Web SEO analysis: A special reference to University web site

S. N. Kulkarni
Department of Computer Science
Sangola College, Snagola, Dist-Solapur (MS), India

D. R. Kawade
Department of Computer Science
Sangola College, Snagola, Dist-Solapur (MS), India

Abstract- In India various universities having their website which contains different information. On the basis of web page analysis present study identifies which website accesses fastly and which accesses slowly. Website data is collected using site-seo-analysis tool. This data is clustered using K-Means clustering for identification of access speed. This data is classified using different classification algorithm supported by WEKA tool. Experimental result shows that IBk algorithm gains highest accuracy, lower error rate and best time. So IBk algorithm is working efficiently for SEO analysis of website.

Keywords – site-seo-analysis, K-Means, clustering, classification, WEKA

I. INTRODUCTION

Large volume of data and information published in numerous web pages due to the phenomenal growth of web. Web pages don't have any fixed structure i.e. they are dynamic in nature. Web page data are not indexed so it is very tedious and time consuming task for searching the data. In this information age most of the people uses internet for different purpose and hence lots of data is available on internet.

This makes sorting of particular information through billions of webpages and displaying the relevant data makes the tough task for the search engine. Semi structured data is more refined and dynamic than the any database systems data. To analysis such type of data is one of the challenging tasks.

The emerging field of web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular in text documents published on the Web. Data Mining is useful to extract meaningful and valuable information from huge amount of data. Web mining is an important area in data mining where we extract the interesting patterns from the contents.

There was a rapid expansion of activities in the Web mining field. Web usage mining is nothing but discovering user access patterns from Web usage logs. Web mining is identifies valuable knowledge from the hyperlinks (Web Structure Mining), also extracting/mining useful information or knowledge from Web page contents (Web content mining). Web pages consist of text, images, audio, video, or structured records such as lists and tables. Web content text mining is one of important research issue which deals with topics such as classification and clustering of text document, subject discovery and tracing, identifying patterns etc. To improving the visibility of website in search engine, Search Engine Optimization (SEO) process is used. For optimization purpose various types search are considered such as image searching, multimedia search, link search etc.

SEO uses algorithm to optimize web page code and contents which is useful for fetching the web page by search engine.

Generally analysis identifies duplicate content, link related structure and inbounds links of website. Also it may identify keywords and descriptions, malware etc. files. These all aspects are useful to identify efficiency of search results on search engines.

SEO is useful for getting higher rank in search engine which is beneficial for marketing strategy. Higher ranking means getting noticed on first few places in search engine.

II. LITERATURE REVIEW

2.1 Watermark embedding algorithm –

For extracting information, processing time is too much. They have defined search engine architecture by using distributed design approach. The architecture is able to navigate a large number of websites and also it produces the capacity to execute special types of large scale web data mining tasks. This architecture may reduce the time or processing power of search engine. [1]

In this paper, using the class distribution of query classification some class features are newly derived for providing ranking for classes. One-click and multiple-click queries are used in this experiment. A one-click query shows results when clicked on whereas multiple-click queries assumed the last click for the target document. [2]

To improve the business dictating author has defined web service analysis. This analysis has provided the work to perform effective keyword analysis based search in terms of text search as well as for image search over the web. [3]

User query relevance factor is used to perform effective search in web environment. Three factors 1.keyword based analysis 2.user recommendation analysis 3.user web service visit analysis are used to analyze relevancy of the query. Ranking criterion is decided on these all factors. Based on this user can get the best web service as well as recommend other for the best service selection.[4]

The author has shown SEO research, which is useful to increase website ranking and to fetch the website traffic. The content, the keyword and the link building is required for efficient SEO structure to support high ranking of the web pages. [5]

To attracting more users towards a website, author has proposed new novel method by using Google's Page Rank algorithm. By using the algorithm Author has optimized the web page in Google to improve visibility and profitable deal for an organization. [6]

III. EXPERIMENTAL WORK

3.1 Working Environment

Experimental evaluation carried on Windows XP operating system with Intel Core i3 3.3GHz processor and 2 GB RAM. WEKA data mining tool is used for experimental work.

3.2 Weka

University of Waikato of New Zealand has developed one data mining tool called as Weka, which is developed in Java. Number of data mining algorithms was implemented in Weka. These algorithms are grouped into different grouping according to the generated rule by algorithm. We can apply specific algorithm directly to generate result on required dataset. Weka is freely available on the Internet and it is handled easily by user. Due to this for classifying website data different classification algorithms from WEKA are applied in the present study. [7]

3.3 Site SEO Analysis tool

Site SEO Analysis website is useful for search engine optimization (SEO) of web pages and website, which is free. To analyze the entire web site, user has to create account and add specific URL of website. Site SEO analysis analyze website and display recommendations for better optimization of the web pages or websites. That will be helpful to give higher rank in search results.

The analysis display factors like a site's architecture, content, linking structure, social media efforts, and its trust on the internet.

Some components are used to deliver or reference the web pages. Those components are referred by architecture.

Almost anything contained on the page with the exception of links on the page is referred by Content. Links on the page are referred by Links because they are more of a combination of web references (Architecture) and content for the users. To determine the reputation of the website trust content is used in SEO analysis.

Analysis report generates a summary and a detailed SEO Analysis. It provides an overall score and important recommendations for better optimization of the web pages or website. [8]

3.4 Data Collection and Preprocessing

For SEO analysis, list of websites of various Universities is taken from UGC website [12] and from Wikipedia [13]. SEO analysis is done on the website www.site-seo-analysis.com. From result of SEO website present study generated dataset which consists of 9 attributes and 103 instances.

This dataset is not suitable for analysis purpose. To solve this problem present study has applied data preprocess ing technique. Some fields are null, we have replaced these fields by 0. Apart from these 9 attributes some attributes such as site name, Total Score, Trust are not useful for the analysis purpose. So we have removed these attributes by using WEKA's Remove filter. Finally attributes Architecture, Content, Links and Page speed are selected for analysis purpose.

3.5 Cluster Analysis

This Preprocessed data is clustered by using WEKA's K-means clustering Algorithm. We have generated 3 clusters. These 3 clusters are renamed as Slow, Fast and Moderate. After experimental work 3 clusters are generated namely cluster 0, cluster 1 and cluster 2. 71(69%), 16(16%), 16(16%) instances are present in cluster 0, cluster 1 and cluster 2 respectively.

After observing the cluster 0 page speed attribute is 0.6 which is very less as compared to cluster 2. Similarly links and architecture values are higher than cluster 2. Speed of web browsing is dependent on links, architecture and page speed attributes. Considering values of these attributes cluster 0 is named as slow.

Cluster 1 consists of 16 instances and named as moderate, because value of page speed is higher than cluster 0 and lower than cluster 2. Similarly value of content, link is lower than the cluster 0.

Page speed value for cluster 3 is higher than cluster 0 and cluster 1, similarly value of link and architecture is less than the cluster 0 and cluster 1. It tends to fast browsing, so cluster 2 named as fast.

For classification purpose present study, cluster 0, cluster 1 and cluster 2 are replaced by Slow, Moderate and fast respectively. This file is stored as result.arff.

3.6 Classification of Website Data

Classification is one of the predictive data mining techniques. Classification works in two steps. Firstly it builds model from training dataset and then it is applied on test dataset. On the basis of accuracy of classification model; rules are applied on the test dataset. Based on the output of algorithm, it is categorized as follows:

Tree Structure: - Output of these types of algorithms is in the form of decision tree. In the tree intermediate nodes are represented by attributes and leafs are represented by class labels. Some of the decision tree algorithms are ID3, C4.5 and CART. [9]

Rule Based Algorithm: - Output of these types of algorithms is in the form of IF-THE type rules like Grammar-based genetic programming algorithm (GGP), AprioriC, IF-THEN, genetic algorithm etc. [9]

Distance Based Algorithms: -In this type of algorithms similarity measures or distance are considered. Similar or nearest items are grouped into classes. K nearest neighbors is one of most well-known algorithm of this type.

Neural Networks Based: - It consist of a set of nodes (units, neurons, processing elements) having in which some nodes acts as input and output nodes. Each node has its own node function as well as weights useful for computation. Output of Neural Network is primarily calculated on the basis of connections and their weights. Multilayer perceptron, a hybrid Genetic Algorithm Neural Network (GANN) etc. are some examples of neural network algorithms. [10][11]

Statistical Technique:-In Statistical classification individual items are placed into groups. Groups are generated based on some sort of the quantitative information of characteristics. These groups are inherent in the items and based on a training set of previously labeled items. Linear discriminate analysis, least mean square quadratic and kernel are examples of this type of technique.

3.7 Experimental Result

Following table shows correctly and incorrectly classified instances for different algorithms. Also it show percentage and time require to build classified model.

Table 1: Time and Accuracy gained by different classification Algorithms.

Sr.No.	algorithm name	correctly classified instances	percentage	incorrectly classified instances	percentage	Time in sec
1	J48	95	92.23	8	7.77	0.02
2	NaiveBayes	97	94.18	6	5.82	0
3	RBFNetwork	98	95.15	5	4.85	0.09
4	IBk	98	95.15	5	4.85	0
5	AdaBoost	98	95.15	5	4.85	0
6	AttributeSelectedClassifier	96	93.2	7	6.8	0
7	HyperPipes	94	91.26	9	8.74	0
8	JRip	98	95.15	5	4.85	0.02

Following Graph shows correctly classified instances of various algorithms.

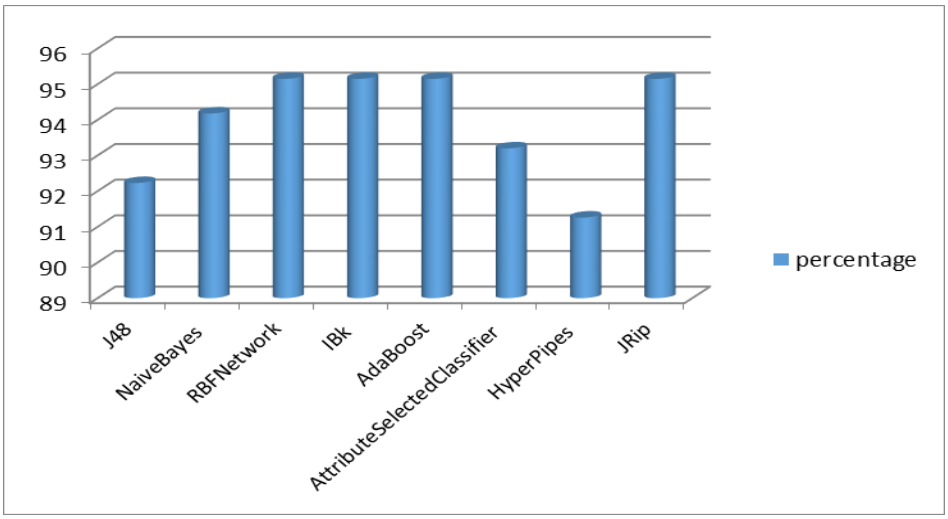

Figure 1. Graph showing correct classified instance Vs different algorithms for SEO analysis dataset

Following table shows kappa statistic and simulation of errors.

Table 2: Simulation of Errors

Sr. No.	Algorithm name	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Average
1	J48	0.8296	0.0719	0.2246	0.22317	0.56253	0.27055
2	NaiveBayes	0.8736	0.0531	0.1636	0.164791	0.40985	0.197835
3	RBFNetwork	0.8959	0.0366	0.1832	0.113566	0.45892	0.198071
4	IBk	0.8947	0.0452	0.1777	0.140385	0.44499	0.202069
5	AdaBoostM1	0.8923	0.1159	0.207	0.360015	0.51857	0.300372
6	AttributeSelectedClassifier	0.8492	0.0683	0.2095	0.21222	0.52473	0.253686
7	HyperPipes	0.8015	0.3817	0.4107	1.185236	1.02871	0.751587
8	JRip	0.8923	0.0545	0.1777	0.169078	0.44515	0.211607

Following graph shows kappa statistic values obtained by various algorithms.

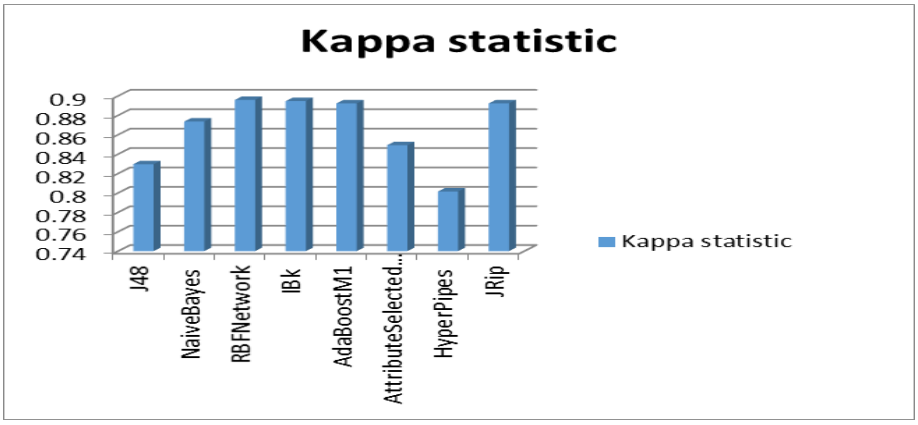

Figure 2. Kappa Statistics of different algorithms

Following graph shows simulation of errors for various algorithms.

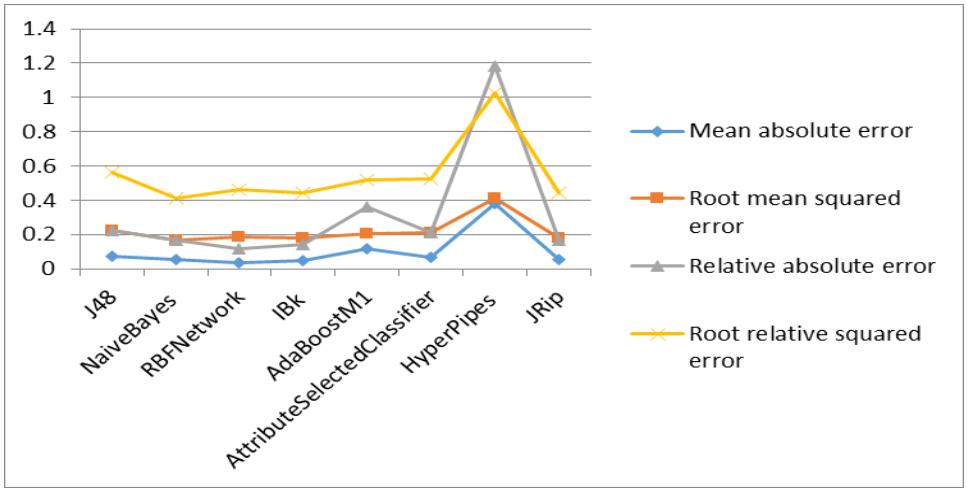

Figure 3. Figure 3: Graph of Simulation of Errors

IV OBSERVATIONS

Accuracy obtained by RBFNetwork, IBK, AdaBoost and Jrip is 95.15 which is higer than other algorithm(See Figure 1 and Table 1) but for Hyperpies algorithm accuracy is lowest i.e. 91.26. 93.93 is average accuracy obtained by all algorithms. 98 instances are correctly classified by RBFNetwork, IBK, AdaBoost and Jrip algorithms which is highest. Lowest correct classified instances obtained in Hyperpies are 94. In terms of accuracy, RBFNetwork, IBK, AdaBoost and Jrip algorithms works best among all other algorithm.

On the basis of both accuracy and time, IBK and AdaBoost are best algorithms for classification of SEO analysis of websites. Because their accuracy is highest as well as time required to generate a model is less than other algorithms. Time required to generate a model in RBFNetwork and Jrip algorithm is more. From time point of view both algorithms are worst for this classification purpose.

Reliability of data collected and validity of the data is differentiated by Kappa statistics. It is used to access particular measuring cases. The average value for Kappa statistics is around 0.87. Kappa statistic value for algorithm RBFNetwork, IBK, AdaBoost and Jrip is approx. 0.89, which is near to 1. According Kappa Statistic, value which is near to 1 is best algorithm.[9] So above these algorithms are best for this dataset.

Table 2 shows different errors obtained by various algorithms. It is discovered that the highest error is found in HyperPipes with an average score of around 0.75 followed by AdaboostM1, J48, AttributeSelectedClassifier and JRip, where the rest of the algorithm ranging normally around 0.27 except NaiveBayes, RBFNetwork, IBK algorithm has error rate 0.20 which is lower. An algorithm having a lower error rate is acceptable because it tends to accurate

classification. In this case NaiveBayes, RBFNetwork, IBK are having lower error rate. IBk algorithms works efficient for SEO analysis of Website because it gives highest accuracy, less time and lower error rate.

On the basis of such considerations, the algorithm uses a different color image multiplied by the weighting coefficients of different ways to solve the visual distortion, and by embedding the watermark, wavelet coefficients of many ways, enhance the robustness of the watermark.

V CONCLUSION

In this digital world, every university in India has their own website which contains huge amount of information. Due to the different contents in website, accessing of website is varied i.e. some websites are accessed very fastly, whereas some are accessing very slowly.

SEO technique is useful to analyze the website. Present work uses site-seo-analysis tool for analysis of university websites in India. The purpose of the paper is to identify which website is accessed very fastly and which is very slowly.

By using SEO analysis tool, we have generated dataset which consists of 9 attributes and 103 instances. Real life data is dirty data so it is necessary to preprocessing of data. After preprocessing, data will be clear data. This clear data is used for clustering analysis. Proposed work uses K-means clustering algorithm implemented by WEKA. By using K-means clustering algorithms, 3 clusters are generated. For classification purpose these clusters are renamed as Slow, Moderate and fast and result is stored in result.arff file.

On the basis of both accuracy and time, IBK and AdaBoost are best algorithms for classification of SEO analysis of websites. Different errors are obtained by various algorithms. From these we have obtained average error rate. We observed that Naive Bayes, RBFNetwork, IBK are having less error rate. For SEO analysis of website we found that IBk algorithm works efficiently in terms of exactness, time and error rate.

We may improve result of SEO analysis by using bagging and boosting techniques in future. Also we may use different SEO analysis tools for better results.

REFERENCES

[1] Rajesh Singh, S.K.Gupta, "Search Engine Optimization - Using Data Mining Approach", 'International Journal of Application or Innovation in Engineering & Management (IJAIEM)'Volume 2, Issue 9, September 2013 ISSN 2319 – 4847

[2] Rajesh Singh, S.K.Gupta,"An approach for Search Engine Optimization using Classification - A data Mining Technique",'IPASJ International Journal of Computer Science (IIJCS)'Volume 2, Issue 4, April 2014 ISSN 2321-5992

[3] Ping-Tsai Chung, "A Web Server Design Using Search Engine Optimization Techniques for Web Intelligence for Small Organizations", Proceedings of IEEE Conference, pp 1-6, 2013.

[4] Minky Jindal, nishaKharb , Data Mining in Web Search Engine Optimization and User Assisted Rank Results,'International Journal of Computer Applications' (0975 – 8887) Volume 95– No.8, June 2014'

[5] Gurpreet Singh Bedi, Ms.Ashima Singh, Analysis of Search Engine Optimization (SEO) Techniques,International Journal of Advanced Research in Computer Science and Software Engineering 4(3), Volume 4, Issue 3, March 2014, pp. 563-566 © 2014, IJARCSSE ,ISSN: 2277 128X

[6] Vinit Kumar Gunjan, Pooja, Monika Kumari,DrAmitKumar,Dr (col.) Allamapparao, "Search engine optimization with Google",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012, ISSN (Online): 1694-0814

[7] WEKA at "http://www.cs.waikato.ac.nz/~ml/weka" accessed on 12-01-2016

[8] www.site-seo-analysis.com

[9] Dipak R. Kawade, Kavita S. Oza, " SMS Spam Classification using WEKA", International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 5 Issue ICICC (May 2015)

[10] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás"Data Mining Algorithms to Classify Students" Educational Data Mining 2008 The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings

[11] Raj Kumar, Dr. Rajesh Verma "Classification Algorithms for Data Mining:A Survey" International Journal of Innovations in Engineering and Technology (IJIET)

[12] www.ugc.ac.in accessed 12-05-2016

[13] https://en.wikipedia.org/ accessed 13-05-2016