A COMPREHENSIVE STUDY OF MINING WEB DATA

Asem Bidyapati Devi[1], Prajwal.G[2], Sameeksha Aithal[3], Samhith.V[4]

[1]*Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India*
[2]*Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India*
[3]*Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India*
[4]*Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India*

Abstract

World Wide Web is a repository of massive amount of data related to various fields. It is difficult to obtain the necessary and relevant information from this vast collection. Many researchers have proposed different methods for fetching out accurate results from the web for the given user queries. In this paper we have made an attempt to consolidate the works done by different researchers in the three fields of Web Mining namely, Web Content Mining, Web Structure Mining, and Web Usage Mining. A simple study of the researches done on these three domains is provided in the paper. This study provides a basic knowledge about the recent past studies on the field of Web mining to the readers, who are interested in further development of this field.

Keywords:*Web Mining, Web Content Mining,Web Structure Mining,Web Usage Mining*

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

1. INTRODUCTION

The usage of internet recently has grown in exponential terms and hence the use of World Wide Web. There is a need for effective retrieval of appropriate information from this World Wide Web as per the user requirements. To accomplish this task researchers and scientists have come up with many techniques and methodologies which led to the introduction of the idea of Web Mining. Web Mining is the application of Data Mining techniques on World Wide Web [1]. Web Mining involves mining process in three branches i.e. Web Content Mining, Web Structure Mining and Web Usage Mining.

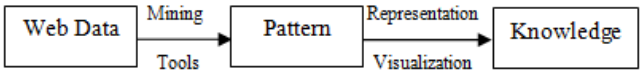

Fig -1: Overview of Web Mining Process

Web Content Mining is the process which concentrates on extracting useful knowledge derived from the content information available within the websites in terms of texts, images, audios and videos.

Web Structure Mining is a stream of Web Mining which explores the web structure information i.e. by perceiving the web pages and links basically as a graph. Mainly this involves the study of hyperlink structure of web pages.

Web Usage Mining is the process of extracting usage patterns and information of Web User's activity during a session for a particular search query. Thus it focuses on the techniques that predict user's behavior while interacting with the Web.

2. RELATED WORKS

2.1 Web Content Mining

Jia Li [1] published a paper named "Using Distinct Information Channels for a Hybrid Web Recommender System". In this paper, the author has developed a Hybrid Recommendation System which uses all the three information channels i.e. web content, structure and usage. The architecture of the proposed system basically contains two modules; an offline component, which pre-processes data to generate the user profile; and an online component, which is a real-time recommendation engine. Here, for building the offline component the approaches used are User and Session Identification, Mission Identification (an improved transaction identification approach), Clustering the Missions to build user profiles and Augmenting and Pruning the clusters to improve user profiles. For mission identification purpose, Content clustering technique is employed which uses a modified DC-tree algorithm to obtain a set of keywords, to identify the interesting cluster and represent each web page by a feature vector (Web Page Feature Extractor). For clustering the missions, Page Gather algorithm is adopted to generate a set of page clusters to build user profiles. For other two approaches, the author has applied the techniques of usage and structure mining. The Online module is constituted with a recommendation engine which responds to trigger (user's current information need) by providing √n best recommendations where n is the number of links, with a maximum of 10.

For evaluating the system, the metrics used are accuracy, coverage, precision, recall, recommendation accuracy and shortcut gain. A powerful tool called VIVIDESK is used to generate logs based on the web pages that are visited in terms of client address, request date and time. The result on the mentioned data-set is summarized as follows:

Table-1: Results from Paper [1]

Data-Set	Metric (%)	Results
40000 web pages, 150000 links, 800000 missions per month, 1500 content clusters per month.	Accuracy and Recall	80-57 & 8-22
	Shortcut Gain and Recall	14-45 & 8-20
	Recommendation Accuracy and Shortcut Gain	45-79 & 64-14
VIVIDESK dataset		
Application missions	Accuracy and Coverage	83-33 & 6-51
Text missions	Accuracy and Coverage	80-31 & 7-52
Application missions	Shortcut Gain and Coverage	2-35 & 6-51
Text missions	Shortcut Gain and Coverage	2-34 & 6-45

The integration of all the three information channels is the first approach in designing the hybrid recommendation system. Hence, this idea led to higher accuracy in recommending relevant information for the given user queries.

Moreno Carullo [2] proposed "Web Content Mining with Multi-Source Machine Learning for Intelligent Web Agents". Here, an automated method suitable for a wide range of domains based on machine learning and link analysis is introduced. For machine learning, three different approaches namely Supervised, Unsupervised and Semi-supervised are applied. Few techniques and tools called Wrappers and Multi-site web content mining algorithm are used. Wrappers is a software tool which permits to view a website as a simple database table by considering data of interest in the form of a structured table. Wrapper induction uses wrapper inductors in information extraction to build a wrapper out of some supervision. For wrappers and page representation, two main approaches, i.e., Plain text approach and Structured approach are used where all available information is exploited. The page representation technique strongly affects the extraction quality and limitations. Multi-site web content mining uses an approach called LEWECOM (Learnable Web Content Mining approach), a general model to recognize a given data-set of interest.

Based on precision, recall, f-measure and usefulness metrics, the evaluation of this intelligent system is done. The data-sets used are WEBNEWS-1 which was collected on 29 daily news websites with a total dataset dimension of 310 pages; 207 pages as training set (TrSnews) and 13,192 numbers of images. The COMMOFF-1 dataset was collected on 600 e-commerce websites with a total dataset dimension of 1,200 pages, 800 pages as training set (TrSoffers) and 61,692 numbers of images. New dataset was built which consisted of 822 pages with a total block count of 172,937 and total anchor count of 1,676 from European E-commerce websites.

Experimental results on different datasets are specified as follows: Feature analysis results of WEBNEWS-1 and COMMOFF-1 datasets are proposed and values for precision, recall and f-measure are tabulated. COMMOFF-1 dataset results with 'product name' field of interest and 'price name' field of interest values for the precision, recall, and f-measure are proposed and the average gain for precision, recall, and f-measure considering the feature usefulness for both fields of interest are also reported in the paper. The results obtained by the overall set of experiments are reported in the table below:

Table-2: Results From paper [2]

Feature set (G)	Feature set (F)	P	R	F_1
	Intitle, fsize, fbold	0.78	0.66	0.72
dice	Intitle	0.87	0.81	0.83
dice	Intitle,fsize,fbold	0.88	0.84	0.86

This paper proposes a suitable model for web content mining to extract accurate information from the web. Different machine learning models, content mining and structure mining techniques have been adapted to design the model. Mainly, the focus is on the page contents related to particular domain on which aforementioned methods are applied and appropriate results are generated.

G. Poonkuzhali, R. KishoreKumar, P. Sudhakar, G.V.Uma, K.Sarukesi [3] proposed a paper titled "Relevance Ranking and Evaluation of Search Results through Web Content Mining" in which they have tried to ease the information retrieval process from the web. Here, they have adapted different methods like pre-processing, used to generate profile of all words and store it in hash table. This includes: Stemming, which is used to compare the root forms of the searched terms to the documents in its database, Stop words removal, which is used to eliminate certain words that do not affect the final result, and Tokenization, that is used to split the words into small meaningful constituents. Term frequency computation is the second method used to compute frequency of all the words. Next method involves computing correlation coefficient between two documents. Final one is to rank the relevant documents. Correlation technique helps to analyze the behavior of two or more variables. Correlation analysis is used to find the related documents from the input document set of some particular category. An algorithm called Correlation algorithm for relevance ranking is applied to find the

correlation between the documents, to remove the redundant documents and to rank the documents.

For evaluation, few Performance Metrics are used, namely Discounted Cumulative Gain (DCG), Relevant Score, Normalized Discounted Cumulative Gain (NDCG), Positional parameter (c) and Positional parameter (I). For testing purpose, 10 documents are taken and named respectively as D1, D2…D10.The experimental results of the proposed algorithm are shown below:

Table-3: Results from Paper [3]

Documents ranked by Human	Ideal Ranking Position (I)	Correlation Ranking Position (C)	Relevance Score (RS)
D4	1	3	10
D1	2	2	10
D2	3	5	9
D7	4	6	9
D5	5	1	8
D6	6	4	8
D3	7	7	6
D8	8	8	4
D9	9	9	0

In this paper, the authors have developed a new algorithm called as correlation algorithm to compare the resulting web pages and to rank them according to their relevance for the given user query. Hence, the results obtained show its accuracy level in ranking the pages to meet user satisfaction.
Zakaria Sulman Zubi [4] has published a paper named "Using Some Web Content Mining Techniques for Arabic Text Classification". In this paper, web content mining is used to extract non-English knowledge from the web. An algorithm called Arabic language independent algorithm is used as a machine learning system to classify various numbers of documents written in a non-English text language. Different methods have been specified like: Pre-processing, which involves some steps namely, Stemming the words (to reduce the number of related words in the document), Weight assignment (assigns real numbers between 0 and 1 to each keyword for its prominence) and K-fold cross validation (used in CK-NN to test accuracy of ATC system). Classification technique is mainly utilized to classify the Arabic language based on its grammar. Classifier Naïve Bayesian algorithm (CNB) is implemented to compute the conditional probability and categorizes the incoming objects to its appropriate class based on its posterior probability. Classifier K-Nearest Neighbor algorithm (CK-NN), a supervised learning algorithm, stores all the training samples as classifiers. It classifies the object based on the attributes and training samples. Based on the above two classifiers, a system called as Arabic Text Classifier (ATC) has been proposed, which uses the comparison results obtained by the two classifiers and selects the best greater average accuracy result rates to start the retrieving process. For evaluation, the algorithms are constituted with different functions and dataset named Arabic Corpora i.e. a corpus of Arabic text documents collected from Arabic newspapers archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and Al-Dostor as well as a few other specialized websites. Entire dataset consisted of 1,562 documents belonging to 6 different categories specified as: Cultural news-258, Sports news-255, Economic news-250, Social news-258, Political news-250 and General news-255. The results of the proposed ATC system along with other systems proposed by different researchers prior to this are shown below:

Table-4: Results from Paper [4]

Researchers	El-Kourdi et al.	Siraj	Sawaf	El-Halees	Our result (ATC)
Training	8560	8560	8560	6740	1562
Test	2780	2780	2780	2605	798
Topics	93	88	90	90	6
CNB	68.78 %	72.02 %	50%	74.41 %	77.3%
CK-NN	75.02 %	82.03 %	62.7 %	85.01 %	86.2%

This paper typically deals with the idea of text classification based on the language used to specify the text documents. Here, the author mainly focuses on classifying the documents written in Arabic language. Hence, the results generated using ATC are more accurate than the others specified in the paper.
Jianfei Gou [5] has worked on the implementation of a Vertical Search engine system and has published a paper on it named "Web Content Mining and Structured Data Extraction and Integration: An Implement of Vertical Search Engine System". To develop this system, he has adapted some methods namely: Raw data acquisition (task of collecting raw data), Data extraction and Information retrieval, a technology in content mining to extract information using some techniques. Various tools are used such as Web crawler, which are programs that use web structure to navigate among different pages using an algorithm to note different URLs' while moving from a website, Data Extractor, which focuses on structured data extraction and the program which does this task is called as a Wrapper, Data Integrator, which accomplishes the task of mapping of raw data from sources into a unified database, and index searcher in which an indexing technique called Inverted Index is applied to index web pages to help design an information retrieval system or a search engine for fast retrieval of pages to users.

For experimentation, they have considered the hot E-commerce, i.e. data of online shopping and group-by industry as the experimental object. They have also presented the search interface available for the users. This vertical search

engine gives out appropriate results for the given user query and an example of a search query is also depicted in the paper. Finally, the author has mainly focused on developing a search engine to be different from other search engines. Here, a search engine named vertical search engine is developed to respond to the user queries more accurately and also saves user time from viewing redundant information. This system also focuses on the page content and then gives out the results with high precision compared to other search engines.

2.2 Web Structure Mining

Nacim Fateh Chikhi, Bernard Rothenburger, and Nathelie Gilles [6] came up with a paper called "A Comparison of Dimensionality Technique for Web Structure Mining" which uses a supervised approach. They made use of Dimensionality Reduction Technique (DRT), which is the process of reducing the number of random variables under consideration and these variables can be divided into feature selection and feature extraction in machine learning. They have made use of Principal Component Analysis (PCA), a statistical procedure which uses orthogonal transformation to convert the set of observations of correlated objects into a set of linearly uncorrelated variables. Independent Component Analysis is further used to linearly transform the original data into statistical independent components. Non-negative Matrix Factorization is employed for approximation of high dimensional data. Lastly, high dimensional data is projected onto lower dimensional subspace, thereby achieving dimensionality reduction by Random Projection method.Accuracy and Normalized Mutual Information Value are the metrics used for comparison. The comparison is carried out on 4,200 pages collection of antique WebKb, 5,360 pages from Wikipedia covering 7 topics and 3,270 pages obtained from the first 200 result from Yahoo Search engine for the word 'Amstrong'.

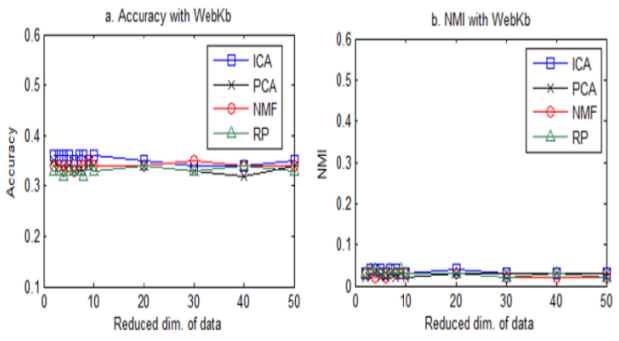

Fig-2: Results of paper [6] for WebKb dataset

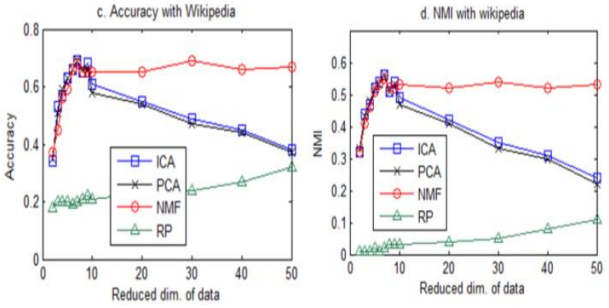

Fig-3: Results of paper [6] for Wikipedia dataset

As shown above, for Web Kb dataset, accuracy ranges from 0.32 to 0.36 and the NMI is very low. For Wikipedia dataset, variable results for the different dimensions of data are shown. From the results, we can conclude that Non-negative matrix factorization (NMF) is a promising approach for web structure analysis because of its superiority over other methods as it has higher accuracy values.

"The anatomy of a Large-Scale Hyper textual Web Search Engine" is a paper by the famous Google founders Sergey Brin and Lawrence Page [7] in 1998. This was the first idea of Google search engine as a prototype making use of both Supervised and Unsupervised approaches. They came up with the idea of Page Rank, an algorithm to rank websites based on its citation importance with people's subjective idea of importance. They made use of a repository which contained full HTML code of every web page which is document-indexed and ordered by docID. Lexicon and Hit lists are also used which are indexed in the form of forward and reverse index barrels, holding a range of word IDs. These techniques are integrated and carefully organized so as to efficiently retrieve the query results. Web Crawler, Indexers, Parsers, Sorters and Barrel Indexers are the tools used for preprocessing large data.Since this was a prototype modeling idea, the performance metrics used here are real-time measures specified as Quality of Search result(relevance of search result), Storage Requirements(to store index, lexicon, repository), System Performance(time taken for preprocessing the data like crawling and indexing), and mainly, Search Performance (time span to respond to search queries). To evaluate this prototype, they made use of 26 Million pages from WWW. The results were promising, with a high quality of search results and with a quick response time of within 2 seconds, but takes a considerable amount of time for preprocessing the web pages.Google appears to be a feasible search engine with a primary objective of providing high quality search results over a rapidly growing WWW. It is also a complete architecture for gathering web pages, indexing them and performing search queries efficiently with a quick response time.

P. Ravi Kumar and Ashutosh Kumar Singh [8] proposed a paper called "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval". Here, they dealt with the previous works done in Web Structure Mining and brought advancements to it. The supervised technique is mainly used with Page Rank [7] and HITS [19] as basis. The advancement is brought with the idea of Weighted Page Rank, an extension to Page Rank Algorithm which assigns a higher rank value to important pages rather than dividing the rank value evenly among its outgoing links. For evaluation, Page Rank values are used as performance metrics. They have explained the idea of a simple hyperlink structure of 3 pages as shown below:

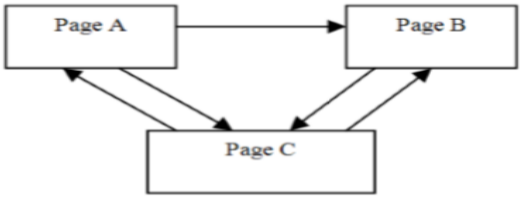

Fig-4: Simple Hyperlink structure used as Dataset in paper[8]

The results of convergence of Page Rank values through various iterations are shown below:

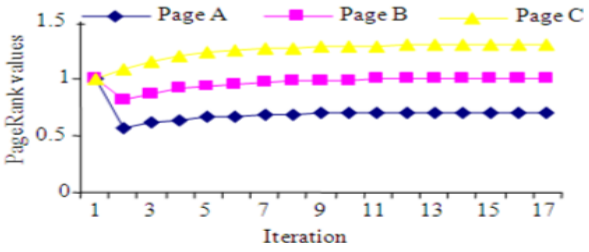

Fig-5: Page Rank Distribution from results of paper [8]

Table-5: Comparison of hyperlink algorithms

Criteria	Algorithms		
	Page Rank	Weighted Page Rank	HITS
Mining technique used	WSM	WSM	WSM and WCM
DP parameters	Back links	Back links, Forward links	Back links, Forward Links and content
Complexity	O(logN)	<O(logN)	<O(logN)
Limitations	Query independent	Query independent	Topic drift and efficiency problem
Search engine	Google	Research model	Clever

The Page Rank Computation results show that the incoming and outgoing links play an important role in the ranking of web pages. Thus, assigning weights and computation of Page Rank for web pages provides better values for Page Rank.

Debora, Leonardi, Millozzi, and Tsaparas [9] came up with a paper called "Mining the Inner Structure of the Web Graph". With Web pages as nodes and the hyperlinks as edges, idea was given that WWW existed in the form of a bow-tie structure for human perception. But these authors used External and semi-external Memory Graph Theoretic Algorithms for the detailed study of the structure. They made use of Power Law which follows the Degree distribution and made series of measurements on CORE, IN and OUT components of the bow-tie graph. They worked on a dataset containing samples from Italian, Indochina and UK domain collected by the "Language Observatory Project" and samples from whole Web collected by the WebBASE project at Stanford in 2001 containing 360 millions of nodes and 1.5 billion edges obtained from different crawlers. The metrics used for quantization of the components are In-degree distribution, out-degree distribution and strongly connected component distribution.

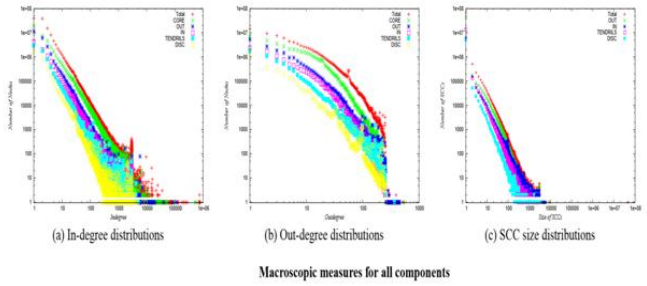

Fig-6: Macroscopic measure of result from paper [9]

The detailed analysis of the obtained results led to the inner form of the Bow-tie structure called as Daisy Chain Structure.

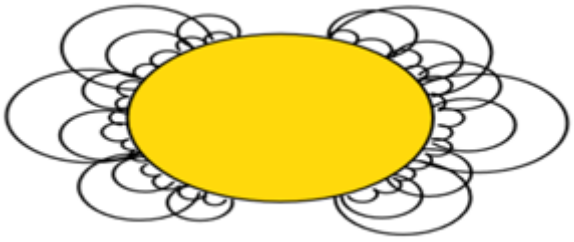

Fig-7: Conclusive detail structure of Web from paper [9]

The detailed structure study of the web reveals important and useful information which can help in effective web mining.

Hung-Yu, Shian-Hua, Jan-Ming, and Ming-Syan [10] proposed a paper called "Mining Web Informative Structures and Contents Based on Entropy Analysis". In this paper, the problem of mining informative structure of news Web site consisting of thousands of hyperlinked documents is studied and a module which provides solution to the problem is designed. Based on HITS algorithm, entropy based analysis mechanism (Link Analysis of Mining Informative Structure) for analyzing the entropy of anchor texts and links by eliminating the redundant hyperlinked structure is introduced to filter for particular information. This LAMIS makes use of page mode versus content block mode idea along with Hybrid ranking of authority and hub. To eliminate unwanted information and links (like advertisements), a new algorithm called Info Discoverer is used. This Info Discoverer analyzes the information measure of content blocks (page set) and an entropy-threshold value is set to classify the page. It also applies link entropy to discover the informative structure generating link entropy values to enhance the effective retrieval of information. The module developed uses Web crawlers, Feature Extractor and Informative structure mining modules as tools. The performance metrics used for the evaluation of this module are Entropy values, Precision and Recall. The datasets used for evaluation are 33,747 websites crawled from 13 Chinese and 5 English News websites as root URL.

Table-6: Results of Link Analysis of HITS and Entropy-Based HITS from Paper [10]

Method	HITS		Entropy-based HITS	
	Authority	Hub	Authority	Hub
P_0	0.535	0.297	0.229	0.142
P_1	0.419	0.524	0.338	**0.756**
P_2	**0.576**	0.160	0.244	0.031
P_3	0.321	**0.553**	**0.622**	0.451
P_4	0.321	**0.553**	**0.622**	0.451

On average, LAMIS increases precision by a ranging factor of 122 to 257 percent, with recall values dropping. In Info Discoverer method precision and recall values are greater than 0.95. This method helps in mining the complex Web-site structure with automatic flow serving as a very good preprocessor of Web miners and search engine application.

2.3 Web Usage Mining

V. Chitraa and Dr. Antony Selvadoss Thanamani [11] proposed “An Enhanced Clustering Technique for Web Usage Mining” in which a methodology to increase log file visibility and representing data in hierarchical clustering using enhanced k-means clustering algorithm is described.The log file entries, robot’s requests and entries with errors are removed by data cleaning. It is followed by user identification where user's IP address, browser and operating system are recorded for consecutive entries which followed by session identification and transaction identification. The data obtained is subjected to enhanced k-means clustering algorithm where the initial cluster points are calculated which are then used as centroids for intra cluster comparisons.

Table-7: Result of paper [11]

Datasets	Metrics	Results
Initial log file consisting of 9464 raw log entries with noisy entries	City block measures	Records after cleaning phase: 1476 Unique users : 124 Sessions: 365

This algorithm is stable and has a shorter running time.K.Poongothai, M.Parimala and Dr. S.Sathiyabama [12] proposed “Efficient Web Usage Mining with Clustering”. Here, the authors aimed at building a robust web usage knowledge discovery system, which extracts the web user profiles at the web server, application server and core application level with fuzzy C means clustering algorithm. It is compared with Expected Maximization cluster system to analyze the Web site visitor trends. By using cookies and forms, visitor behaviour and profiles in an e-commerce site can be analysed. Experimentation conducted with CFuzzy means and Expected Maximization clusters in Syskill Webert data set from UCI, shows that EM shows 5% to 8% better performance than CFuzzy means in terms of cluster number.

Table-8: Results for the system proposed in paper [12]

Datasets	Metrics	No of instances	Results
Monash University's Web site, Syskill Webert data set	EM precision	1012	0.745
		842	0.682
		286	0.535
	CFUZZY precision	1012	0.713
		842	0.643
		286	0.526

EM approach concludes that the higher the number of clustering lower is its values and vice versa. Also, the percentage of user usage visit coherence precision in the EM is approximately 11% higher than C Fuzzy means clustering algorithm for difference values of attributes.

Sheng-Tang Wu and Yuefeng Li [13] proposed “Pattern-Based Web Mining Using Data Mining Techniques” in which several data mining techniques like association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining are compared with

discovered patterns like SCPM,NSCPM and pattern taxonomy method to retrieve useful information for the users. PTM was proposed to replace keyword-based methods by closed sequential patterns which decrease the dimensionality but increases in efficiency of the system. The Sequential Pattern Mining algorithm (SPM) applies the pruning scheme to avoid non-closed patterns in sequential patterns discovery resulting in a set of closed sequential patterns with relative supports greater than or equal to a specified minimum support. Non-sequential patterns mining (NSPM) from a set of textual documents is used where non-sequential patterns are taken as frequent item sets.

Table-9: Conclusive result of paper [13]

Method	t20	b/e	F1	IAP	MAP
SPM	0.401	0.343	0.385	0.384	0.361
SCPM	0.406	0.353	0.390	0.392	0.364
NSPM	0.412	0.352	0.386	0.384	0.361
NSCPM	0.428	0.346	0.385	0.387	0.361
PTM	0.490	0.431	0.440	0.465	0.441

The SCPM and NSCPM data mining method, which adopts closed patterns, are more consistent around the low recall situation.

Zhenglu Yang,Yitong Wang,Masaru Kitsuregawa [14] proposed "An Effective System for Mining Web Log" that deals with log preprocessing, sequential pattern mining, visualization using LAPIN_WEB algorithm and implementation of a visualization tool to display mining results and predict users' future behavior. Lexicographic tree employing Depth First Search (DFS) strategy is used as the search path of the algorithm. Item set Extension case (IE) (in which a user clicks two pages at the same time in common sequential pattern mining) does not exist in Web log mining. Hence, we have Sequence Extension case (SE). Here, by scanning the database once, all the 1-length frequent patterns are sorted and *SE* item-last position list is constructed in ascending order based on each 1-length frequent pattern's last position.

Table-10: Results for the paper [14]

Dataset	#Users	#Items	Min. Len	Max.len.	Avg.len.	Total size
DM Research	12193	8846	1	10745	28	56.9 M
MSNBC	989818	17	1	14795	5.7	12.3 M

LAPIN_WEB is very effective and outperforms existing algorithms up to an order of magnitude. The visualization tool could be further used to make final patterns easy to interpret and thus improve the presentation and organization of web-site.

Romero, Sebastián Ventura, Amelia Zafra, Paul de Bra [15] proposed "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems". It consists of a specific Web mining tool and a recommender engine integrated into the AHA! System that helps users find relevant information. It consists of three phases: data preparation, pattern discovery and recommendation. The first two phases are performed off-line. Here, data preparation will transform Web log files and profiles into refined data and pattern discovery uses technique, such as clustering, sequential pattern and association rule mining. Finally, recommendation is performed online which uses the discovered patterns to provide personalized links or contents. Sequence mining algorithms like Apriori, GSP and PrefixSpan are used and compared with the shortcut recommendation rules discovered (number of rules discovered and the average value of the support and confidence of the rules) varying the minimum support threshold (from 0.3 to 0.03). Clustering algorithm is used to find out the number of clusters using k-means algorithm by varying the k value (number of clusters) from 2to 5 in which the data is uniform in all the clusters. Again, sequential mining and combination of clustering and sequential mining are compared using PrefixSpan algorithm varying the minimum support threshold (from 0.3 to 0.1) using all data. On the other hand, the same algorithm using the data obtained from the k means algorithm for 2 and 3 clusters GSP and PrefixSpan discover a lower number of rules with higher support and confidence values higher values. Balanced data is obtained when we use lesser number of clusters (2 and 3 clusters). Combination of clustering and sequential mining has higher values of confidence and support.

Rahul Mishra and Abha Choubey [16] proposed "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining" where they used the FP-growth algorithm for obtaining frequent access patterns from the web log data. They have compared the Apriori and FP growth algorithm. The FP-growth algorithm uses FP-tree data structure and a divide-and-conquer approach to summarize the representation of the database transaction into a set of smaller problems. The Apriori algorithm searches for large itemsets during its initial database pass and applies the same for discovering other large datasets during subsequent passes. The results are shown in table 11.

FP-growth method is more efficient and scalable for mining both long and short frequent patterns than Apriori algorithm.

R. Cooley, Pang-Ning Tan and Jaideep Srivastava [17] proposed "WebSIFT- The website information filter

system"where the web logs are filtered in extended NSCA format. WebSIFT system provides an option of converting server sessions into episodes. The server session includes sequential pattern analysis, association rule discovery, clustering or general statistics algorithms using WEBMINER prototype. Three server logs – access, referrer and agent, the HTML files that make up the site and the optional data such as registration files and remote agent logs are preprocessed to construct a user session file which is converted to the transaction file which with the data mining techniques generates rules and patterns in pattern discovery phase. With the pattern analysis, tools such as the information filtering, OLAP and knowledge query mechanism like SQL generate the final mining results. Beliefs with Mined Evidence algorithm (BME) and Beliefs with Contradicting Evidence algorithm (BCE) are used for identifying frequent item sets and interesting item sets respectively. The WebSIFT system has been implemented using a relation database, procedural SQL, and Java programming language. Java Database Connectivity (JDBC) drivers are used as interface with the database.The results are shown in table 12.

Information filtering is needed as there is no guarantee that both the algorithms will be satisfied simultaneously.

Table-11: Result of FP growth algorithm used in paper [16]

Datasets	Metrics	Results
1000 entries from the log files of NASA Kennedy Space Centre	Support	
	50k records	FP:10& Apriori:850
	80k records	FP:400&Apriori:1800
	30k records	FP:10&Apriori:520
	Number of hits	
	Most visited page	10
	Top downloaded gif file	55
	Frequently downloaded gif file	12
	Top downloaded page	49

Table-12: Results obtained in paper [17]

Datasets	Metrics	Results
The log physical size:19.3MB of 102838 entries from Web server log of the university of Minnesota	Support of 0.1% threshold	frequent itemsets:700 interesting itemsets:21 out-of-date information:2; poor page design:1

P. Nithya, P.Sumathi [18] proposed a paper called "An effective web usage analysis using fuzzy clustering" whichpre-processes the data by removing local and global noise which includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally web robots using Fuzzy algorithm in which it finds a cluster centre (centroid) to reduce the number of clusters. Here, they considered every data point with respect to each centroid depending upon the distance between them.

Table-13: Final results for paper [18]

Datasets	Metrics	Results
Anonymous Microsoft Web Dataset 37711 records.	λ-threshold 0.7	After removing records with local and global noise, graphics and videos format such gif, JPEG, etc.,: 29862 records
		After checking the status code and method field, the total of 26854 records is resulted
		After applying robot cleaning process: 18452 records are resulted

Taking $\lambda = 0.6$ for the optimal threshold, the result of the algorithm was the most ideal and the algorithm is faster and need less storage space.

3. CONCLUSIONS

In this paper, we have made an attempt to bring out a few important and ground breaking ideas which radically changed the web mining field, after going through many papers and have presented them in a simplified manner. This paper gives a generalized knowledge regarding the current state of the researches done in this field. From this knowledge, we can say that there is a need for the emergence of effective systems equipped with the methodologies of the three streams of Web Mining. Thus, further studies or researches should be on developing such type of systems for managing efficiently the problem of information retrieval from the World Wide Web.

ACKNOWLEDGEMENTS

We would like to thank Dr. Syed Shakeeb Ur Rehman, Principal, Sri Jayachamarajendra College of Engineering and Dr. C. N. Ravi Kumar, HOD, Department of Computer Science and Engineering for their support. We are extremely

grateful and thankful to our beloved professor Dr. Anil Kumar K. M for his guidance, and providing us this opportunity. We are also thankful to our parents and friends for their motivation and encouragement.

REFERENCES

[1]. "Using Distinct Information Channels for a Hybrid Web Recommender System", Jia Li University of Alberta PhD Thesis 2004.

[2]. "Web Content Mining with Multi-Source Machine Learning for Intelligent Web Agents", Moreno Carullo, Universit`a degli Studi dell'Insubria Facolt`a di Scienze MM.FF.NN.Dottorato di Ricerca in Informatica, Dec-2010.

[3]."Relevance Ranking and Evaluation of Search Results through Web Content Mining",G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi, March-2012.

[4]."Using Some Web Content Mining Techniques for Arabic Text Classification", Zakaria Sulman Zubi , Computer Science Department, Faculty of Science, Al-Tahadi University, P.O Box 727, Sirt, Libya, zszubi@yahoo.com(year not known).

[5]. "Web Content Mining and Structured Data Extraction and Integration: An Implement of Vertical Search Engine System", Jianfei Gou, University of Illinois at Urbana-Champaign, jgou2@illinois.edu (year not known).

[6]. "A Comparision of Dimensionality Technique for Web Structure Mining", Nacim Fateh Chikhi, Bernard Rothenburger, and Nathelie Gilles, IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)Silicon Valley, California, USA, November 02-November 05, ISBN: 0-7695-3026-5.

[7]."The Anatomy of a Large-Scale Hypertextual Web Search Engine", Sergey Brin and Lawrence Page, Computer Science Department, Stanford University, Stanford, CA 94305, USA, Computer Networks, vol. 30 (1998), pp. 107-117.

[8]. "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", P. Ravi Kumar and Ashutosh Singh, Department of Electrical and Computer Engineering, Curtin University of Technology, Sarawak Campus, Miri, Malaysia, 2010, American Journal of Applied Sciences 7 (6): 840-845, 2010 ISSN 1546-9239.

[9]. Debora Donato, Stefano Millozzi ,Stefano Leonardi Universita di Roma,"La Sapienza"and Panayiotis Tsaparas University of Helsinki, Eighth International Workshop on the Web and Databases (WebDB 2005), June 16-17, 2005, Baltimore, Maryland.

[10].Hung-Yu,Ming-Syan, Electrical Engineering Department National Taiwan University Taipei, Taiwan, ROC Shian-Hua, and Jan-Ming Institute of Information Science Academia Sinica Taipei, Taiwan, ROC, CIKM'02, November 4-9, 2002, McLean, Virginia, USA, 2002 ACM 1-58113-492-4/02/0011.

[11]. "An Enhanced Clustering Technique for Web Usage Mining", V. Chitraa and Dr. Antony Selvadoss Thanamani International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 1 Issue 4, June - 2012.

[12]. "Efficient Web Usage Mining with Clustering", K. Poongothai M. Parimala and Dr. S. Sathiyabama, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011 ISSN (Online): 1694-0814 www.IJCSI.org.

[13]. "Pattern-Based Web Mining Using Data Mining Techniques", Sheng-Tang Wu and Yuefeng Li, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 3, No. 2, April 2013.

[14]. "An Effective System for Mining Web Log", Zhenglu Yang Yitong Wang Masaru Kitsuregawa, Institute of Industrial Science, The University of Tokyo 4-6-1 Komaba, Meguro-Ku, Tokyo 153-8305, Japan fyangzl, ytwang, kitsureg@tkl.iis.u-tokyo.ac.jp (year not known).

[15]. "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems", Cristóbal Romero , Sebastián Ventura , Amelia Zafra , PauldeBra, Department of Computer Sciences and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain, Department of Computer Sciences, Eindhoven University of Technology, PO Box 513, Eindhoven, The Netherlands(2009).

[16]. "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", Mr. Rahul Mishra, Ms. Abha Choubey, Volume 2, Issue 9, September 2012 ISSN: 2277 128X.

[17]. "WebSIFT- The website information filter system", R. Cooley,Pang-Ning Tan,Jaideep Srivastava, {cooley,ptan,srivasta}@cs.umm.edu Department of Computer Science University of Minnesota June 13,1999.

[18]. "An effective web usage analysis using fuzzy clustering", P. Nithya, P.Sumathi, nithi.selva@gmail.com ARPN Journal of Science and Technology ISSN 2225-7217 VOL. 3, NO. 7, July 2013.

[19]. "Authoritative Sources in Hyperlinked Environment", Jon M. Kleinberg, Dept. of Computer Science, Cornell University, Ithaca NY 14853, ACM-SIAM Symposium on Discrete Algorithms, 1998, and as IBM Research Report RJ 10076, May 1997.