Feature Based Identification of Web Page Noise through K-Means Clustering

S. S. Bhamare, B. V. Pawar

Abstract*: Web pages has pieces of information which are of unequal importance like navigational bar, copyright notice, links, advertisement etc. and these are considered as noise or insignificant items of web page for web mining. Web page informative content is only useful for performing effective web mining task and presence of noise on web page can hamper the result of this task. Web page has several features including information location, occupied area and its contents. Content data in different portions of an internet web page has dissimilar significance weights according to its location, occupied location and content that are considered to be features of the web page. The position of contents and importance of contents play a vital role in identification of noise in web pages for removal. In this paper web page feature based method is proposed for identification of noise from web pages. K-means clustering technique is applied to classify main content information and noise content information into two clusters of web pages based on these features. For performance evaluation of clustering method, accuracy, precision, f-measure, and recall are calculated.

Keywords: Noise, Feature Extraction, Clustering, HTML Tag, Tag Weight, Web Pages.

I. INTRODUCTION

In the huge World Wide Web network, web pages contain large amounts of informative data. The researchers always want only useful content from the web pages that useful content needs to be processed. Data mining on web become a main task for detecting useful data from the web. Usually web information has large amounts of noise data and that data is not useful for mining such as navigation bars, links, advertisements, copyright notices etc. Demarcating important information from noisy content is essential because the noise misguides user interest. Performance of Web mining can be improved by identifying and removing noise from Web pages.

This paper proposes web page feature based method which is used for identification and removal of noise from web pages and helps efficient web mining operations.

This method group's data into two clusters such as noise data and non-noise data using two feature variables (i.e. final tag content weight and location feature weight) of web pages through k – means clustering technique. Web page clustering automatically categorizes data into different groups.

Revised Manuscript Received on January 5, 2020
***** Correspondence Author
S. S. Bhamare, School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, India
Email: ssbhamare.nmu@gmail.com
B. V. Pawar, School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, India
Email: bvpawar@hotmail.com

In performance evaluation, metrics such as Accuracy, Precision, F-measure, and Recall are used.

II. RELATED WORK

Noisy web page data cleaning is an important task, by retrieving and extracting main content data and eliminating noise data from Web pages. Many researchers have worked in this area.

Yong Zhang et al. proposed a different method to find the main content data block of the web page using web page purification based on improved Document Object Model (DOM) and statistical learning. In this approach produced block tree structure which is helpful for information retrieval and information extraction and web page classification using statistical learning [2].

Li Xiaoli et al. suggests a new technique to removes noise data in web page classification. Initially it displays the presentation of a web page based on HTML tags, then it uses a new distance formula and removes the noise data using similarity measure [3].

Cai Deng et al. proposed Vision Base Page Segmentation (VIPS) algorithm. It uses full web page layout features and some experimental rules to partition the web page at the semantic level. In this method main restriction is performing visual rendering and partition of web pages is resource intensive [4].

Zhao Cheng-li et al. uses new style tree model to identify and remove web page noise it uses new. It determines whether element node is noisy or not through information based measures. This proposed technique is able to increase the mining outcome considerably [5].

Bhamare et al. discussed different supervised and unsupervised web page noise cleaning techniques [6].

III. SYSTEM FLOW DIAGRAM OF PROPOSED METHOD

The process flow diagram (Figure 1.) shows overall web page noise detection and removal process of proposed method.

This proposed web page feature based method consists of following four steps,

i) **Feature extraction:** This step of proposed method determines and extracts possible important features from webpage's.
ii) **Feature selection:** In this step we find the best required features set for proposed method.

iii) **Clustering:** In this step of clustering web page data is separated into two clusters, Cluster with noise data and Cluster with non-noise data.

iv) **Performance evaluation:** In this step, metrics such as Accuracy, Precision, F-measure, and Recall are used for performance evaluation.

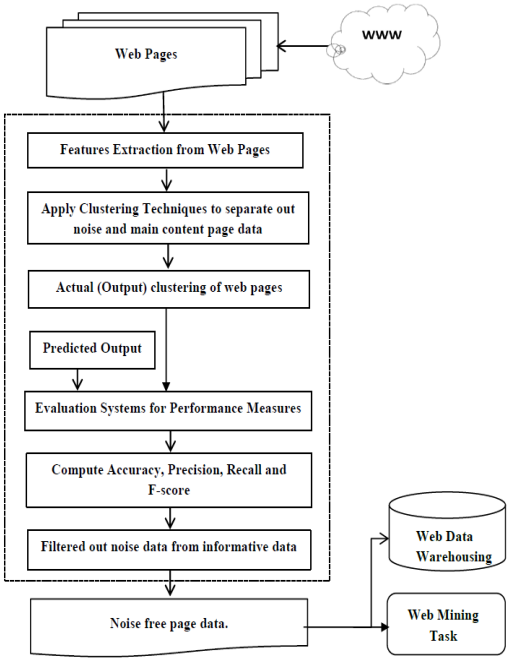


Figure 1: Process flow diagram of proposed web page feature based method

The above steps help to accurately identify noise data and non-noise data from web pages for effective web mining operations.

A. Web page feature extraction

Web page content position and spatial features give clues to recognize noise content and main or informative content of web pages. Here, in this proposed method two features are used for clustering of web page data.

Feature 1: The first feature relates to HTML Tag information of web pages i.e. Tag Weight of each tag including their corresponding tag through first computing a tag weight using text density of each tag based on importance of content held by the tag, importance values are assigned to each tag as shown in Table: 1. Tag Weight will help to find the noise which is more formatted containing a small text and main content usually lengthy and less formatted [9].

Feature 2: The second feature is rooted on location based content importance of web pages, in different set of web pages of different categories; it is found that most of the web pages organize their information in similar structure or in the same location of web pages. These features are useful to decide which part is the important parts and which parts is unimportant of web pages. Typically, Designers of web pages like to place relevant and important information in middle or center of the web page and also place the copyright notice on the footer or on the bottom side of web pages and the navigation bar on the left hand side or on the header of web pages. With the help of this location based content information, it is possible to distinguish their content importance and assign importance values to them. Importance values assigned to content tags of location blocks are as shown in Table 2. Therefore, location content features are also included in this experiment [4] [10].

Table 1

Tag Weight (importance Value) Assigned	Web Page Tag Content Importance
0	Used for tag without noise (i.e. Main Content Tag), also range of values 0.01 to 0.50 are used to assign those tags of which the information is relevant to the main content. Common main content or informative tags contains <table> , <form> , , <tr>, <td>, and <body> etc.
0.01 to 0.50	
1	Used for tag with noise content, also a range of values 0.51 to 0.99 is used to assign those tags of which the substance is not relevant to the main content such as advertisement, copyright, decoration, etc. Common noise content tags include <a>, <img>, <font> and etc.
0.51 to 0.99	

Table 2

Different Location levels (Importance Values) of Web Pages	Location Based Content Importance
Level 1 (Value 1) (Top Panel)	Noisy information such as advertisement, copyright, decoration, etc.
Level 2 (Value 2) (Right Side Panel)	Useful data, but not exactly related to the subject of the web page or main content for example directory list, navigation bar etc.
Level 3 (Value 3) (Left Side Panel)	Related information to the web page theme, but not with prominence, for example topic index, related topics etc.
Level 4 (Value 4) (Bottom Panel)	Noisy information such as advertisement, copyright notice, declaration, etc.
Level 5 (Value 5) (Central Panel)	The most prominent part of the page, such as headlines, main content, etc.

These features are extracted through web pages such as tag weight or text density and location importance. Importance values can be assigned to these features according to their importance as shown in Table 1 and 2.

The following table (3) shows sample feature vectors of two categories of web pages,

Table 3: Sample features of two categories of web pages

Source Category	Final Tag Feature Weight of web pages (Feature 1)	Final Location Feature Weight of web pages (Feature 2)	Source Category	Final Tag Feature Weight of web pages (Feature 1)	Final Location Feature Weight of web pages (Feature 2)
CNNIBN_ Technology	1	1	Times of India_Sport	0.8	1
	1	2		0.85	2
	1	2		0.6	2
	0	5		0.9	2
	0	5		0.65	2
	0.75	1		0.77	1
	0.85	1		0.75	1
	0.3	2		0.65	1
	0.7	2		0.88	1
	0.42	3		0.67	2
	0.35	3		0.95	1
	-	-		-	-

B. Clustering of web page data

This proposed method uses K- Means clustering algorithm for clustering web page data. K Means Clustering are the simplest unsupervised learning algorithms used to solve clustering problems. Clustering or Cluster analysis is the task through which it makes a group of similar items or objects in one cluster and other than those are groups in another cluster.

In Clustering data are grouped into two clusters such as Noise data and Non-Noise data, using two feature variables i.e. Final Tag Weight and Location Feature Weight of web pages.

IV. IMPLEMENTATION AND RESULTS

This experiment is carried out on web page feature dataset of different Web sites of web pages.

A. Dataset Used

The experimental evaluation is done on three news web sites, Times of India, ABB News and CNNIBN of different categories of web pages such as Main, Technology and Sports Pages.

In this experiment, we collect feature samples from 22 to 25 web pages of each category i.e. Technology, Sports and Main pages from CNNIBN, ABB News and Times of India web sites.

B. Performance Evaluation of WPFB Method through Clustering Techniques

This proposed method is implemented using MATLAB, for evaluating performance of clustering techniques. Here Accuracy, Precision, F-measure, and Recall as a quality measure are used. These measures are derived from confusion matrix, in which prediction class data is compared with actual class data, below table shows confusion matrix.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

The above quality measures can be determined by equations 1, 2, 3, and 4.

Classwise calculation:

$$Accuracy\ (Class) = \frac{(TP\ (Class) + TN\ (Class))}{Total\ Samples} \quad \text{(Eq. 1)}$$

$$Precision\ (Class) = \frac{TP\ (Class)}{(TP(Class) + FP(Class))} \quad \text{(Eq. 2)}$$

$$Recall\ (Class) = \frac{TP\ (Class)}{(TP(Class) + FN(Class))} \quad \text{(Eq. 3)}$$

$$F\_Score\ (Class) = \frac{2 * TP\ (Class)}{(2 * TP(Class) + FP(Class) + FN(Class))} \quad \text{(Eq. 4)}$$

Average calculation:

$$Accuracy = mean(Accuracy)$$
$$Precision = mean(Precision)$$

$$Recall = mean(Sensitivity)$$
$$F\_Score = mean(F\_score)$$

V. FEATURE BASED NOISE IDENTIFICATION FROM WEB PAGES THROUGH K-MEANS CLUSTERING TECHNIQUE

This proposed web page feature based method (Algorithm 1) uses well known K-Means clustering technique for noise detection and performance evaluation. K-Means clustering can division number of items or objects into n number of clusters in which each items or object belongs to the cluster with the nearest mean. This clustering technique produces exactly n different clusters of best possible distinction [13].

A. The proposed algorithm

Algorithm: Feature Based Noise Detection through K-Means Clustering

Input: Two feature vectors

: **Call K-Means Clustering for two clusters with two features**

Output: Results, Accuracy of calculation, Precision, Recall and F-Score

---

Begin

For clustering of Web pages,

Step 1: Read Feature Vector .xls File into file1

i.e. input file = Clusterfeature.xls

Step 2: Assign two feature columns of file1 into feature file2

i.e. feature vector = input file (Feature1, Feature 2)

Step 3: Assign expected output of Feature file1 into outputfile1

i.e. exp_output vector = file1 (expected output)

Step 4: Call K-Means Clustering for two clusters on file2 with two features

Step 5: Assign actual clustering output into outputfile2

i.e. ***actual_output vector = Kmeans (feature vector, 2)***

Step 6: Call function for Confusion Matrix Statistics with passing two parameter values

i.e. confusion matstat (outputfile1, outputfile2)

Step 7: Compute **Accuracy, Precision, Recall and F-Score**

End

Algorithm 1: Proposed algorithm for web page noise identification & removal and evaluating performance of clustering.

B. Performance Evaluation of Each Category of Web Pages through K- Means Clustering Technique

The Experimental results of K-Means clustering are shown with sample features and outputs (expected and actual) in Table 4. The web page data are assigned into two different clusters.

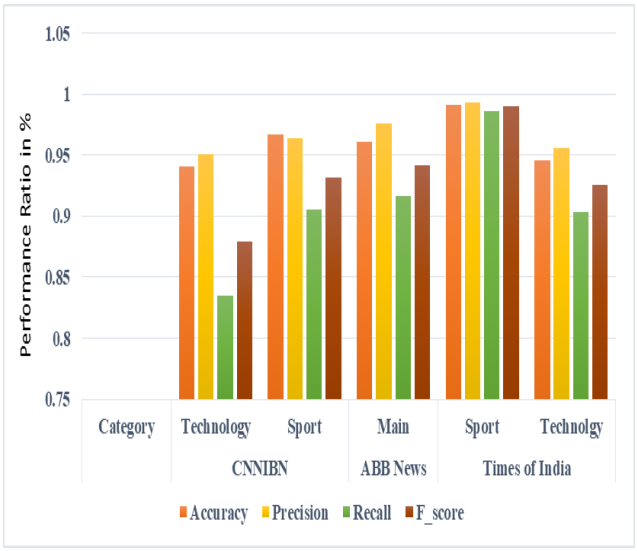
Table 4: Sample K means Clustering Results of one category of web pages

Source Category	Final Tag Feature Weight of Web Pages (Feature 1)	Final Location Feature Weight of web pages (Feature 2)	Expected Output	Actual Output Using K-Means
CNNIBN_ Technology	1	1	2	2
	1	2	2	2
	1	2	2	2
	1	2	2	2
	0	5	1	1
	0	5	1	1
	0.75	1	2	2
	0.85	1	2	2
	0.3	2	2	2
	0.52	2	2	2
	-	-	-	-

The category-wise performance evaluation shows high accuracy for sports pages while technology pages contain more unidentified noise. The following Table 5 and Graph 1 demonstrate category-wise evaluation of K-Means clustering techniques.

Table 5: Category Wise Performance Measures of K-Means Clustering

Name of Web Site	Source Category	Accuracy	Precision	Recall	F_score
CNNIBN	Technology	0.9407	0.9503	0.835	0.8796
	Sport	0.9665	0.9641	0.9053	0.9317
ABB News	Main	0.9609	0.9757	0.9167	0.9421
Times of India	Sport	0.9908	0.9932	0.9861	0.9896
	Technology	0.9458	0.9553	0.9029	0.9253



Graph 1: Category-wise Performance Measures of K-Means Clustering

VI. RESULT DISCUSSION

The experimental evidence of K-Means clustering shows category-wise performance evaluation done through Accuracy, Precision, Recall and F_Score. In these results, the sports pages of TIMES OF INDIA contains noise which is identified by the algorithm correctly with 99% accuracy while on CNNIBN pages sometimes the location noise cannot identify so the accuracy is 96%. The technology pages of both TIMES OF INDIA and CNNIBN contains similar issues of importance identification so misleading the results. The technology pages noise removal accuracy is 94%. The main page contents have the both issues of location and importance with 96% results. The overall performance evaluation of noise removal is 96% accuracy as well as precision with 90 percent sensitivity that gives recall.

Considering both precision and recall the F_socre is calculated which is 93%. The drop of recall with respect to precision is the reason for the less value of F_Score.

VII. CONCLUSION

The Web pages of different categories like sports, technology as well as main page of news are considered for robust evaluation of the system. The position of contents and importance of contents play a crucial role in identification of noise in web pages. The clustering of web page contents effectively classify noise from important data contents in web pages. The effectiveness of the features and learning mechanism is evaluated using accuracy, precision and recall and F Score.

The K- Means clustering algorithm can identified page contains noise correctly with considerable accuracy and gives the better results.

REFERENCES

1. S. Sumathi et al., Introduction to Data Mining and its Applications, Studies in Computational Intelligence, Volume 29, Springer.
2. Yong Zhang et al., Algorithm of Web Page Purification Based on Improved DOM and Statistical Learning in 2010, International Conference On Computer Design And Applications (ICCDA 2010).
3. Li Xiaoli et al., Innovative Web Page Classification through Reducing Noise, Journal of Computer Science and Technology Volume. 17 No. 1. Jan-2002.
4. Cai Deng et al., Vision Based Page Segmentation (VIPS) Algorithm, Microsoft Technical Report: (MSR-TR-2003-79), 2003.
5. Zhao Cheng-Ii et al., A Method of Eliminating Noises in Web Pages by Style Tree Model and Its Applications, Wuhan University Journal of Natural Sciences Volume.9 No.5 2004.
6. S.S.Bhamare, et al., Survey on Web Page Noise Cleaning for Web Mining, International Journal of Computer Science and Information Technologies (IJCSIT) Volume 4 Issue 6, Nov-Dec. 2013, ISSN: 0975-9646.
7. Isabelle Guyon et al., an Introduction to Variable and Feature Selection, Journal of Machine Learning Research, Volume. 3, No.2, pp. 1157-1182, 2003.
8. Gupta, S. et al., DOM based Content Extraction of HTML Documents, in the proceedings of the Twelfth World Wide Web conference (WWW 2003), Budapest, Hungary, May 2003.
9. S.S. Bhamare et al., An Efficient Method of Web Page Noise Cleaning for Effective Web Mining, International Journal of Computer Applications (0975 – 8887) Volume 146 – No.3, July 2016.
10. Song. Liu et al., Learning Block Importance Models for Web Pages, WWW 2004, May 17-22, 2004, New York, NY USA. ACM.
11. Bing Liu, Web Data Mining (Exploring Hyperlinks, Contents, and Usage Data), Springer.
12. Sivakumar P , Noise Free Information Retrieval Using Web Content Mining On Web Pages, Ph.D. Desertation-2013
13. The K-means Clustering technique documentation, [Online] Available: https://in.mathworks.com/help/stats/k-means-clustering-12.html
14. B.D. Davision, Recognizing Nepotistic links on the Web. Proceeding of AAAI 2000.
15. N. Kushmerick. Learning to remove Internet advertisements, AGENT-99, 1999.
16. S.H. Lin et al., Discovering informative content blocks from Web documents. In Proceeding of SIGKDD-2002, 2002
17. YI L. et al., Web Page Cleaning for Web Mining through Feature Weighting, Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI-03), 2003.

AUTHORS PROFILE

S. S. Bhamare, is Ph.D. in Computer Science from North Maharashtra University, Jalgaon, Maharashtra, India. He has completed his M.Sc. in Computer Science from NMU, Jalgaon. He is working as an Assistant Professor in School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University (formerly Known as North Maharashtra University), Jalgaon. His total teaching experience of 17 years and published more than 12 papers in reputed peer reviewed national and international journals & conferences. His research area includes Web Mining, Information Retrieval and IOT.

B. V. Pawar, is Ph.D. in Computer Science from North Maharashtra University, Jalgaon, Maharashtra, India. He has completed his M.Sc. in Computer Science from NMU, Jalgaon. He is working as a Professor and Head, Department of Information Technology, School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University (formerly Known as North Maharashtra University) , Jalgaon.. He has total teaching experience of more than 30 years and published more than 170 research papers in reputed peer reviewed national and international journals, books & conferences. He has successfully guided more than 12 Ph.D. students. His research area includes Natural Language Processing, Information Retrieval, Web Mining and DBMS. He has successfully carried out many research projects funded by various funding agencies like UGC, New Delhi, Rajiv Gandhi Science and Technology Commission (RGS&TC), Govt. of India/Maharashtra etc.