Cognitive Analysis over Web Server Log Emphasis on Dimension Reduction: A Review

[1]Naresh Kumar Kar, [2]Megha Mishra, [3]Subhash Chandra Shrivastava
[1]Research Scholar, [2]Associated Professor, [3] Associated Professor
[1]Department of Computer Science & Engineering,
[1]Rungta College of Engineering & Technology, Raipur, Chhattisgarh, India

Abstract : Aim behind web mining is to ascertain and retrieve valuable and fascinating patterns from a hefty dataset. There is enormous attentiveness in the field of web mining, web log data encompasses different varieties of information, including web documents data, web structure data, web server log, and web user data. WM(Web Usage Mining) has numerous application areas for example link likelihood, site restructuring, web personalization and web pre-fetching. Most significant phases of WM are the rebuilding of user sessions with the help of heuristics algorithms and ascertaining valuable outlines from these web sessions with the assistance of pattern identification techniques alike apriori or analogous ones. In this paper we will discuss different phases of web mining, different methods used for the same and their bottleneck emphasis on dimension reduction for cognitive analysis, an experimental evaluation given for dimension reduction.

***Index Terms* – WM, ML, DM.**

I. Introduction

The World Wide Web can be considered as a gigantic library. By and by, a considerable measure of reports in this enormous library is not organized by a specific request. Another reason is that the web is a profoundly powerful data source. Notwithstanding astounding development; World Wide Web's data is additionally refreshed much of the time. News, stocks and markets, organization ads and Web benefit focuses refresh their pages frequently. The World Wide Web serves to a wide assorted variety of client networks. The Internet as of now interfaces around 50 million workstations and the client network of these frameworks is expanding quickly. Web clients may have distinctive foundations, interests and use purposes. It is additionally expressed that just a little bit of data in the web is pertinent or helpful. Any web client can be keen on just little bit of the web.

The preprocessing step of WM can fluctuate with respect to source data and computation time for preprocessing mechanism. There is two categories of preprocessing as “Reactive” and “Proactive”, in reactive processing requests handled by the web server and in "proactive" preprocessing take place for the duration of the interactive browsing of the web site by the web user. In “proactive” approaches, the raw data is placid when web server is processing client entreaties. Proactive approaches are much apposite for the server pages which are dynamically created. Moreover, in proactive methodologies, suggestion of an agent with a session is resolute for the duration of the communication of user with web site. On the other hand, in reactive approaches, the accessible data is primarily server logs comprising information about client requests. Reactive approaches are typically functional on static web pages, for the reason that the content of dynamic web pages alters as per the time, it is grim to forecast the association between web pages and attain evocative navigation path patterns. There are diverse phases of attainment of pattern discovery using WM.

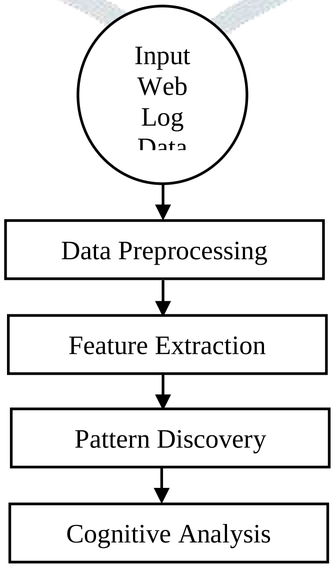


Fig. 1. Different Phases for Cognitive Analysis

There are three general classes of knowledge which can be exposed by WM:

- Web action, from server logs and Web browser action following.
- The associations among the pages, Web graph.
- Web content, for the information found on Web pages and within reports.

Data Preprocessing: Some of the data are inadequate, unreliable and contain noise. The data pre-treatment is to lug on a confederacy transformation to those dataset and the dataset will become assimilate and dependable, therefore results the dataset which meant for mining. In the data pre-treatment, primarily comprise data noise removal, user proof of identity, session credentials and path accomplishment. Basically, Data Preprocessing extracts text format data form log file and store clean data into database. It eliminates the inappropriate and duplicate log entries and it also modifies or eliminate sun ethical records from the dataset files. There are three classes of inapt or redundant data meant to be removed.

Feature Extraction: In this phase we need to select features which are relevant to our study i.e. cognitive analysis. Feature selection is an essential role in improving the eminence of learning algorithms in data mining and machine. This has been broadly deliberated in supervised learning, whereas it is still comparatively infrequent researched in case of unsupervised learning.

Pattern Discoveries: After the difference in the information in the log record into an arranged information, the example disclosure process is under gone. Pattern Discovery finds pattern, Classify data by applying mining techniques. Pattern analysis finds knowledge from the discovered pattern of the absorbing patterns by removing the extraneous patterns. Pattern Analysis includes the authentication and understanding of the mined patterns. Validation can be cast-off to eradicate the inappropriate patterns and to excerpt the curious patterns from the output of the pattern detection procedure. The output is in mathematic form which is not appropriate for direct human understandings.

Web usage mining applications

In E-Commerce, WM gives valuable data that can enhance client, deals and advertising support. A few utilizations of Web Usage Mining are as underneath.

- Improving the plan of web based business site as indicated by client's perusing conduct nearby with the end goal to all the more likely serve the requirements of clients.
- Personalizing sites as indicated by person's advantage and make progressively changing specific site for guest.
- Developing a security framework that can identify the interruption and to limit the client's entrance to certain online substance.
- Understanding clients' need and holding them by giving altered items, enhancing fulfillment with help of following perusing conduct in online business.
- Evaluating the adequacy of promoting by breaking down vast number of purchaser standards of conduct.

Further in this paper in section II we will discuss different literature, in section III motivation towards research, in section IV we will provide a tabular comparison among different literature, in section V will discuss earlier methods of dimension reduction, in section VI we will conclude our study.

II. Literature Survey

Yonas Gashaw et. al. said that in the present data time, the Internet is an amazing stage as the information storehouse that assumes an extraordinary job in putting away, sharing, and recover data for learning revelation. In any case, as there are incalculable, dynamic, and noteworthy development of information, web clients confront enormous issues as far as the applicable data required. Subsequently, poor data accuracy and recovery are a piece of the most sizzling ongoing exploration territories in this day and age. In spite of the voluminous of data dwelled on the web, profitable educational learning could be found with the use of cutting edge information mining procedures. Affiliation rule mining, as a system in information mining, is one approach to find visit designs from different information sources. In this paper, three of the chief affiliation rule digging calculations utilized for incessant example finding in particular, Eclat, Apriori, and FP-Growth inspected on three arrangements of value-based databases formulated from server get to log document. The correlation is set aside a few minutes and memory utilization angles. Dissimilar to most past research works, discoveries, in this paper, uncover that every one of the calculations has their very own fittingness and specificities that can best fit contingent upon the information size and bolster parameter edges [IEEE 2017].

Jayanti Mehra et. al. said that Pre-preparing, Pattern Discovery and Pattern Analysis. It has moreover existing a few methodologies for instance measurable examination; grouping, affiliation rules and successive example are individual used to decide designs in web utilization mining In this paper portrays information preprocessing procedure of web use mining, after fulfillment of information preprocessing, any sort of unimportant data can be deal with. Author have additionally proposed a calculation and its execution for web log preprocessing in web utilization mining. Each page has been allotted with an individual token. As per this token and recurrence, information mining method (Classification, Association Rules, and Clustering) can be connected [IJAER 2018].

Rajinder Singh Rao et. al. said that The knowledge found from WM can be used to enhance web design, introduce personalization service and facilitate more effective browsing. The various applications of web usage mining are: robots detection and removal, extracting user profiles, recommendation systems, Personalization of Web Content, Prefetching and Caching, Ecommerce etc. Web usage mining is an effective technique to extract knowledge from the unstructured data. With the help of web log data the required data can be sorted out and one can judge its popularity by deriving the interested and not interested ones. The objective of this paper is to provide a review of web usage mining [IRJET 2017].

B. Naveena Devi et. al. said that the rising popularity of electronic trade makes information digging an irreplaceable innovation for a few applications, particularly online business aggressiveness. The World Wide Web gives rich crude information as web get to logs. Presently a days numerous business applications using information mining methods to separate valuable business data on the web developed from web seeking to web mining. This paper presents a web utilization mining smart framework to give scientific classification on client data dependent on value-based information by applying information mining calculation, and furthermore offers an open administration which empowers guide access of site functionalities to the outsider [Elsevier 2011].

Ketul B. Patel et. al. said that the movement on World Wide Web is expanding quickly and colossal measure of information is produced because of clients' various communications with sites. Web Usage Mining is the utilization of information mining strategies to find the helpful and intriguing examples from web use information. It backings to know often got to pages, foresee client route, enhance site structure and so forth. With the end goal to apply Web Usage Mining, different advances are performed. This paper talks about the procedure of Web Usage Mining comprising steps: Data Collection, Pre-handling, Pattern Discovery and Pattern Analysis. It has likewise introduced Web Usage Mining applications and some Web Mining programming [IJCT 2012].

Michael L. Raymer et. al. said that Pattern acknowledgment for the most part necessitates that objects be portrayed as far as an arrangement of quantifiable highlights. The determination and nature of the highlights speaking to each example have a significant bearing on the accomplishment of resulting design arrangement. Highlight extraction is the way toward getting new highlights from the first highlights with the end goal to decrease the expense of highlight estimation, increment classifier proficiency, and permit higher order exactness. Numerous current element extraction methods include direct changes of the first example vectors to new vectors of lower dimensionality. While this is helpful for information perception and expanding arrangement proficiency, it doesn't really decrease the quantity of highlights that must be estimated since each new component might be a straight mix of the majority of the highlights in the first example vector [IEEE 2000].

Min Jiang et. al. said that domain adaptation learning (DAL) examines how to play out an assignment crosswise over various areas. In this paper, we present a kernelized local– worldwide way to deal with take care of area adjustment issues. The fundamental thought of the proposed technique is to consider the worldwide and neighborhood data with respect to the areas (e.g., most extreme mean inconsistency and intra class separate) and to change over the space adjustment issue into a bi-question streamlining issue through the bit strategy. An answer for the improvement issue will enable us to distinguish a dormant space in which the dispersions of the diverse areas will be near one another in the worldwide sense, and the neighborhood properties of the marked source tests will be protected. In this manner, exemplary grouping calculations can be utilized to perceive unlabeled target area information, which has a noteworthy contrast on the source tests. In light of the examination, we approve the proposed calculation utilizing four unique wellsprings of information: engineered, literary, protest, and facial picture. The trial results show that the proposed strategy gives a sensible way to enhance DAL calculations [IEEE 2017].

B. Santhosh Kumar et. al. said that Web Usage Mining is the utilization of information mining systems to find intriguing use designs from Web information, with the end goal to comprehend and better serve the necessities of Web-based applications. Use information catches the character or starting point of Web clients alongside their perusing conduct at a Web webpage. Web utilization mining itself can be arranged further contingent upon the sort of use information considered. They are web server information, application server information and application level information. Web server information relate to the client logs that are gathered at Web server. A portion of the run of the mill information gathered at a Web server incorporate IP addresses, page references, and access time of the clients and is the principle contribution to the present Research. This Research work focuses on web utilization mining and specifically centers on finding the web use examples of sites from the server log records. The correlation of memory utilization and time use is looked at utilizing Apriori calculation and Frequent Pattern Growth calculation [IJANA 2010].

III. Motivation

WM has numerous leads which makes it more tempting to enterprises including the government organizations.

- This technology has underway the e-commerce to do custom-made marketing, which sooner or later turn out to great in trade volumes. Government organizations are by means of this technology to classify intimidations and fight counter to terrorism.
- The anticipating capacity of mining applications can be helpful for society by perceiving criminal exercises.
- The organizations can build up better client relationship by demonstrating them precisely what they require.
- Companies can comprehend the necessities of the client better and they can react to client needs quicker.
- The organizations can discover, draw in and hold clients; they can likewise spare the generation costs by utilizing the procured knowledge of client needs.
- They can make more benefit by target evaluating dependent on the profiles made.

IV. Comparison

S. No.	Author/Title/Publication	Dimension Reduction	Algorithm Used	Description
1.	Yonas Gashaw et. al./Performance Evaluation of Frequent Pattern Mining Algorithms using Web Log Data for Web Usage Mining/ IEEE 2017	-	Apriori, FP Growth	Reasoned that every one of the calculations has their very own particular and idiosyncrasy of region of execution where they can best fit especially dependent on the span of the information and the help edge esteems. No calculation can produce visit item sets when the help

				edges progress toward becoming as little as 0.0001 or lower.
2.	Jayanti Mehra et. al./ An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining/IJAER	Data Preprocessing but dimension not applied	Page access frequency	Proposed a calculation and its execution for web log preprocessing in web use mining. Each page has been dispensed with an individual token.
3.	B.Naveena Devi et. al./Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce/ Elsevier 2011	-	API for user behavior information	Investigations led on web logs demonstrate the suitability of our methodology. Be that as it may, much work is as yet expected to add greater usefulness to web mining administrations, to make web use mining more valuable in the electronic business space.
4.	B.Santhosh Kumar et. al./Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms/IJANA 2010	-	Aprioro, FP Growth	This Research work focuses on web utilization mining and specifically centers around finding the web use examples of sites from the server log documents. The examination of memory use and time use is analyzed utilizing Apriori calculation and Frequent Pattern Growth calculation.
5.	Ketul B. Patel et. al./Process of Web Usage Mining to find Interesting Patterns from Web Usage Data/IJCT 2012	Data cleaning	Path Analysis	This paper talks about the procedure of Web Usage Mining comprising steps: Data Collection, Pre-preparing, Pattern Discovery and Pattern Analysis. It has likewise exhibited Web Usage Mining applications and some Web Mining programming

V. Dimension Reduction

In WM the delinquent is that not all features are significant. Some redundant features are there, certain may be inappropriate, and certain can even irrelevant clustering results. Furthermore, plummeting the number of features increases comprehensibility and ameliorates the problem that some unsupervised learning algorithms break down with high dimensional data.

Concluding that dimension reduction over input datset has following leads:

- It decreases the data storage, space and time required.
- Eradication of multi-co linearity increases the feat of the machine learning algorithm.
- It winds up less demanding to consider the information when lessened to low measurements, for example, 2D or 3D.

There are some reasons that why we need dimension reduction over data instance because:

- It is easy and expedient to accumulate data from sources.
- Data gathers in an unmatched speed.

Henceforth information preprocessing has huge influence for compelling machine learning and information mining applications. Dimensionality decrease is a viable way to deal with scaling back information.

- The mainstreamML and DM techniques may not be helpful for high dimensional data and Query exactness and efficiency disgrace speedily as the dimension upturns.
- Data compression.
- Noise elimination: +ve effect on query precision.

Earlier methods used for dimension reduction:

i) PCA-LRG: is a PCA-based method that selects features associated with the first *k* principal components. It has been shown that by Masaeli et al. that this method achieves a low reconstruction error of the data matrix compared to other PCA-based methods.

ii) FSFS: is the Feature Selection using Feature Similarity method with the maximal information compression as the feature similarity measure.

iii) LS: is the Laplacian Score (LS) method.

iv) SPEC: is the spectral feature selection method using all the eigenvectors of the graph Laplacian.

v) MCFS: is the Multi-Cluster Feature Selection method which has been shown to outperform other methods that preserve the cluster structure of the data.

vi) GreedyFS: The basic greedy algorithm presented in this paper (using recursive update formulas for ***f*** and ***g*** but without random partitioning).
vii) PartGreedyFS: The partition-based greedy algorithm.
viii) Genetic algorithm

In fig.-2 document Classification job is to classify unlabeled documents into categories and the bottleneck is there are thousands of terms, all are not related to job hence we need to apply dimensionality reduction so as to improve the performance of desired data mining application. There are several applications of Dimensionality Reduction:

- CRM
- Text mining
- Image retrieval
- Microarray data analysis
- Classification of Protein
- Face identification
- Handwritten digit identification

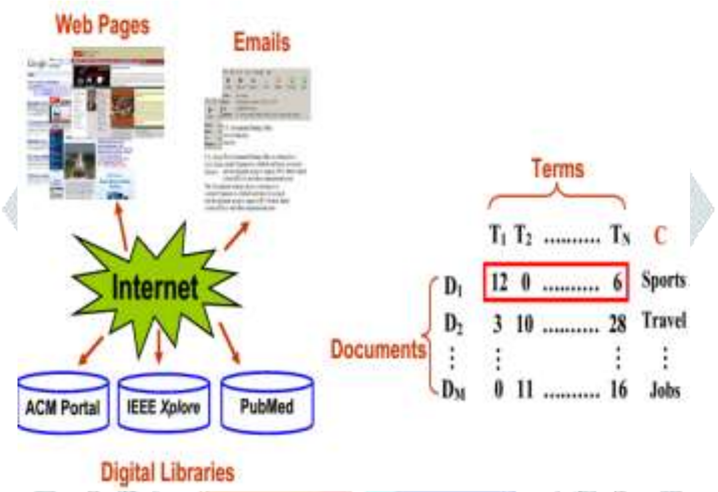


Fig. 2 Document Categorization

S. NO .	Name of Paper	Year of Publication	Author	Abstract / Conclusion
1	An Efficient Greedy Method for Unsupervised Feature Selection.	2011	Ali Ghodsi,Mohamd S. Kamel and Ahmed K. Farahat	* This paper proposes a novel strategy for unsupervised element choice, which effectively chooses includes in a ravenous way. * covetous calculation depends on an effective recursive equation for figuring the recreation blunder.
2	An automaticapproachforontology-basedfeatureextractionfrom Heterogeneoustextualresources	2012	Carlos Vicientn, DavidSa´nchez, AntonioMoreno	*pre-explained contributions to which content has been mapped to their formal semantics as indicated by one or a few learning structures (e.g. ontologies, scientific classifications). * area free, programmed and unsupervised strategy to identify pertinent highlights from heterogeneous printed assets, partner them to ideas displayed in a foundation philosophy.
3	Greedy Column Subset Selection for Large-scale Data Sets	2013	Mohamed S. Kamel, Ahmed K. Farahat, Ahmed Elgohary, Ali Ghodsi	* information experts to comprehend the bits of knowledge of the information and investigate its shrouded structure. * This paper shows a quick and exact ravenous calculation for vast scale section subset determination. *This paper likewise displays an exact and ecient MapReduce calculation for choosing a subset of

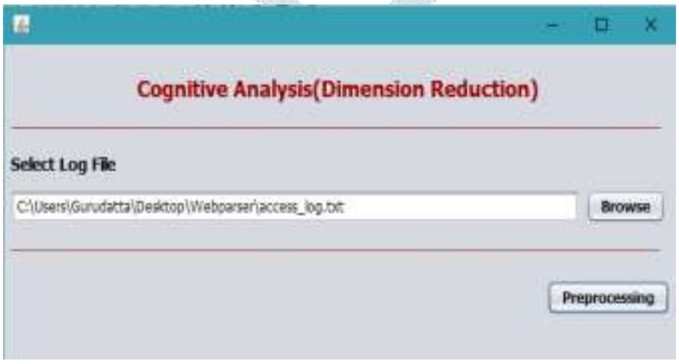
				sections from a greatly appropriated matrix.
4	A Fast Greedy Algorithm for Generalized Column Subset Selection	2013	Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel	* the choice of a couple of segments from a source lattice A that best inexact the range of an objective network B * We characterize a summed up variation of the section subset determination issue and present a quick voracious Calculation called Greedy Generalized CSS for unraveling it.

VI. EXPRERIEMNTAL EVALUATION

For implementation of Dimension reduction over log data done in JAVA 1.8, algorithm used for the same as follows:

Dimension Reduction

```
Read Log file.
//Tokenize each line by replacing "-", "\", "\\", "[]" -> ""
for each line in log file
        line.replaceAll("-", "")
        line.replaceAll("[\"\\]\\[-]","")
        update log
end loop
for each line in updated log file
        Read line in log file
        Tokenize read line
        for each token
                if token contain Null Value
                        Remove Line and Update file
                end if
        end loop
end loop
```



C:\Users\Gurudatta\Desktop\Webparser\access_log.txt

Fig.-3 Main UI of Implementation

Fig.-3 shows the first UI of implementation which takes input as log file, fig-4 chart of total data and reduced dimension data.

Precision of different dimension reduction technique					
Database	**PCA**	**ICA**	**RM**	**NR**	**Genetic**
CRAN	0.186	0.186	0.111	0.131	0.27
MED	0.253	0.253	0.174	0.197	0.31

Table-1 Precision Comparison [6]

Table-1 shows the precision value of different dimension reduction techniques.

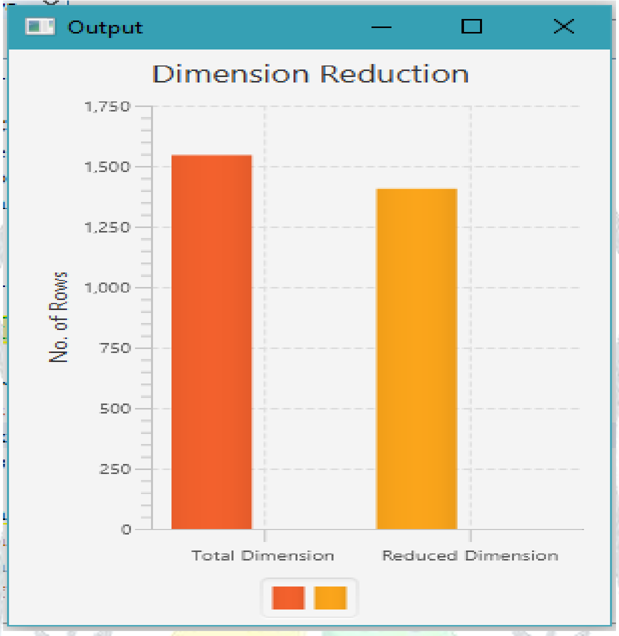


Fig.-4 Dimension Reduction

VII. Conclusion

Increasing popularity of internet draw an attention towards this study. Web server log data contains huge information which are meant for cognitive analysis such as we can find out customer buying pattern, gain customer attention, sentiment of web users. By applying web usage mining we can identify -ve and +ve association we will help for business intelligence. Earlier methods available are not time efficient and not even more accurate by applying better dimension reduction technique we can make time efficient and accurate system. In future we will use Genetic algorithm for dimension reduction which provide higher precision.

References

[1] Shiming Xiang ; Zisha Zhong ; Kun Ding Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification IEEE 2015.

[2] P.Miruthula1, S.Nithya Roopa Unsupervised Feature Selection Algorithms: A Survey IJSR 2015.

[3] Wee-Hong Ong, Leon Palafox, Takafumi Kosek investigation of Feature Extraction for Unsupervised Learning in Human Activity Detection Volume 2, Number 1, pages 30–35, January 2013.

[4] Liang Du, Yi-Dong Shen Unsupervised Feature Selection with Adaptive Structure Learning 2015.

[5] Ahmed K. Farahat Ali Ghodsi Mohamed S. Kamel An Efficient Greedy Method for Unsupervised Feature Selection 2011 11th IEEE International Conference on Data Mining.

[6] Vishwa Vinay, Ken Wood, Natasa Milic-Frayling , A Comparison of Dimensionality Reduction Techniques for Text Retrieval, Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05) 0-7695-2495-8/05 $20.00 © 2005