GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

OPPORTUNITIES AND FUTURE SCOPE IN TEXT DATA WEB MINING

HOD Yuvraj R.Gurav*[1] and Lecturer Suhas A.lakade[2]

*[1,2]Computer Engineering, Ashokrao Mane Polytachinc,wathar, Kolhapur,Maharatra ,India.

ABSTRACT

This paper is a work on overview on the current procedures of text data web mining and the issues identified with it. The World Wide Web goes about as an intelligent and prominent approach to exchange data. Because of the tremendous and assorted data on the web, the clients can't make utilization of the data adequately and effortlessly. Web mining is a utilization of data mining which has turned into an imperative region of exploration because of limitless measure of World Wide Web administrations as of late. The point of this paper is to give the past and current technique in Web Mining. This paper likewise reports the synopsis of different strategies of web mining drew closer from the accompanying points like Feature Extraction, Transformation and Representation and Data Mining Techniques in different application areas. The overview on data mining strategy is made regarding Clustering, Classification, Sequence Pattern Mining, and Association Rule Mining. The exploration work done by various clients delineating the upsides and downsides are talked about. It additionally gives the review of advancement in examination of web mining and some vital exploration issues identified with it.

Keywords**: feature extraction, clustering, classification**,** Association rule mining, Data pre-processing, Text mining**.

I. INTRODUCTION

The web is extremely tremendous, different, adaptable, and dynamic. The World Wide Web keeps on becoming both in the huge volume of traffic and the size and unpredictability of Web destinations. A large portion of the substance in the web are unstructured in nature, however almost no work manages unstructured and heterogeneous data on the Web. The rising field of web mining goes for finding and separating important data that is covered up in Web-related information, specifically in content records distributed on the Web. Data Mining includes the idea of extraction important and significant data from vast volume of information[13].

In Web sites for the most part three sorts of data are taken care of 1.Content 2.Structure 3.Log information. Taking into account these sorts of data the Web Mining comprises of 3 procedures specifically Web Content Mining, Web structure Mining and Web Usage Mining [6]. The web structure mining primarily manages the structure of the web sites[4]. Web Usage mining includes mining the utilization qualities of the clients of Web applications. It is in a semi-organized arrangement with the goal that it needs heaps of pre-preparing and parsing before the genuine extraction of the required data. This paper gives the overview of web mining procedures. Data mining process comprise of a few stages namely[7] Domain Understanding, Data choice, Data pre-preparing and cleaning, Pattern revelation, Interpretation and Reporting. The format of this paper for the up and coming segments will be as segment 2 will give a diagram about the inspiration. In segment 3 writing survey will be introduced. Segment 4 is about pre-processing strategy. In area 5 paper conclusion and future bearings will be presented[8]. We give the web mining methods study as appeared in the fig 1

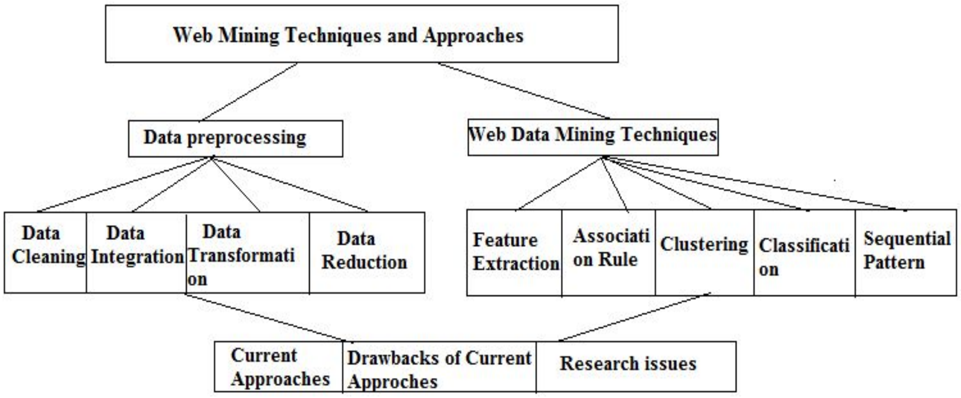

Fig1. Web Mining Techniques

The proposed plan accomplishes the below mentioned objectives:

• Issues identified with Data pre-processing, pattern revelation, web usage mining
• Identifying the issues amid mining of information from Feature Extraction, Transformation and Representation and Data Mining Techniques in different application spaces.
• Discussion of existing methods for web mining in content and text.

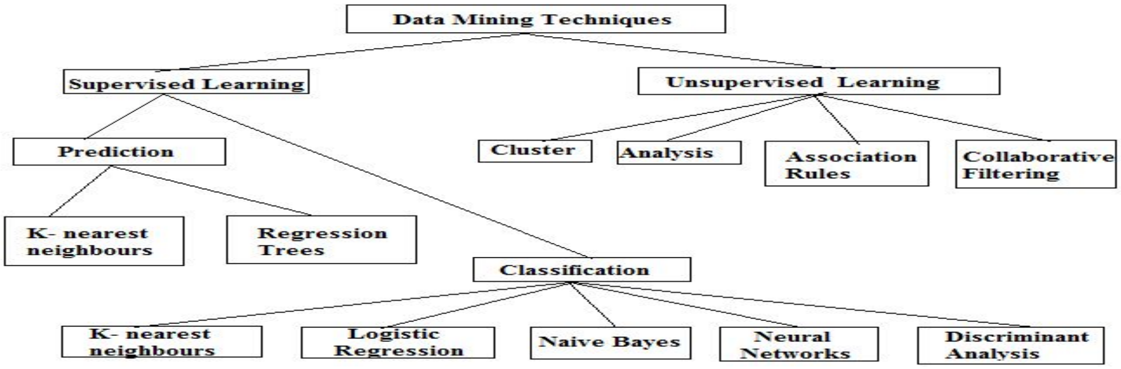

Fig.2 Data Mining Techniques

II. WEB MINING LIMITATIONS AND SOLUTIONS

Web mining is a method in data mining that consequently recovers removes and breaks down the data from web. Yang and Wu et al, (2006) talk about the different issues to be tended to in data mining. The significant issues incorporate Automated Data Cleaning, Over Fitting, Under Fitting and Oversampling of information, Scaling up for high dimensional information, Mining grouping and time arrangement information. A survey was led and given by k d chunks and a considerable lot of the scientists proposed the critical work for examination as Scaling up Data Mining algorithms for colossal information, mining content and automated data cleaning as the real issues talked about with most elevated priorities[9].

1. Data Preprocessing Techniques

Web log preprocessing is the initial step that is vital to enhance the effectiveness and nature of the web information in light of the fact that just about 65% of the time is taken in pre-preparing and these pre-handled information are given as a data to the following stages pattern revelation and pattern investigation. There are numerous methods accessible for pre-processing subsequent to quite a while. Web Server Logs keeps up a background marked by page demands. Data about the solicitation, customer IP address, demand date/time, page asked for, HTTP code, bytes served, user agent are stored. Proxy Server Logs a storing component which lies between client browser and Web servers. To enhance productivity and nature of pattern mined and to to avoid noisy and dirty data different pre-preparing systems are accessible like Data cleaning, Data coordination, Data changes, and information reduction [5].

Data cleaning- It is expected to evacuate noise from the information. The mains issues of information cleaning are missing values, noise, irregularities and copy elimination [3]. The methods utilized as a part of missing qualities are classification, regression, impedance based tools utilizing Bayesian formulation, Decision Tree Induction. Binning, Smoothing, Regression, Clustering is valuable to expel the noisy information from the database.

Data Integration - To union information from various sources into a coherent information store, for example, data warehouse or data cube we utilize this procedure. There are various issues to consider amid data integration, scheme integration can be dubious. This is reffered to as the entity identification problem. Repetition is another critical issue. A third imperative issue in data integration is the recognition and determination of information worth clashes.

Data transformation - Data transformation includes the methods like Smoothing, Aggregation and Normalization.

2 Data preprocessing Demands [13]:

•Data cleaning is by all accounts troublesome for semi structured and unstructured data most of the data is structured .So more work must be done in cleaning semi-structured data.

- Data transformation is an essential stage, but still researcher searching for correct tool
- Limited interoperability.
- For duplication avoidance many techniques used today but more work required in this area
- Query handling is troublesome on heterogeneous information.

3. Study on Pattern Extraction Techniques

Preprocessing of Web Usage Mining include data cleaning, user identification, session identification , path completion, session recreation, transaction identification and formatting[1]. Presently more modern systems for discovering and analysis of pattern are developing. These tools fall into two primary classifications: Pattern Discovery Tools and Pattern Analysis Tools. Pattern discovery draws upon strategies and algorithms created from a few fields, for example, statistics, data mining, machine learning and pattern recognition. They are statistical analysis, association, rule mining [9], clustering, classification and sequential pattern mining. The works done by various authors are arranged into Association rule mining Clustering, Classification and Sequential pattern mining [10].

Association rule mining: Association Rules finds all set of items that have support greater than the minimum support and after that utilizing large item sets to create desired rule that have certainty more prominent than the minimum certainty. An algorithm for discovering ass rule named as AIS was proposed by R.S.Agarwal et al. in 1993. The burden of the AIS algorithm is that it results in unnecessary generation and numerous candidate item sets. The Apriori algorithm exploits the way that any subset of a frequent item set is also frequent item set [4]. The inconveniences are that numerous scans must be done on the database and it has complex time and memory expending.

The upside of AprioriTid algorithm is that the number of entries might be littler than the number of transaction in the database, particularly in the later passes however the expense of exchanging ought to be considered. AprioriHybrid Algorithm Apriori shows improvement over AprioriTid and AprioriTid shows improvement over Apriori in the later passes. FP –Tree algorithm examines the database just twice however it is by all accounts troublesome in incremental and interactive rule mining[9]. Custom fabricated Apriori algorithm that is productive and does successful pattern analysis.

Algorithms Used	Author	Advantages	Disadvantages
Association rule hiding	M.Mahendran	Hide crucial information	--
Multi objective Association rule mining	Yuping Wang	Reduce time consumption and improve performance	To extend it to immediately useb categorical data set
Apriori TID	A Cegret et .al.	Reduce multiple scan	Cost of switching
FP Tree	Han & Pei	Scans are limited only twice	Hard in iterative mining process
RARM	DAS Ng & Woon	Speedy,effictive,scalable	Hard in iterative mining process
Improved Apriori	WANG T ong et.al	Less complexity time	Memory space should be considered
AIS	R.S.Agrawal et. al.	Efficient	min support Items are liminated.
Apriori	Q. Zhao et. al.	Reduce search space and memory cost	Time and memory consuming

Table I Terature Review Association Rule Mining

Fuzzy clustering tech can be utilized to find group that have share similar interest by looking at information assembled in web servers. Jain and dubes et.al in 1998 and Kaufmann et al. 1990 proposed the Agglomerative and Divisive calculation to perform hierarchical clustering. It was adaptable and simple to handle yet obscure and did not visit intermediate clusters. The disclosure of user navigation pattern utilizing SOM is proposed by Etminai et al[4]. SOM is utilized to pre-process the web logs for removing common patterns. Kobra et al. utilized Ant Based

Clustering algorithm to separate frequent pattern for pattern discovery and the outcome was shown in an interpretable format. N.Sujata has proposed another system to enhance web session cluster quality from k-means with genetic programming. The k means was utilized for clustering and GA to enhance the cluster quality. Y.R.Gurav et.al. [12] proposed k- means with fuzzy which simple and effective but time consuming for reducing noisy data.

Algorithms / technique Used	**Author**	**Advantages**	**Disadvantages**
Algomerative Divisive	Jain &kofiman	Easy , flexible	Do not visit intermediate cluster
k- means ,k-mediods		Easier	Not scalable
SOM	Pola britos	Simple	Time consuming
Hierarchical Algomerative clustering	r. shah	Handle large dataset ,efficient	Not powerful
k- means plus fuzzy	Y.R.Gurav et. al.	Simple ,reduce noise	Time consuming

Table II Literature Review Clustering

Bestavros et al. [9] introduced a Markov demonstrating application for web information. To foresee the subsequent link inside of a specific timeframe that a client may take after, the fast order request Markov model. Chen et al. [1] presented the idea which can be portrayed as the sequence of client's request document up to the last one preceding backtracking. Pie et al. [2] presented a WAP-mine algorithm. This technique is quicker than ordinary strategies. WAP-mine is effective than GSP-based arrangement in a wide edge. Mortazaviasl et al. [3] presented a novel projection based algorithm Prefix Span, which bolster sequential pattern mining..

Technique	**Algorithms Used**	**Advantages**
Sequential pattern	General	-
Sequential pattern mining	High utility	Scalable
Classification	Prefix scan	Scalable
WAP tree Association rule	Conditional search strategy	Scalable
Association rule	hasing and purning	Scalable

Table III Literature Review Sequential Pattern Mining

Prediction of group membership for data instancesis done in classification, A few noteworthy sorts of classification strategy [4]including decision tree, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic.

Technique	**Algorithms Used**	**Advantages**	**Disadvantages**
Classification	Naïve Baysian	Simple and efficient	Bad against complex problems

Table IV Literature Review Classification

Pattern Analysis [1]: This is the last stage in the Web Usage Mining process. After the pre-processing and pattern discovery, the got use patterns are investigated to channel uninteresting data and concentrate the helpful data. The strategies like SQL processing and OLAP[8] can be utilized.

4. Pattern Discovery Demands [13]:

- Clustering and Classifications of the documents are done by using many techniques, but more work required for classification and to improve cluster quality
- There is no effective algorithm for pattern extraction.
- Identification of precise user is impractical for mining reason
- The precise sequence of pages client visit is hard to reveal from server site.
- Security, security issues

III. FUTURE DIRECTIONS

Web cleaning is the most imperative procedure as analysts say 65% of the time is spent on information pre-preparing. The web use mining algorithms are more proficient and exact. Yet, there is a test that must be contemplated. In any case, information cleaning gets to be troublesome with regards to heterogeneous information. Keeping up exactness in classifying the data should be focused. In spite of the fact that numerous characterization systems exist the nature of clustering is still an inquiry to be replied[13]. In addition mining rules from semi structure and unstructured as in the semantic web turns into an awesome challenge. This prompts time and memory consumption [4][8][11]. Research work must be focused on these issues as web information run the Web. Privacy of user is also focused in data preprocessing .

IV. CONCLUSION

In this paper we have examined about the research issues and the downsides of the current procedures. More research work should be done on the web mining area as it will control the web sooner rather than later. Semantic web mining is to be focused that is advancing which offers us to defeat the cons of web mining. In spite of the fact that different algorithm and methods have been proposed still work must be done in finding new tools to mine the text data in web.

REFERENCES

1. *Vashi Ms. Dipa Dixit, Fr.CRIT , M Kiruthika," PREPROCESSING OF WEB LOGS", (IJCSE) International Journal on Computer Science And Engineering, Vol. 02, No. 07, 2010, 2447-2452.*
2. *Dr. Sohail Asghar, Dr. Nayyer Masood," Web Usage Mining: A Survey On Preprocessing Of Web Log File Tasawar Hussain", 978-1-4244-8003-6/10@2010.*
3. *Theint Theint Aye "Web Log Cleaning For Mining Of Web Usage Patterns".*
4. *Chidansh Amitkumar Bhatt · Mohan S. Kankanhalli, "Multimedia Data Mining: State Of The Art And Challenges" Published Online: 16 November 2010© Springer Science+Business Media, LLC 2010.*
5. *Hemant Kumar Singh2,Brijendra Singh1, ,"WEB DATA MINING RESEARCH: A SURVEY", 978-1-4244-5967-4/10/$26.00 ©2010 IEEE.*
6. *Xiaoyan Gao ,Rajni Pamnani, Pramila Chawan 1 Qingtian Han, , "Web Usage Mining: A Research Area In Web Mining".*
7. *Wenguo Wu, "Study On Web Mining Algorithm Based On Usage Mining", Computer - Aided Industrial Design And Conceptual Design, 2008. CAID/CD 2008. 9th International Conference On 22-25 Nov.2008.*
8. *http://www.kdnuggets.com*
9. *J.Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology .Volume 11 Issue 4 Version 1.0 March 2011.*
10. *Xue Y ,Chen L, Mao X,Wei P,Ishizuka M (2012) Mandarin emotion recognition combining acoustic and emotional point information. Appl Intell 37(4):602–612.*
11. *Shang F, Jiao LC, Shi J, Wang F, Gong M (2012) Fast affinity propagation clustering: a multilevel approach. Pattern Recognition 45(1):474–486.*
12. *Y.R.Gurav et.al." Fuzzy Based Text Document Clustering Using Side Information in Data Mining", Global Journal of Science and Researches(GJESR) Volume 2, Issue 7, 2015.*
13. *D. jayatchumy et.al."web mining research issuesand future directions –a survey "IOSR-JCE Volume 14,Isuue3, 2013.*