Educational Web Mining using Weka

V S Kumbhar[#1], K S Oza[*2]

[#]Assistant Professor Department of Computer Science, Shivaji University
Vidyanagar, Kolhapur-416004, Maharashtra, INDIA

Abstract**—** *In this digital era almost every domain is overloaded with voluminous data. This voluminous data need to be processed to get some interesting patterns. Proposed work deals with education domain. Here educational websites data is collected for analysis using site analyzer tool. Analysis is focused on search engine compatibility of websites. There are many parameters related to search engine optimability but proposed work deals only with website accessibility parameter. An open source data mining tool Weka is used to cluster and then classify the data. Class labels are identified as best, good and average depending on website user friendliness.*

Keywords) — *educational websites, site analyzer, search engine, Weka.*

I. Introduction

Internet has removed the geographical bar and has made the world a single cyber village. All information hunts ends on internet. As we know websites are important component of internet. Websites provides a good platform for marketing. Educational institutes can use this platform for attracting students and marketing their courses. Another advantage of educational websites is their institute is open to the whole world without any geographical barriers. Only having website for the institute is not sufficient for business purpose their website should be visited by many students and academicians. Website should have a good and easy interface to the visitors. It should also provide good accessibility to its content. To make these websites reach to the masses they should be search engine friendly and user friendly.

The traditional learning environment of blackboard teaching has evolved over time to power point presentations, smart boards, smart class rooms, online learning, etc. Now it's a trend of eLearning. eLearning provides 24X7 learning , which is as per the ease of students. Here students don't need to be present physically in the class room as per the time table. They can enroll for online courses at more than one institute. They enjoy the freedom of learning any course from any Institute. Before enrolling for the course they can take a survey of all the Institutes offering the courses of his/her interest. For survey students can visit the Facebook page of the institutes, discussion forums, student's feedback posted online etc.

There are so many new technologies coming up. Institutes website should have good collection of courses related to these new technologies. The courses offered online should have good accessibility. This means student should be able to navigate and access website data easily. Website pages should not take more time to download otherwise students might migrate to other similar websites and institute might lose a student. At abstract level website accessibility is a vital factor to be considered in popularization of institutes. Rest of the paper is organized in following way: Section –2 surveys the existing literature, section-3 describes the data collection for analysis, section-4 talks about the work carried out and it is followed by conclusion.

II. Literature Review

S. Balam et. al. have extracted the web information by using various clustering techniques. They have focused on the study and analysis of web content mining techniques, tools and research issues [1]. Monika Yadav and et al. defines the web mining as the application of data mining techniques to extract valuable knowledge from the web content, structure, and usage. How the web technologies are used for making business through websites and for marketing have been presented [2]. Neeraj Raheja and et al have proposed the use of web usage mining to record user behavior. These records are further used to extract data which helps in search engine optimization. They proposed an approach in which web logs are used in cluster forms, reduces the searching time [3]. Pallavi and Sunila Godara worked on iris plants, Haberman's Survival, and Wine Recognition datasets. Using k-means algorithm the datasets is clustered into different clusters and error is also checked [4]. Bharat Chaudhari, Manan Parikh used banking data having 11 attributes and 600 records. The baking data is loaded into Weka and using different clustering algorithms it is grouped into different clusters. The performance of these clustering algorithms is compared. The performance of k-means algorithm is better than hierarchical clustering algorithm [5]. Bhoj Raj Sharma and Aman Paula stated that simple k-means algorithm achieves highest efficiency with consideration of Euclidean Distance and Manhattan Distance methods [6]. Sonam Narwal and Mr. Kamaldeep Mintwal have clustered data repository and J48 classification was well performed on the same data [7].

III. DATA COLLECTION

For research work 173 educational websites data (domain is engineering colleges data in Maharashtra) was collected. Out of these 173 educational websites we could get only 85 websites which provided detail data for analysis. Site-analyzer [9] is a free tool available on internet was used for data collection. This tool gives over all analysis of websites with different parameters. There are many parameters which helps in overall analysis of a website like networking, text, multimedia, accessibility etc. Proposed work focuses only on accessibility of websites so data related to accessibility parameter was collected using this site analyzer tool. The site-analyzer tool performs a multi-criteria SEO (Search Engine Optimization) analysis. Here we collected the data regarding the different accessibility parameters and stored in excel file. This collected data was preprocessed and filtered and then loaded it into Weka for further analysis. The SEO analysis allows us to check the accessibility of web pages and helps to improve the accessibility of the web pages. Due to this it is possible to improve the natural rank of the educational website and also in increase the potential number interested visitors to the educational website. Web accessibility is one of the powerful assets of the website. This accessibility can be visualized from different angles for example one of the angle may be from readability like whether the contents on the website are human eye friendly from font size, font type, color etc. Here in this work page accessibility is focused with following parameters.

1) ***Page weight:*** It is weight of the HTML code. It is recommended that the page weight should be minimum, as it reduces the download time. Page should not have lots of data on it. It can have more links thus making the page weight proper to be considered by search engine while ranking it for indexing.
2) ***Compression:*** Compression of web page is a simple and effective way to save bandwidth and speed up any website. The aim of page compression is to optimize loading time of the web page by reducing the amount of data to be downloaded. It is activated on server-side.
3) ***Page caching:*** Page caching means web page is stored locally as a back-up on user's computer or server. Page caching speedup page access and save download time. Page caching also provides cached resources even when original server is down.
4) ***Download time:*** This criterion allows seeing the average time needed by visitors to download the full content of web page such as images, javascripts and CSS files based on the page weight and global average connection speed.

IV. EXPERIMENTAL WORK

An open source free data mining tool WEKA [8] is used for experimental work. The educational dataset is loaded into Weka. This data set is clustered using simple kmeans clustering algorithm. For clustering only three parameters are considered viz. page caching, compression, and download time. The number of cluster selected is three. The dataset is analyzed by using k-means clustering algorithm and clustered it into three different clusters. There are 89 instances used for the analysis.

Fig.1Screen shot of clustering of educational website data

1) ***Clustered instances:*** In Weka cluster number starts with zero but for simplicity and analysis purpose we have numbered them from one. Following are the statistics of the datasets in each cluster.

Cluster-1	46 (54%)
Cluster-2	26 (31%)
Cluster-3	13 (15%)

2) ***Features of websites in Cluster-1:***The intra cluster analysis shows that websites in cluster-1 are fast because of download time is too minimum as compared to other clusters but page compression is little bit higher than other clusters. So the class label for this cluster is decided as Fast. All the websites belonging to this cluster are fast in accessibility.
3) ***Features of websites in Cluster-2:*** The intra cluster analysis is done. It shows that websites in cluster-2 are moderate because page compression is less than other clusters but download time is more than cluster-1.So the class label for this cluster is decided as Moderate. All the websites belonging to this cluster are moderate in accessibility.
4) ***Features of websites in Cluster-3:*** The intra cluster analysis shows that websites in cluster-3 are slow to download as page compression is little more as compared to cluster-1 and download time is too more as compared to other clusters. So the class label for this cluster is decided as Slow. All the websites belonging to this cluster are slow in accessibility.

Output of Weka clustering is stored in result.arff file and cluster labels are renamed according to above analysis. Here we replace Cluster-1 as Fast, Cluster-2 as Moderate, and Cluster-3 as Slow. The same file is used for classification of the educational websites.

Classification of educational websites: After replacement of the cluster labels, the part of result.arff file is as:

```
8,83,0,4.99,Fast
9,87,0,615.47,Slow
10,79,0,1.14,Fast
11,68,0,230.58,Moderate
12,44,0,426,Moderate
13,67,0,99.69,Fast
```

This result.arff file is classified using four different classification algorithms available in Weka. Here J48, RBFNetwork, Naïve Bayes, and SMO classifier algorithms are used to classify the web accessibility data. The following tables: table 1 and table 2 summarize the result of the classification algorithms when applied on educational dataset. The table 1 shows results based on accuracy and time. However, table 2 displays the result based on errors.

Algorithm used	Correctly classified instances	Percentage of Correctly classified instances	Incorrectly classified instances	Percentage of incorrectly classified instances	Time taken to build model (in sec.)
J48	82	96.4706	3	3.5294	0
RBFNetwork	82	96.4706	3	3.5294	0.04
Naïve Bayes	82	96.4706	3	3.5294	0
SMO	76	89.4118	9	10.5882	0.09

Table 1 Result based on accuracy and time

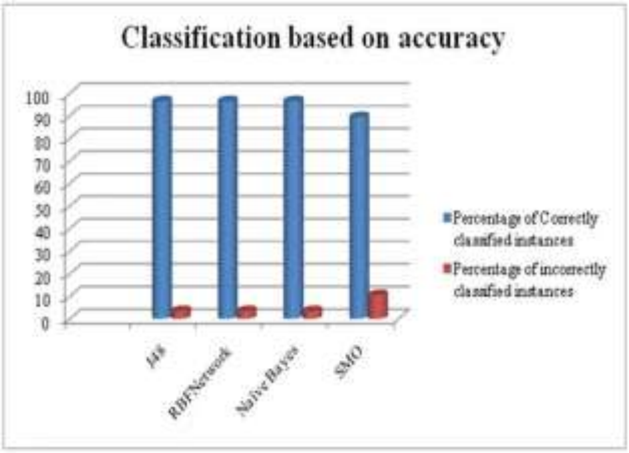

Fig. 2 Classification based on accuracy

Algorithm used	Kappa Statistics	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)
J48	0.9404	0.0235	0.1534	5.9424	34.5270
RBFNetwork	0.9407	0.0282	0.1544	7.1118	34.7467
Naïve Bayes	0.9406	0.0683	0.1572	14.2569	35.3832
SMO	0.8124	0.2510	0.3207	63.386	72.1801

Table 2 Result based on errors

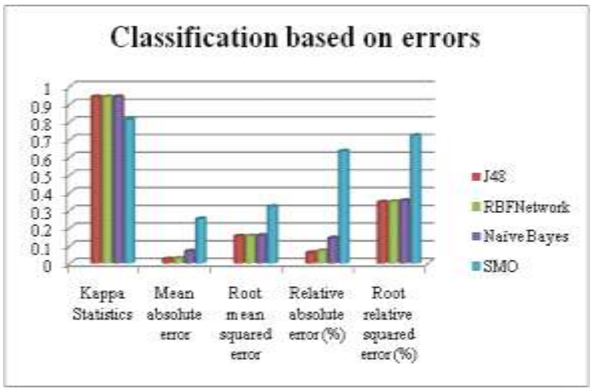

Fig. 3Classification based on errors

From fig. 2 and table 1, it is clearly observed that J48 and Naïve Bayes classifiers give the highest accuracy which is 96.4706%. Out of 85 instances J48 and Naïve Bayes classifiers classifies 82 instances correctly. Also on the basis of time taken to build the model these algorithms took zero second which is less than other classification algorithms. On the basis of both accuracy and time, J48 and Naïve Bayes classifiers are best for classification of educational websites dataset.

As per data in table 2, the kappa statistics of J48, RBFNetwork, and Naïve Bayes algorithms are nearly same. On the basis of kappa statistics these three algorithms works best. However, the other error terms in the table 2, J48 classifier algorithm has less error rate compared to RBFNetwork and Naïve Bayes algorithms. On the basis of error, J48 classifier has less error rate as compared to other classification algorithms. So J48 gives the best performance for classifying educational website dataset. The performance of SMO algorithm is too poor because accuracy is less and time is higher as compared to other algorithms. Similarly the kappa statistics is less and error rate is high compared to other algorithms.

V. CONCLUSIONS

The educational websites are referred by so many students all over world. The web accessibility plays very important role while anyone accessing educational websites. The present study identifies accessibility of educational websites as fast, moderate, or slow. Present study uses Weka as a tool for clustering and classification of data. For the identification of the same present study uses simple k-means clustering algorithm on the educational website data. The clustering algorithm categories the educational websites into fast, moderate, and slow. Then these categorized educational website data is classified by using J48, RBFNetork, Naïve Bayes, and SMO algorithms. The experimental results shows that J48 and Naïve Bayes algorithms accuracy are high and time required to build model is less. On the basis of error J48 algorithm has less errors compared to other algorithms. The present study shows that for classification of educational websites, J48 algorithm works best. The work can be extended for other domains also.

REFERENCES

[1] S. Balam, P. Ponmuthuramalingam, "A Study of Various Techniques of Web Content Mining Research Issues and Tools", International Journal of Innovative Research & Studies, Vol 2 Issue5, ISSN: 2319-9725, May-2013, PP 508-517.

[2] Monika Yadav, Mr Pradip Mittal, "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, Issue 3, ISSN: 2277 128X, March-2013, PP 683-687.

[3] Neeraj Raheja, V. K. Katiyar, "Efficient Web Data Extraction Using Clustering Approach in Web Usage Mining", International Journal of Computer Science Issues, Vol. 11, Issue 1, No.2, January, 2014, ISSN(Print):1694-0814, ISSN(Online): 1694-0784, PP. 216 -225.

[4] Pallavi, Sunila Godara, "A Comparative Performance Analysis of Clustring Algorithms", International Journal of Engineering Research and Applications, Vol 1, Issue 3, ISSN: 2248-9622, PP 441-445.

[5] Bharat Chaudhari, Manan Parikh, "A Comparative Study of Clustering algorithms using weka tools", International Journal of Application or Innovation in Engineering and Management, Volume 1, Ussue 2, October 2012, ISSN: 2319-4847, PP 154-158.

[6] Bhoj Raj Sharmaa and Aman Paula, "Clustering Algorithms: Study and Performance Evaluation Using Weka Tool", International Journal of Current Engineering and Technology, ISSN 2277 – 4106, Vo. 3, No.3, PP 1094-1098.

[7] Sonam Narwal and Mr. Kamaldeep Mintwal, "Comparison the Various Clustering and Classification Algorithms of WEKA Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013 ISSN: 2277 128X, PP 866-878.

[8] www.cs.waikato.ac.nz/ml/weka/

[9] www.site-analyzer.com/