing the web pages which contains less text content on the basis of features extracted from both the URLs present in web page along with its own web page text content. Several experiments were conducted on the data set “WebKB” using K-NN, SVM and Naive Bayes machine learning algorithms which shows the effectiveness of the proposed approach achieving higher accuracy in predicting the category of the testing web pages. From the results of the experiment it is found that the proposed algorithm achieves higher accuracy 90% when K-NN is employed on the given data set with an observation that K-NN performs much better in accuracy than the other two algorithms SVM and Naive Bayes algorithm.

Gautam R. Raithatha (2014)

In this paper, the main focus is to extract the concepts and conceptual relationships from unstructured textual data using web content mining to create formal representation of collection of concepts and their relationships. The main objective of this research is to create a web with semantics, the information available from the unstructured or semi-structured web data has to be extracted and converted to a structured form that can be interpreted by computers. After analyzing different web mining techniques for extracting knowledge from web data for creating semantic web, it can be concluded that the unstructured data present on the web can be scanned to create ontologies to populate the knowledge base of the search engine.

Gluseppe Attardi, Sergio Di Marco, Davide Salvi (1998)

The focus of this paper is to discuss different techniques to retrieve documents on the World Wide Web by using the methods of search engines, through keyword-based queries or by catalogue, which organise documents into hierarchical collections. As it is very difficult to obtain the exact required contents from the web, automated classification is needed to extract the information (“indexed content”) over the web which matches the URL and keyword given by the user which represents the documents for which they are searching for.

Gokul Patil, Amit Patil (2011)
This paper discusses algorithm about the methodology to trace desired website or any web page according to users request and fetch data of particular domain over web by extraction on web mining. The application of VSM algorithm is used due to its capability for text classification which has more accuracy, better applicability and less manual interference. For text extraction and classification technique involves algorithms which recognize the reference keywords with the associated category and store them into a vector and then the classification takes place on the basis of depending characteristic items, to provide graphics interface or to guide operation. Due to its excellent result of the collection, it can be used to create knowledge database, and it can also be treated as a small scaled vertical search system, which makes it distinct from other search engines.

**Govind Murari Upadhyay, Kanika Dhingra(2013)**

The research paper gives briefings about the web content mining techniques, processes and its applications in the current cut-throat business environment as well in research and extracting contents for educational purposes. It also demonstrates the application of web content mining playing an important role by extracting rich set of contents and uses those contents in the decision making in the corporate environment, education and research. Thus by adopting various Data mining techniques and web content mining tools used to extract useful information or knowledge from web page contents, it makes possible to search contents over the web faster with complete exactness.

**Indra Devi Mahadevan, Selvakuberan Karuppasamy(2012)**

The need of automated classification arises due to the large amount of web contents available over internet which is growing at rapid speed and to keep track of latest update becomes difficult. Hence in this patent design, a Web page classification system for automatically generating context features for Web pages and classifying the Web pages based on the automatically generated context features is implemented. The Web page classification system includes a context feature generating module configured to generate at least one context feature for the Web page, wherein one context feature includes at least two of uniform resource locator
(URL) features, title features, and Meta features; and a classifying module classifies the web page based on at least one generated context feature.

**John Nicholas Gross (2014)**

The invention states the aggregation methodology about the news contents, search engines and other automated system where it is expected to predict stories, events or similar content related to a query or topic. The objective of this invention is to detect & differentiate document contents, particularly in different geographical regions, which proved its efficiency by considering web contents as parameters from different web sources like Google News, CNBT etc., holding information about the events of Hollywood, recent accidental activities, financial market aspects with a goal to fetch similar type of content which talks on common news topic.

**K Sudheer Reddy, Dr. G. Partha Saradhi Varma, M. Kantha Reddy (2012)**

Here the research work discovers the patterns which are most relevant and interesting by using a Web usage mining process. The objective of this research is to discover user’s behaviors, who have visited the web sites for a lesser or many number of times and in which domain. The methodology is developed to divide the log file recursively in order to discover the behaviors and to characterize them as clusters (similar behaviors are grouped into a cluster), which is based on sequential patterns. The key advantage of the proposed method is to analyses the Web Usage Mining with minimal support as a composite problem that can be solved by succeeding divisions (clusters). By adopting this approach, we could establish that the boundary between the data quantity and the quality of results can be obtained by extracting behaviors with a minimal representativeness.


The research here states that due to noisy information like advertisement, links, headers, footers etc., available over internet it becomes difficult for required data extraction from web, thus to overcome this issue a combination of two proposed techniques: automatic extraction and hand crafted rule generation are discussed to resolve such complication. This hybrid approach is proposed to extract main content from Web pages. A HTML Web page is converted to DOM tree and features are extracted and with the extracted features, rules are generated. The performance of classification is measured with metrics like precision, recall, F-measure and accuracy.
Karan Sukhija(2015)

This paper gives a brief idea about the web mining which is widely used to dig-out the necessary contents from several web resources. The research conducted here is described in 3 phases: (1) to define how web mining is implemented in research area by focusing on mining research and retrieval research. (2) It categorizes the Web mining as content mining, structure mining and usage mining. (3) It focuses on natural language processing unstructured data over the web by adopting various techniques. The major outcome from this result is the efficiency obtained from the personalization of web content by having track of earlier retrieved pages by a user that result in identification of the distinctive behavior of the user and to make forecast about the interest area of the user.


This paper describes about the well supervised learning technique called “Multilayer Perceptron” which extracts the features by HTML structure & URL traversal which the motto to optimize the feature selection (to reduce feature). Despite there are several learned techniques like decision tree, k-nearest neighbor, one r, rbf kernel etc, but multilayer perceptron technique is been proved efficient one because a vast difference is seen among the result of categorization before and after features reduction to classify page properly. Here web page categorization has been carried out using feature selection methods such as filter model, wrapper model and hybrid model in addition with web page ranking algorithm with an idea to reduce redundancy in features available in current web page to be classified.

Kollu Bhanu Prakash, M A Dorai Ranga Swamy, Arun Raja Raman(2013)

In this paper, a solution is described for a challenge where majority of the documents on the web are written in English language which makes it hard for people in India where there is huge diversity in regional languages to understand and utilize the content. A proposed model is discussed which takes inputs in both English and Telugu, an Indian regional language in both printed and handwritten formats. Words having common content are chosen and neural network is used to normalize the output. On experimental basis a sample page from a physics textbook dealing with magnetism is taken for consideration for this paper by implementation of two
different methods: statistical interpretation and ANN, which assess the content irrespective of the language of the input texts as well as extended to extract and analysis images and other media formats to extract content, interpret it and provide user with the content in telegu language.

**Lambodar Jena, Narendra Kamila (2013)**

This research was conducted to categories all the web pages & extract all the relevant dates enclosed within a web page with its relevant content as well as title. The major tasks performed during this research are categorization of web page, date extraction, title extraction, as well as content extraction. The objective of this classification method is to extract the content from the web page and provide the users with the latest content in a descending chronological order as the extraction method is developed to fetch content on the account of date and due to this it becomes possible that new updates can be tracked on last update differential basis. Keeping in mind the different format of dates (e.g. mm/dd/yyyy or dd/mm/yyyy) are utilized in different regions of world makes information retrieval more complex hence the document classification, entity extraction like date, and filling templates that correspond to given relationships between entities, are all central text mining operations are been studied in depth.

**M. Hanumanthappa, M. Narayana Swamy (2014)**

The research of this paper is conducted to analyze the following aspects: Growth of data in Indian languages, Need of text mining for Indian languages, Literature survey on Indian language text mining, Application. It is observed that text document may contain a few structured fields, and also unstructured text components, without knowing what could be in the documents, it is difficult to formulate effective queries for analysing and extracting useful information from the data. To compare the documents and rank the importance and relevance of the document the users need tools. Therefore, text mining has become popular and essential for Indian languages.

**Makoto Tsukada, Takashi Washio, Hiroshi Motoda (2001)**

By considering the web searching service provided by Yahoo JAPAN, this research demonstrates how web page classification can be performed by using the machine learning methods i.e. the system finds on its own about the new key nodes found during the search by the
user and match it with the existing one and if the match is not successful then that node is considered as a new category and then by evaluating the error rate, recall and precision the web page is allocated into appropriate category. Using machine learning methods have become almost mandatory because the no. of pages of web is growing day by day and to perform classification manually is near to impossible, as it takes longer time to be evaluated first and then reflected over the web where the information becomes outdated.

**Mansi Sood, Harmeet Kaur (2014)**

The research emphasis on providing the user’s with the article of their own choice & interest along with the news suggestion features. Here the major work describes the comparative study of available recommender system built in past & while building a classifier, a special attention is drawn towards the challenge for news importance & quick updating of dynamic news updates from all dimension. A combination of textual content analysis is conducted with machine learning techniques can be used to find similarity between information content and users’ interest areas.

**Menaka S, Radha N (2013)**

In this Paper, a proposed method employs text mining algorithms to extract keywords from journal papers, which is supported by the WordNet dictionary to calculate the semantic distances between the keywords. If the extracted keywords are having the highest similarity, then documents are classified based on extracted keywords using the machine learning algorithms - Naive Bayes, Decision Tree and k-Nearest Neighbor. The performance analysis of machine learning algorithms for text classification demonstrate that the Decision Tree algorithm gives better results based on prediction accuracy when compared to other two algorithms.

**Michael Curtiss, Krishna Bharat, Michael Schmitt (2005)**

Here the inventive development of application was conducted to identify a source with which each of the links is associated and rank the list of the links based on the quality of its content relevance. The invention concludes that the ranking of news article has improved based on the degree of news source associated with the news article where there is no limit for ranking the news article using the applied system.
Mohammed Hamed Ahmed Elhiber, Ajith Abraham (2013)
This paper discusses the process of Web Usage Mining adopting certain steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented several approaches such as statistical analysis; clustering, association rules and sequential pattern are being used to discover patterns in web usage mining. The core discussion is made on three vital steps in WUM such as preprocessing, pattern discovery and pattern analysis and for user future request prediction several algorithms of pattern discovery techniques like Fuzzy C-Means and techniques of clustering, DT and SVM techniques of classification etc. are employed and comparative analysis is been made about the exact prediction about the pattern of users navigation

Monika Henzinger, Brain Milch, Bay-wei Chang, Sergey Brin (2005)
Here the main objective is to first generate the query (comprising key words) which are essential to search news from the different sources like Live TV Streaming, Broadcast URLs, etc., and then using bunch of algorithms a depth study is carried to evaluate best algorithm to find the news and classify it into exact category. As the result it is measured that 84% - 91% of the news articles found were relevant with the query, 64% exact match was found to the topic of broadcasts and 70% of article has the close relevance to the topics specified in the query.

MyoMyo Than Naing (2013)
A web page consist of data in many formats like text, audio/video, targeted advertisement etc., which is required to be personalized in response to the query passed as parameter for further classification. In this paper, an innovative approach is discussed which starts working by findings on Intermediate category in offline mode and then the algorithm goes for online extraction of target category on the basis of keyword specified within the submitted query. At last the resultant data-set is ranked on account of its relevance with the target labels.

Ning Zhong, Yuefeng Li, and Sheng-Tang Wu (2012)
The research is on an innovative and effective pattern discovery technique which includes the processes of pattern implementation and pattern evaluation, to improve the effectiveness of using and updating discovered patterns to find relevant and interesting information. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. Certain experiments are carried on RCV1 data collection and TREC topics to demonstrate that the proposed solution achieves encouraging performance. These
techniques include association rule mining, frequent data-set mining, sequential pattern mining, Maximum Pattern mining and closed pattern mining. With a conclusion an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining.

**Olivier Mirandette, Marc Tremblay, Eric Melin(2012)**

The method is designed to categorize URLs (Uniform Resource Locator) of the web pages accessed by multiple users over an IP (Internet Protocol) based data network. Here the real time data are collected from IP data traffic occurring on the IP based data network and extract parameters from the collected real time data along with parameters including URL of the web page by evaluating the URL’s association with the matching category.

**Pooja Mehtaa et al(2012)**

The research describes Web Personalization as an area of the Web usage mining that can be defined as delivery of content tailored to a particular user or as personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in the content delivery framework to manipulate what information is disseminated to users. The method web content personalization consist of four phases: Resource Finding - the task of retrieving indented Web documents, Information Selection and Pre-processing - automatically selecting and pre-processing specific information from retrieved web resources, Generalization - automatically discovers general patterns at individual web sites as across multiple sites and Analysis - validation and interpretation of the mined patterns. It is found that the concept of web personalization has a wide impact on business application which is recently applied in the field of E-Commerce application in order to identify user’s choice and recommend them their suitable needs.

**Pravin M. Kamde, Dr. Siddu P. Algur(2011)**

The purpose of this paper is to provide a systematic overview of multimedia mining with clean intension to represent the issues in the application process component for multimedia mining followed by the multimedia mining models used in retrieval, indexing and classification of multimedia data with accuracy in information gathering carried in different dimensions. An interesting research direction on web content mining is the integration of heterogeneous
information sources. Several limitations are discussed regarding multimedia mining, text mining and image mining in which one of the issue is too much content is lost when the sequence of multimedia is ignored and this is resolved by proposed model of multimedia mining.

**Punam Bajaj, Payal Joshi, Anchal Garg (2014)**

In this paper, we will discuss how to identify obstacle noises to improve efficiency of web mining, and to eliminate noises using simple LRU algorithm. While conducting the research it was found that much useful information is surrounded by noises such as banners, privacy notices, advertisements etc., these noises effects web pages performance and efficiency. Web page noises have been categorized into two parts- Global Noises and Local Noises. The result of Least Recent Used algorithm records that it takes less time and less complex algorithm for web mining.

**Sarika Y. Pabalkar (2012)**

The research model extracts required news and classifies particular news from webpage into certain class or multiple classes and Then again subs classifying it in to specific class. The classification process consists of two phases (1) Extraction, required information is extracted by checking maximum text density from the text contents from root of a web page. (2) Text Classification Techniques, employs Vector Space Model (VSM) as a technique for classification which it confirms categories of samples that need to be divided by method of computing text. The developed system will provide the exact category of information to the final user that provides different personalized services according to different user’s requirements.

**Sonal Vaghela, M.B.Chaudhary, Devendra Chauhan (2014)**

Here from ample of availability of web mining techniques like term occurrence number, term frequency and document frequency are discussed with the prime focus on term frequency. This research work classifies web pages of four categories conference, course, department and student which results in educational universities' catalogs. Hence by generating catalog, it provides users categorized view of information and is more effective for users to find desired information.

**Sujit V. Chaudhari, Shrikant Lade (2013)**
To conduct Text Pattern Mining, there are two approaches been discussed (1) term based approach and (2) phrase based approach. But due to its less accuracy in knowledge discovery from text document, a proposed method utilizes pattern approach with the set of keywords, which is an innovative and effective pattern discovery technique by which research articles, news articles classification of different field are done and more than 80% of documents are identified successfully. This pattern based approach is executed in four phases: Text Pre-processing, Pattern taxonomy process, Pattern deploying, Pattern evolving. As this approach results in developing the effective pattern is done both with the help of relevant or positive document or irrelevant or negative document and a set of keywords that contain text of those relevant fields.

**Supreet Dhillon, Kamaljit Kaur (2014)**

In this paper, the efficiency of classification algorithms on its productivity is evaluated on the bases of some factors like accuracy, precision, session based timing, recall. To conduct web usage mining several algorithm are discussed and its listing includes Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machine, Neural Networks, Rule Based Classifier and K-Nearest Neighbor Classifier. On comparison of these entire algorithms concludes a frame for web usage mining based on classification algorithms including their features and limitations and it is found that Naive Bayesian performed well with respect to all the factors in compare with Decision Tree classifier and SVM also perform well as compared to others.

**V. Sujatha, Dr. Punithavalli (2011)**

The research paper aims to discover frequent patterns in Web log data is to obtain information about the navigational behavior of the users, which can be used for advertising purposes, for creating dynamic user profiles etc. In this paper four types of clustering approaches are investigated in web log files to improve the quality of clustering for user navigation pattern in web usage mining systems, for predicting user’s intuition in the large web sites. The research work conducted is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user’s navigation behavior.

**Xiaofei He (2010)**
This patent describes a method that comprises a technique to retrieve a huge amount of web pages, selecting a plurality of terms from each web page, assigning a laplacian score to each term, filtering each plurality of terms by each term laplacian score to form filtered set of terms, and categorizing each web page of said plurality based in the filtered set of terms. Here, method employs an analysis system that applies a ridge regression algorithm to the filtered set of terms, and predicts each web page to a relevant category.

**Xiaoguang Qi, Brain D. Davison (2009)**

This article describes a comparative background of classification features of specific Web page, having major focus on web page classification and also a detailed description about research from future prospective. Ultimately, Web page classification is a process of accumulating / allocating several web page(s) into a well-defined label known as Category. Here, different types of classification are stated down which acts as a methodology to perform classification and they are broadly defined as Subjective classification which assigns particular web page on the basis of some predefined domain like “Technology”, “Sports”, “Top Stories”, etc. and on other hand there is Functional classifications which perform classification on the basis of activity performed by that web page like “Registration page”, “Admission page”, etc.

**Yogendra Jain, Sandeep Wadekar (2011)**

This article discuss the problem of exact information retrieval through the explanation and implementation of searching algorithms and its hierarchical classification techniques for increasing a search services understanding the search query, provided within are the set of techniques and matrices for performing temporal analysis on query logs. This logs are further analyzed to recognized the pattern and detect the classifying trends by providing value added information along with the search query.