iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** V **Month of publication:** May 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Framework To Integrate Feature Selection Algorithm For Classification Of High Dimensional Data

Durga.S[#1], Lokeshkumar.R[#2]
[#] *Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam*

Abstract: ***The explosive usage of social media produces huge quality of unlabeled and high-dimensional data. The data characteristic with this choice has been tested to be powerful in handling excessive-dimensional facts for effective learning and data mining. In this high-dimensional unsupervised function choice stays a tough task due to the absence of label facts based on which feature relevance is frequently assessed. The specific characteristic of social media statistics further complicate the difficult hassle of unsupervised characteristic selection which makes invalid and identically allotted assumption. In this context bringing approximately new demanding situations to unsupervised characteristic selection algorithmsis a big task. In this paper, we proposed a multiple trouble of function choice for social media records in an unmonitored scenario.Next, analyze the variations among social media data and traditional attribute-fee statistics which looks into the family members extracted from linked statistics to be exploited for selecting applicable functions. Finally, advocate a novel unsupervised feature choice framework, WSLA(Web Server Log Analyzer), for related social media information. Systematically style and implement the general experiments to assess the planned framework on info sets from real-global social media internet sites.The empirical study reveals the learing space of unsupervised feature selection is more powerful and can be extended to different without labeled data with additional information.***

Keywords— Feature Selection; Classification; Web minig:High Dimensional Data; Data Preprocessing; WSLA(Web Server Log Analyzer)

I. INTRODUCTION

The massive and high-dimensional social media data challenges traditional data mining tasks such as classification and clustering due to curse of dimensionality and scalability issues. One traditional and effective approach to handle high-dimensional data is feature selection which aims to select a subset of relevant features from high-dimensional feature space that minimize redundancy and maximize relevance to the targets (e.g., class label). Feature selection helps improve the performance of learning models by alleviating the curse of dimensionality, speeding up the learning process, and improving the generalization capability of a learning model.

A. Web Server

A computer program that is responsible for accepting HTTP requests from clients, which are known as web browsers, and serving them HTTP responses along with optional data contents, which usually are web pages such as HTML documents and linked objects (images, etc.).

B. Log

Usually web servers have also the capability of logging some detailed information, about client requests and server responses, to log files; this allows the webmaster to collect statistics by running log analyzers on log files.

C. Post Parsing Report

The log files are parsed and all the reports are generated after that - usually on a scheduled basis. This can put great strain on a computer as the parsing and reporting are done in one go.

D. HTTP

Every web server program operates by accepting HTTP requests from the client, and providing an HTTP response to the client. The HTTP response usually consists of an HTML document, but can also be a raw file, an image, or some other type of document (defined by MIME-types); if some error is found in client request or while trying to serve the request, a web server has to send an

error response which may include some custom HTML or text messages to better explain the problem to end users.

E. Logging

Usually web servers have also the capability of logging some detailed information, about client requests and server responses, to log files; this allows the webmaster to collect statistics by running log analyzers on log files. In practice many web servers implement the following features also: Authentication, optional authorization request (request of user name and password) before allowing access to some or all kind of resources. HTTPS support (by SSL or TLS) to allow secure (encrypted) connections to the server on the standard port 443 instead of usual port 80. Content compression (i.e. by gzip encoding) to reduce the size of the responses (to lower bandwidth usage, etc.). Virtual hosting to serve many web sites using one IP address.Large file support to be able to serve files whose size is greater than 2 GB on 32 bit OS. Bandwidth throttling to limit the speed of responses in order to not saturate the network and to be able to serve more clients.

Main key performance parameters (measured under a varying load of clients and requests per client), are:

1) Number of requests per second (depending on the type of request, etc.);
2) Latency response time in milliseconds for each new connection or request;
3) Throughput in bytes per second (depending on file size, cached or not cached content, available network bandwidth, etc.).

II. RELATED WORK

In this research, we collected the users' accesses from the user search keyword and graph is generated from it. By applying PageRank algorithm mined from graph to assign importance to accessed pages[8]. In next step, unwanted nodes and weak edges were removed from graph. At last, graph is decomposed into further sub graphs, which depict the behavior of users surfing. The usage of client log file in WUM has become ineffective so it is important to mention that some sort of cleaning was performed to remove the inconsistent and noisy data. User's interests can be mined in better way by grouping the interests based on page visited in a particular time interval.

The analysis of log file is used for proper management of bandwidth and server capacity. Preprocessing step is complex and laborious task. Here discussed the various types of log file in detail based on 19 attributes. In this paper the proposed algorithm two algorithms, first algorithm is to read the log file from any of the three given log file formats and convert the log file data into a database. Second algorithm is to filter out the all the un-interested attributes of web log file. Only "URL" attributes was declared interested. Date, Time, IP Address, and User Agent are some other useful attributes were also dropped. By dropping out such important attributes, the reliability of later phases of Web usage mining cannot be secured[5]. We come to the conclusion that proposed algorithms for data cleaning and data filtering techniques are very weak and needs to be modified.

The data cleaning process on web log file is carried out removing records with graphics and videos format such gif, JPEG, Removing records by checking the status code, removing records applying robot cleaning process. Web log files then the unwanted data's are removed and size of web log file will also become less[15]. When compared to other approaches, novel technique reduces the cleaning time. Preprocessing converts the captured log data in to the valuable information which can be given for further pattern discovery. Preprocessing steps includes Extraction the attributes from the web log which is located in web server, Web logs should be cleaned and removing the repeated and irrelevant data's, Manage the data and put it in relational database or data warehouse[16]. The information after Data Preprocessing can be given to pattern discovery process which includes three different steps in data mining techniques that is clustering, classification and association rules, since irreverent information is removed, so that it speeds up the execution time and provides with valuable information to the users. From the analysis of web log file and Meta data of page contents user profile were derived. In this paper server log file was taken with some mandatory fields such as User-ID, Requested pages and Content-related meta-data.

Preprocessing on log file as basic activity of Web usage mining[11]. To remove irrelevant entries from log file applied two approaches, First approach the data cleaning technique, second is path completion technique at preprocessing level User identification technique was also performed based on IP Address, operating system and browsing software. Session identification step was performed based on log file attributes such as IP addresses, referrer null pages. All the steps of preprocessing were well supported by the proposed algorithms. If the research were able to give some more leverage to WUM users at this level, such as Log files are designed for not for data mining techniques. It is also designed for debugging purposes; User identification and session identification are also very important and essential step[7]. In this research user's sessions were identified by applying website ontology based on the structure of web site and features extracted from the pages. Users are identified based on IP Address and user

stays for long time in inactive mode. A user was represented as triplet of IP Address, date, and time of visit and set of records visited by user in that period of time. The maximal forward reference algorithm MF is used to convert the user access transactions into database. MF algorithm has its own draw back such as we can miss the number of important log transactions as well. Sessions were not grouped further to help in reducing the complexities of other WUM steps.

III. IMPLEMENTATION RESULT

Output refers to the result and information that are generated by the system. It is the main reason for developing the system and based on this, the usefulness and applicability of system are evaluated.Outpurts from computers system are required primarily to communicate the results of processing to users. Efficiently designed outputs enhance the understandability of the information.
According to the requirements of the system, various types of outputs are considered and designed as follows.
Internal outputs,whose destination is within the organization and which require careful design because they are the users main interface with computer.
Interactive outputs, in which the user communication with the Computer is essential.
FTP
Query Engine
GUI
Log Parser

A. Web Server Log Analyzer

The Web Server Log Analyzer is a fast and powerful access log analyzer compared with the LUFS. In this paper, it will give information about your site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. It can be easier to distinguish visitors behind proxies or NAT routers. It can also provide real-time analytics and integrated with email and social media analytics. It is used to give more details about the visitors.

TABLE 1
PERFORMANCE INDICATOR

Technique	Log days	Input parameter	Output parameter	Log records
LUFS(Linked Unsupervised Feature Selection)	30	5	2	2000
WSLA(Web Server Log Analyzer)	100	18	18	5000

B. Capturing Feedback Sessions From Multiple Log File

The log files is parsed to a database in the background. A report is only generated when requested. This type of analyzer is usually more suited for many users as it places less strain on a server.

C. Preprocessing Feedback Session

The log files are parsed to a database in the background. A report is only generated when requested. This type of analyzer is usually more suited for many users as it places less strain on a server.

D. Classification Among Feedback Sessions

Usually web servers have also the capability of logging some detailed information, about client requests and server responses, to log files; this allows the webmaster to collect statistics by running log analyzers on log files.
Virtual hosting to serve many web sites using one IP address.Large file support to be able to serve files whose size is greater than 2 GB on 32 bit OS.Bandwidth throttling to limit the speed of responses in order to not saturate the network and to be able to serve more clients.

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

E. Clustering Report

The functions involved in the development of Real-time Web server Log Analyzer with on-demand Reporting are:

1) Daily Traffic Report
2) Countries Report
3) Accesses Report
4) Searches Report
5) User Agents Report

TABLE 2
COMPARISON WITH DIFFERENT TECHNIQUE

Algorithm	Accuracy	Execution Time	Precision	Recall
Filter	63	64	69	92
Embedded	65	49	72	85
Wrapper	70	37	81	79
Hybrid	73	34	87	75
LUFS	79	29	93	66
WSLA	87	25	96	63

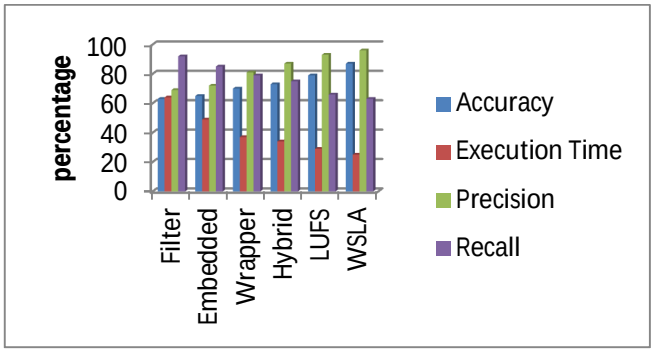


Fig. 1. Compare with different approach

IV. CONCLUSION

The log files ar parsed and every one the reports ar generated then - typically on a scheduled basis. this couldplace nice strain on a pc because the parsing and coverage ar tired one go. During this paper, we have a tendency to use period of time, on-demand coverage within which the log files is parsed to a info within the background. A report is simply generated once requested. this sort of instrument is typically a lot of suited to several usersbecause it places less strain on a server.

V. FUTURE ENHANCEMENT

For further enhancement there are certain modules left over. Those are automations of operational management. So, these modules can be developed for future purpose The software supports most of the user requirements in a friendly manner and also easy to enhance the system in the future. The proposed system is applicable for fulfilling the current requirements but it required modifying to meet future requirements. The software is developed to handle any future complexities and up gradations with minimal effort and changes.

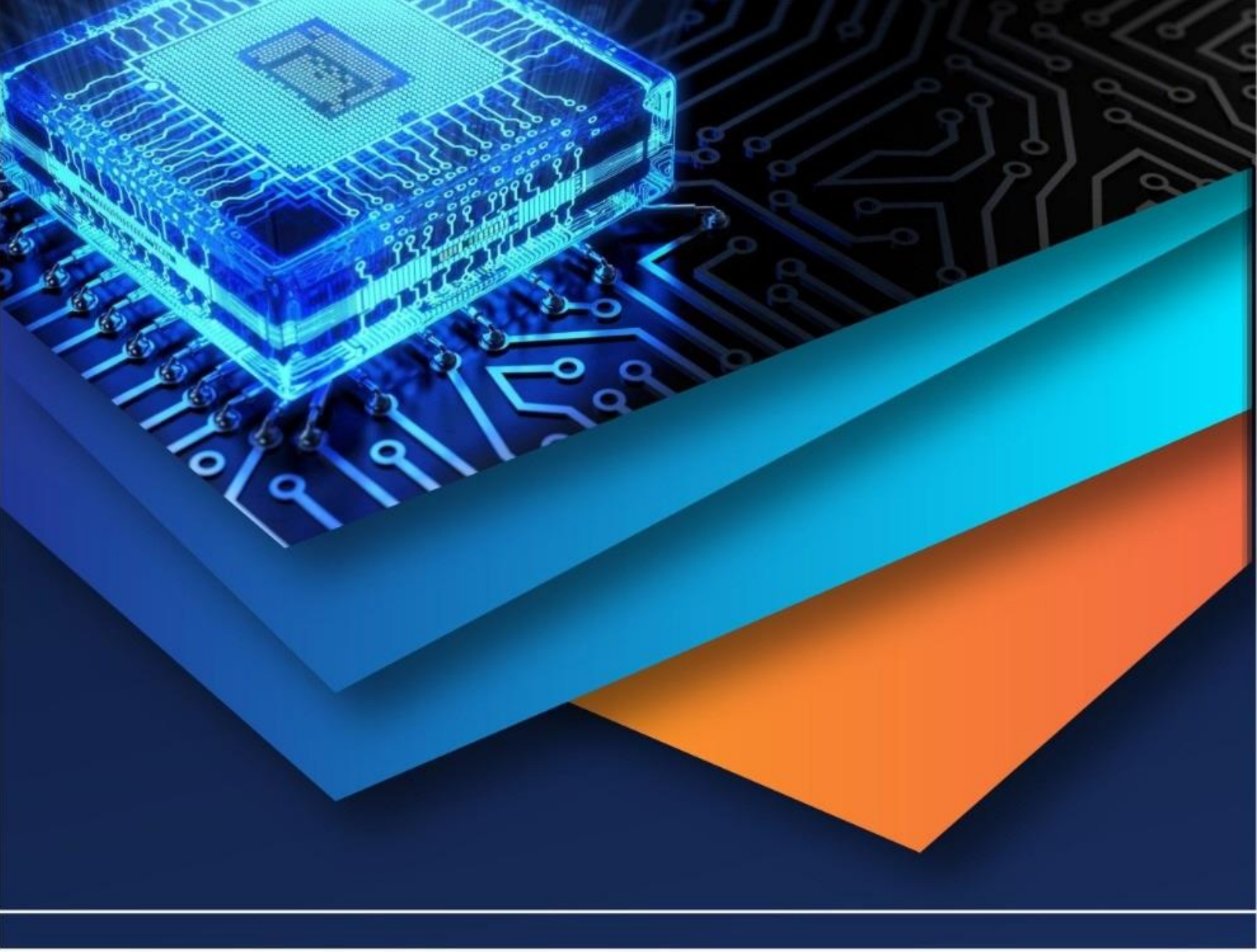
REFERENCES

[1] Jiliang Tang and Huan Liu, Fellow," An Unsupervised Feature Selection Framework for Social Media Data",IEEE transactions on knowledge and data engineering, vol. 26, no. 12, December 2014
[2] J. Tang, A. Salem and H. Liu, "Feature selection for classification: A review," in Data Classification: Algorithms and Applications Boca Raton, FL, USA: CRC Press, 2014.
[3] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in Proc. 16th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2010, pp. 333–342.

[4] Suneetha, K. R. and D. R. Krishnamoorthi . "Identifying User Behaviour by Analyzing Web Server Access Log File." Published in IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[5] Yun, L., W. Xun, et al. "A Hybrid Information Filtering Algorithm Based on Distributed Web log Mining". Third edu 2008 International Conference on Convergence and Hybrid Information Technology 978-0-7695-3407-7/08 © 2008 IEEE DOI 10.1109/ICCIT.2008.39.

[6] Castellano, G., A. M. Fanelli, et al. LODAP: "A Log Data Preprocessor for mining Web browsing patterns". Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence,Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16- 19 , 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY