Predicting Outliers and Ranking Web Documents Using Correlation Analysis

Raheemaa Khan, Mohammed Saleem Irfan Ahmed, Ahmad Alenezi

Abstract: ***Web Content Mining plays a vital role in recent days as people rely on internet for each and every information. People download and upload the documents frequently in turn the data in the web grows tremendously. This loads the servers and consecutively there are large numbers of duplicate documents as a result retrieving the relevant document becomes more tedious. In this proposed work, an algorithm is developed, where the documents are preprocessed, term frequency is calculated, correlation coefficient is found, outliers are identified and the duplicate documents are eliminated and at last the ranking is done and given to the user. The proposed research work is Proportional Correlation coefficient algorithm. It is applied to eliminate the outliers and rank the web documents. To enhance the effectiveness of the search engine this processes is done for web content mining. Using this process the user obtains the efficient results. The experimental analysis provides better accuracy in detecting outliers by comparing the proposed algorithm along with the existing methods.***

Index Terms: Web Content Mining, Correlation Coefficient, Ranking, Outliers, Correlation

I. INTRODUCTION

These days internet plays a vital role in all walks of life. The World Wide Web has all the information stored in it irrespective of the subjects and the data in the server is piling up as everyone is uploading and downloading the information. As the information is heaped in the server, there are large numbers of redundant documents in the web. So accessing the right information from the search engine becomes very hectic. Here the data mining is essential to mine the required data.

Data mining is the extraction of unknown information from large data bases; it helps to focus on the most important data in the data warehouses [1]. When the information is mined from web, it is termed as Web mining. It is the process of mining the knowledge from the web with the help of data mining techniques and algorithms [2]. The web consists of text, structured data such as lists and tables, and even images, videos and audios. The Web mining is categorize into three types

Revised Manuscript Received on December 22, 2018.

Raheemaa Khan, Research Scholar, Department of Computer Science, Bharathiar University,Coimbatore Tamil Nadu, India rlrkhan@gmail.com, Orcid: 0000-0003-2846-9364

Mohammed Saleem Irfan Ahmed, Associate Professor, Department of Computer& Information Sciences, College of Science and Arts, Al Ula Branch,Taibah University, Al Ula, Madhina drmsirfan@gmail.com

Ahmad Alenezi,Assistant Professor, Department of Computer& Information Sciences, College of Science and Arts, AlUla Branch,Taibah, University, Al Ula, Madhina alenezi2015@gmail.com

- Web content mining is the process of extracting the information from the intent of Web pages and Web documents, in the description of text, images and audio or video files.
- Web structure mining is that the method of analyzing the nodes and association structure of an internet site with the assistance of graph theory.
- Web usage mining is that the method of mining patterns and knowledge from server logs

This paper puts limelight on the outlier mining on the web document content. Usually, outliers are the information or record that deviate such a lot or detached from different records which could are engendered employing in contrast to mechanism or the observation that are unreliable compared to the other different observations. [3]

II. RELATED WORK

Web mining is a blooming research area on solving issues while retrieving and streamlining the information on the net. Web Content Mining aims in extracting the knowledge from the web documents, web content, hyperlinks and server logs with the help of data mining techniques.

Ali et al[4] presents an overview of the major developments in the area of detection of Outliers in numerical datasets. It incorporates projection pursuit approaches as well as Mahalanobis distance-based procedures. They also explore principal component-based methods, which is relevant for high dimensional data. Page Content Ranking method is newly adopted method for Web Content Mining [6].

N-gram techniques are appropriate in web content outlier mining as it supports partial matching of strings for outlier detection [9], [10], [12]. But it is actually very slow for huge datasets because of the large number of n-gram vectors originated in the time of mining the web content outliers [12].The most widely used method for text retrieval process is TF.IDF Term Weighting Technique. More number of approaches has been proposed but to implement these techniques in web mining is very limited [13][14].

G.Poonkuzhali et al presented the mathematical approach for mining web content outliers on set theoretical and signed approach [15]-[16]. Also these authors developed an algorithm for eliminating redundant web content through correlation method. [17] [18]. A new algorithm called WCOND-Mine algorithm was developed for mining web

Published By:
Blue Eyes Intelligence *Engineering*
& Sciences Publication

content outliers and it uses vector space model for dissimilarity computation. This approach conjointly uses n-grams without assuming the existence of a domain dictionary [19], [20]. Another approach was suggested by merging generalized pattern mining algorithm and clustering concepts [21].

All these works on web content mining were adopted for search engines to mine the web content. In this proposed work, the Proportional correlation and Reflective Weighted Correlation has been applied to get rid of the duplicates and outliers from the retrieved web pages. It is more efficient than existing algorithms.

III. OUTLINE OF THE PAPER

Section 2 presents the related works. Section 3 presents outline of the paper. Section 4 presents architectural design of the proposed work Proportional Correlation. Section 5 presents the algorithm for ranking web documents. Section 6 presents experimental results. Section 7 presents performance evaluation. Finally section 8 presents the conclusion and future work.

IV. PROPORTIONAL CORRELATION FOR RELEVANCE RANKING IN WEB CONTENT

Fig.1 portrays the architecture of the proposed proportional correlation coefficient algorithm

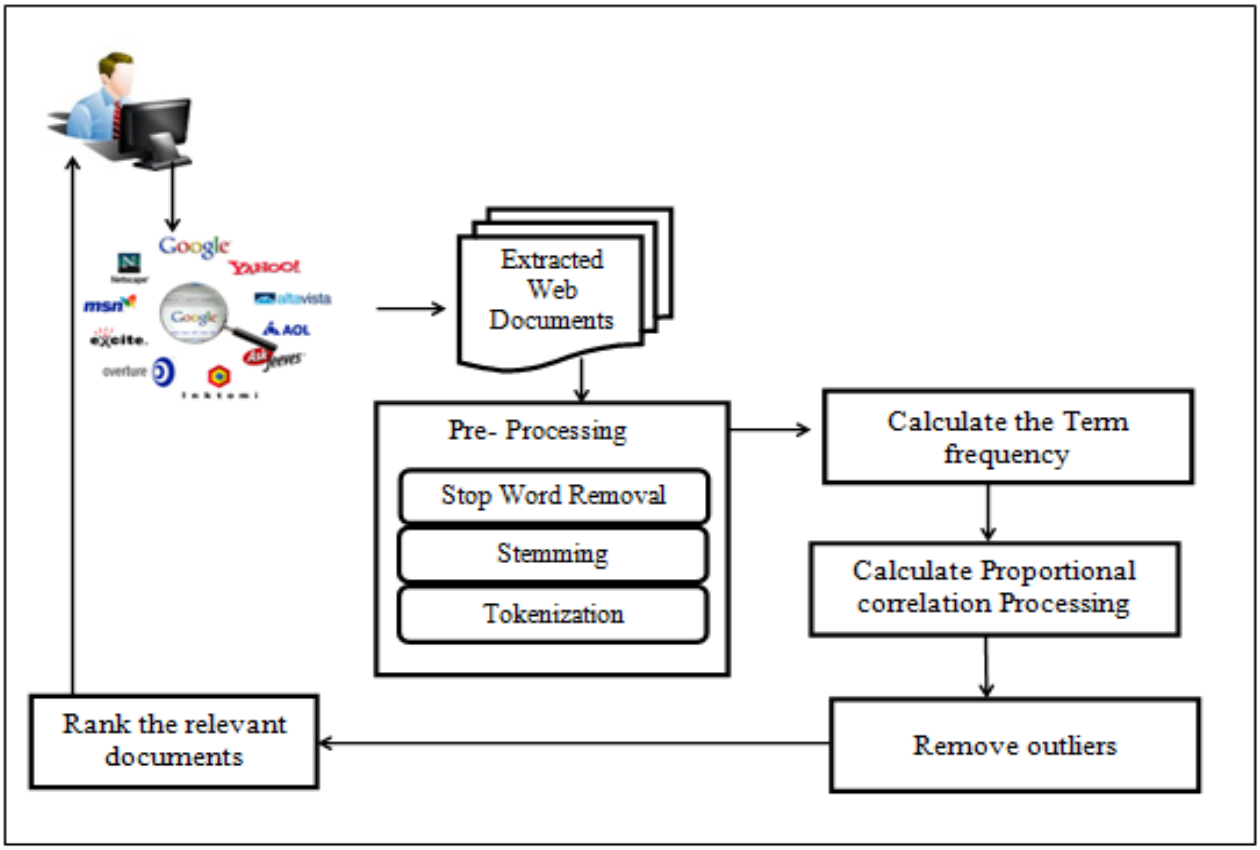


Figure 1: Architecture of the Proposed Proportional Correlation Coefficient Algorithm

The query is given in the search engine by the user. On the basis of the query, the documents where are retrieved from the web servers. Majority of the documents pulled out from the web servers may not be applicable to the user query. So those extracted documents are preprocessed. Preprocessing comprises of stop words removal, stemming, filtering and tokenization.

Stop words are the words which are often used and has

less meaning, are also less important than keywords (a, an, the, on, etc.,). Stop word removal filters stop words from the built-in stop word list for the record. Stemming is replacement of word suffixes to its root form (removed, removing, removal to the foundation word remove). Filtering is restricting of words with minimum and maximum length. Normally after stemming, the length of the word will become minimum as two or three. Those phrases may be removed as it is not significant. Hence, in the proposed work, the minimum length is set as 3 and the maximum length is set as 15. Tokenization is the process of breaking the textual content of a record into a sequence of words, terms, symbols, or different significant elements. For further processing those tokens are used.

Once the preprocessing is completed, the words in the document

Retrieval Number:F10970476S519/19©BEIESP

Published By:
Blue Eyes *Intelligence Engineering*
& Sciences Pssublication

are matched with the keywords. Those words are

taken for the term frequency calculation. Then the Proportional Correlation coefficient is found using Canberra metric method.

Proportional Correlation coefficient =

$$\sum \frac{(xi - yi)}{(xi + yi)} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

Then based on the total Proportional Correlation, the term having highest frequency is ranked accordingly. The last step is relevance calculation using the statistical method. Finally, a mined web document is obtained as per the end users desire.

V. PROPOSED ALGORITHM FOR PROPORTIONAL CORRELATION FOR RELEVANCE RANKING IN WEB CONTENT

Input: Web document.

Method: Statistical Method

Output: Extraction of unique web document.

1. Input the query Qr to Search Engine.
2. Extract the Documents D_i, where $1 \le i \le n$, n is the number of retrieved documents.
3. Pre-process the obtained documents by discarding the stop words, stemming, and tokenization.
4. List the preprocessed keywords K_n from the given query.
5. Calculate the total frequency of all the key terms $TF(K_n)$.
   a. For each document pair D_i and D_j, Calculate the term frequency of all the key terms in documents D_i and D_j and is denoted as $TF(K_n)_i$ and $TF(K_n)_j$.
   b. Compute the Weights for all the key terms $WTF(K_n)_i$ in the document pair D_i and D_j based on the $TF(K_n)$ where

$$WTF(K_n)_i = TF(K_n)_i / TF(K_n).$$

6. Finally prepare the correlation matrix using Proportional Correlation coefficient for each document pair using the formula.

Proportional Correlation coefficient =

$$\sum \frac{(xi - yi)}{(xi + yi)}$$

where $x_i = WTF(K_n)_i$ and $y_i = WTF(K_n)_j$

7. Analyze the calculated coefficient value of two documents Di and Dj. If it is 1, then Djis duplicate document and the document can be removed.
8. At the end, the Total Correlation Value is calculated by summing the coefficient of the document with all the other documents, based on which the ranks are assigned.

VI. EXPLANATION

The explanation of the above algorithm is as follows: the Term Frequency (TF) for the terms Database, Knowledge, Structure, Technique, Text from each document D1, D2, D3, D4,D5 and D6 are computed. The sample TF value is given in Table 1.

Table 1. Sample TF values in each document

Terms / Doc	D1	D2	D3	D4	D5	D6	TF(Ki)
Database	5	1	1	4	1	1	13
Knowledge	3	3	1	2	4	3	16
Structure	17	8	8	7	11	8	59
Technique	10	6	11	15	4	6	52
Text	8	5	8	23	1	5	50

Weights are assigned for each and every element of the table is shown in the Table 2.

Table 2. Sample Proportional Values of TF for each document

Terms / Doc	D1	D2	D3	D4	D5	D6
Database	0.38462	0.07692	0.07692	0.30769	0.07692	0.07692
Knowledge	0.1875	0.1875	0.0625	0.125	0.25	0.1875
Structure	0.28814	0.13559	0.13559	0.11864	0.18644	0.13559
Technique	0.19231	0.11538	0.21154	0.28846	0.07692	0.11538
Text	0.16	0.1	0.16	0.46	0.02	0.1

Now the Proportional correlation Coefficient is calculated and the redundant document is found using the formula Wij = $\sum \frac{(xi - yi)}{(xi + yi)}$

Table 3.ProportionalCorrelation coefficient for all document pairs

Terms / Doc	D1	D2	D3	D4	D5	D6
D1	0.00	1.50	1.47	0.04	1.94	1.50
D2	1.50	0.00	0.02	1.40	0.56	0.00
D3	1.47	0.02	0.00	1.50	0.48	0.02
D4	0.04	1.40	1.50	0.00	1.54	1.40
D5	1.94	0.56	0.48	0.54	0.00	0.56

Retrieval Number:F10970476S519/19©BEIESP

Published By:
Blue Eyes Intelligence *Engineering*
& Sciences Publication

D6	1.50	0.00	0.02	1.40	0.56	0.00

Since the Wij value of D2 and D6 is the identical, the D6 document is a duplicate document and therefore it may be eliminated. D1, D2, D3, D4 and D5 are ranked Proportionately for each term and is listed in the below Table 4.

Table 4.Ranking for relevant document

Variables	TCV	Rank
D1	6.4530	1
D2	3.4800	5
D3	3.4900	4
D4	5.8800	2
D5	4.0800	3

From the above table, D6 is the replica document is eliminated. And ranking is performed on the **Total Correlation Value**. Based on that, D1 is ranked first, then comes D4, then D5, then D3 and at final is D2.

VII. EXPERIMENTAL RESULT

In the experimental evaluation dataset is created with 100 web documents, in which 80 web documents are applicable and 20 web documents are duplicate or irrelevant as outliers. Primarily based on the proposed techniques the input files are preprocessed and the duplicate documents have been removed. The Proportional correlation coefficient and Reflective Weighted Correlation are computed for all of the documents. The replica files are carefully removed if the document having their coefficient cost as 1.

As a result of the proposed method, Precision, False rate, Accuracy of the methods in detecting outliers are analyzed. The test is conducted by various the number of documents which include relevant documents and outliers. The table 5 indicates the precision, recall of and accuracy of the proposed system.

Table 5. Calculation of Precision, False Rate and Accuracy

Trials	Doc Size	Relevant Doc	Actual Outlier	Outlier Deducted	Precision (%)	False Rate (%)	Accuracy (%)
1	10	8	2	1	50.0	50.0	90.00
2	20	16	4	3	75.0	25.0	95.00
3	30	24	6	4	66.7	33.3	93.33
4	40	32	8	6	75.0	25.0	95.00
5	50	40	10	8	80.0	20.0	96.00
6	60	48	12	9	75.0	25.0	95.00
7	70	56	14	10	71.4	28.6	94.29
8	80	64	16	11	68.8	31.3	93.75
9	90	72	18	13	72.2	27.8	94.44
10	100	80	20	15	75.0	25.0	95.00

From the above table, the experiments are preformed analyzed the Precision of the proposed method Proportional Correlation coefficient is high enough.

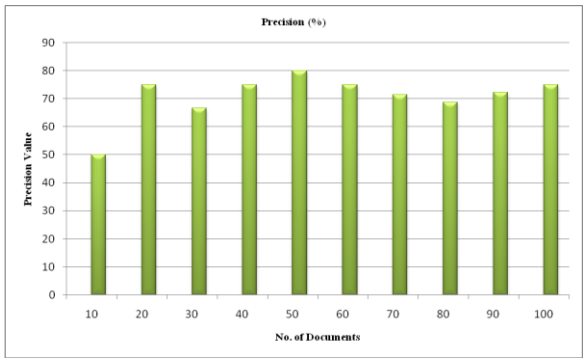


Figure 2: Precision at various count of documents

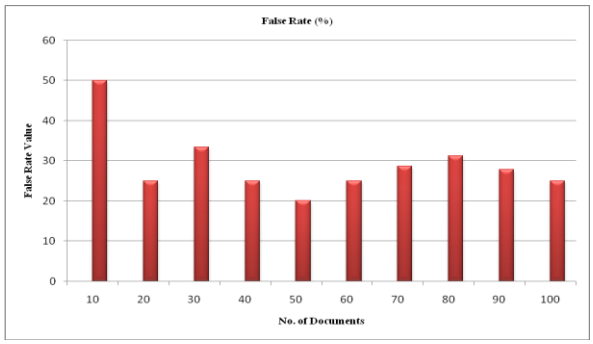


Figure 3: False rate at various count of documents

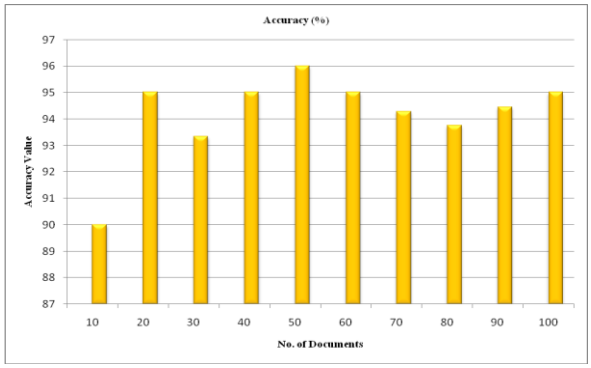


Figure 4: Accuracy at various count of documents

Figure 2, 3, and 4 depicts the results of the precision,false rate and accuracy comparison as a bar chart respectively.

From the above analysis, it states that the precision for the proposed method is high while compared to the prevailing strategies. Also, the proposed method reduces the false rate thereby increasing the efficiency in detecting and removing outliers when the data set is more the precision is more efficient.

Based on the experimental analysis the Precision and accuracy of the proposed system is high enough. Also the false rate is low and the accuracy is high for the proposed method from the other methods such as Ranked Correlation analysis.[3], [4], Proportional Approach and n-gram method [9], [10] Precision and False rate is calculated by varying the number of

Retrieval Number:F10970476S519/19©BEIESP

Published By:
Blue Eyes *Intelligence Engineering*
& Sciences Pssublication

documents having outliers and the result is compared with existing algorithms. The table 6 shows the precision and false rate at each trial for the proposed and existing methods.

Table 6. Comparison of Precision & False Rate with existing methods

Doc Size	PCA (Proposed)		KCA		RCA		WEIGHTED APPROACH		N-GRAM	
	Precision	False Rate	Precision	False Rate	Precision	False Rate	Precision	False Rate	Precision	False Rate
10	50	50	50	50	0	100	0	100	0	100
20	75	25	75	25	50	50	50	50	25	75
30	67	33	67	33	50	50	50	50	33	67
40	75	25	75	25	63	38	50	50	38	63
50	80	20	70	30	60	40	50	50	40	60
60	75	25	67	33	50	50	50	50	33	67
70	71	29	64	36	50	50	43	57	36	64
80	69	31	63	38	50	50	44	56	38	63
90	72	28	61	39	44	56	44	56	33	67
100	75	25	65	35	50	50	45	55	35	65

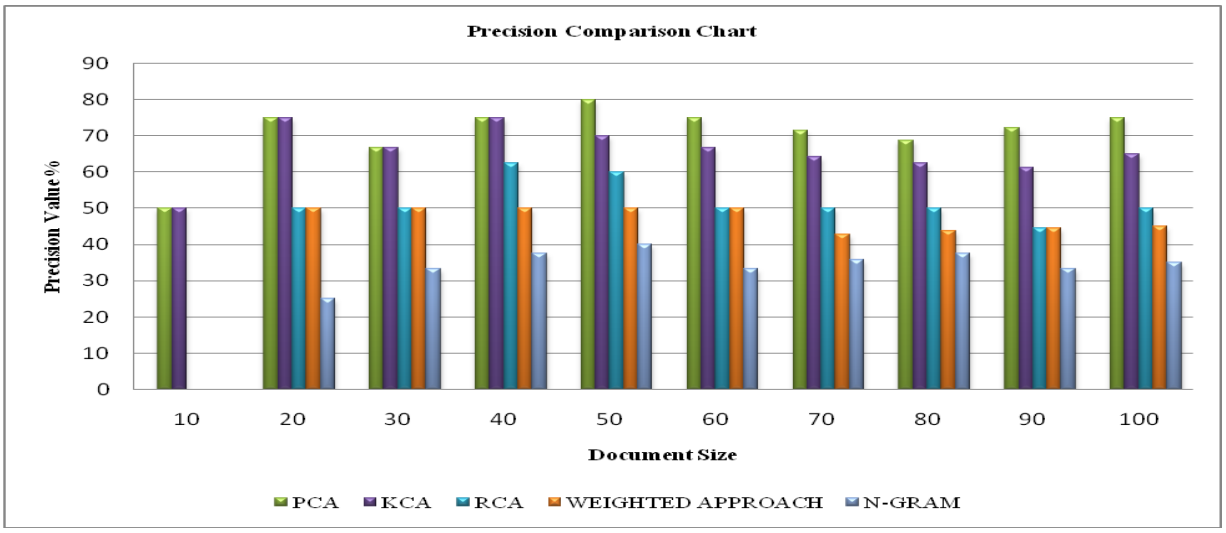
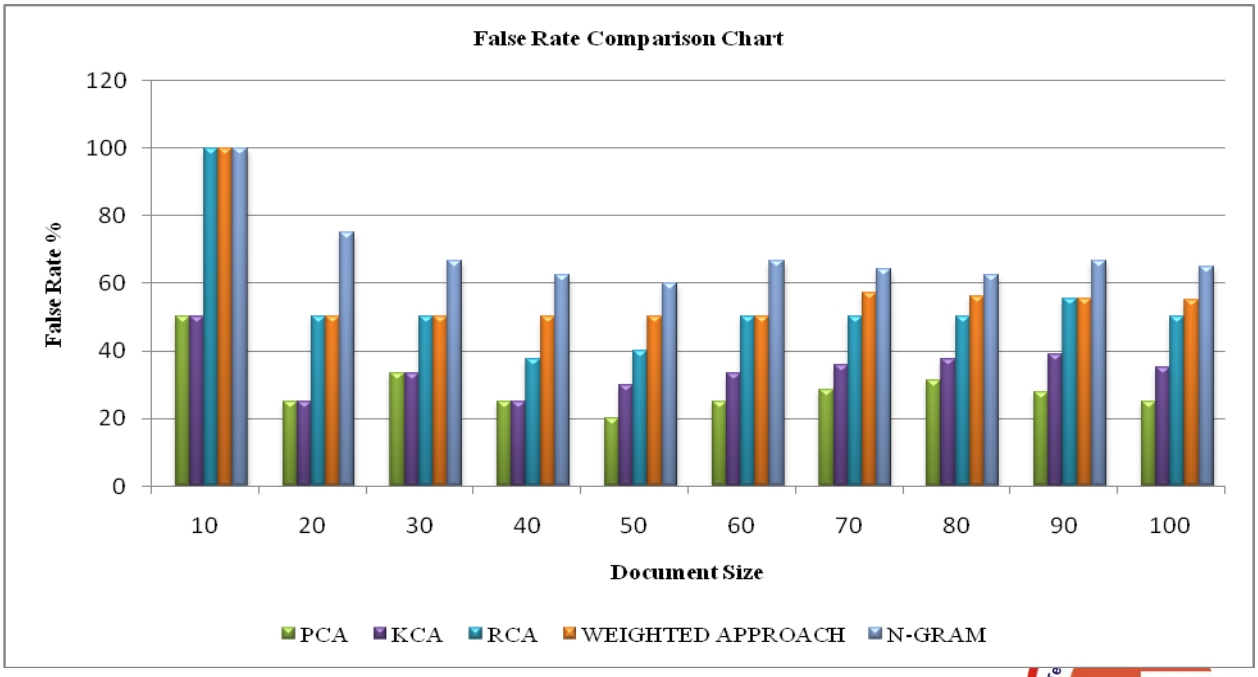


Figure 5: Precision Comparison Chart



Retrieval Number:F10970476S519/19©BEIESP

Published By:
Blue Eyes Intelligence *Engineering*
& Sciences Publication

Figure 6: False Rate Comparison Chart

From the Fig. 5 and 6, it is understood that the precision for the proposed method is high when measured with the existing methods such as Kendalls Correlation Analysis, Rank Correlation Analysis, Weighted Approach and N-Gram analysis. Apparently, the proposed method reduces the false rate consequently the efficiency in detecting and eliminating the outliers is enhanced.

VIII. CONCLUSION

The information in the World Wide Web are growing tremendously. People depend of web for each and every information, so it became necessary to mine the contents of the web as the data is huge. Mining the web content is developing research area in data mining. There are several tools are introduced to extract the relevant information without any duplication. This paper uses the analytical methods based on correlation analysis to find out the outlier documents from the web documents. The experimental results have proved that the proposed algorithm proved that the algorithm is efficient in finding outliers compared to the existing algorithms. The future work aims at enhancing the performance of the web content mining for the images and the hyperlinks.

REFERENCES

1. http://www.thearling.com/text/dmwhite/dmwhite.htm
2. https://www.techopedia.com/definition/15634/web-
3. mining
4. S.SathyaBama, Dr.M.S.Irfan Ahmed, A.Saravanan, "A Mathematical Approach For Improving ThePerformance Of The Search Engine Through Web Content Mining", Journal of Theoretical and Applied Information Technology 20th February 2014. Vol. 60 No.2
5. Mrs. R.L. Raheemaa Khan, Dr. M.S. Irfan Ahmed, , Mr. A. M. Riyad, "A Novel Analytical Approach for Identifying Outliers from Web Documents", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12156-12161
6. J. Pokorny, J. Smizansky, "Page Content Rank: An approach to the Web Content Mining", Proceedings of the IADIS International Conference on Applied Computing, Vol 2, 2005, pp. 22-25.
7. J. Pokorny, J. Smizansky, "Page Content Rank: An approach to the Web Content Mining", Proceedings of the IADIS International Conference on Applied Computing, Vol 2, 2005, pp. 22-25.
8. Hawkins D. "Identification of Outliers" Chapman and Hall, London
9. Breunig, M.M., Kriegel, H-P., Ng R.T., and Sander, J. "LOF: Identifying Outliers in Large Dataset" Proc. of ACM SIGMOD 2000, Dallas, TX 2000.
10. M. Agyemang, K. Barker, and R.S. Alhajj, "Framework for Mining Web Content Outliers," ACM Symposium on Applied Computing, pp.590-594, 2004
11. M. Agyemang, K. Barker, and R.S. Alhajj, "Mining web content outliers using structure oriented weighting techniques and n-grams,"Proceedings of ACM SAC, New Mexico, 2005.
12. M. Agyemang, K. Barker, and R.S. Alhajj, "WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents," Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC), 2005.
13. M. Agyemang, K. Barker, and R.S. Alhajj, "Hybrid approach to web content outlier mining without query vector. Springer –Berlin, Vol.3589, 2005
14. A. Khan, B. Baharudin and K. Khan, "Efficient feature selection and domain relevance term weighting method for Document Classification ,"Second International Conference on Computer Engineering and Applications IEEE, 2010.
15. C. Deisy, M. Gowri, S. Baskar, S.M.A. Kalaiarasi, and N. Ramraj, "Anovel term weighting scheme MIDF for Text Categorization," Journal of Engineering Science and Technology Vol. 5, No. 1 pp. 94 – 107,2010.
16. G Poonkuzhali, K Thiagarajan and K Sarukesi,Set theoretical Approach for mining web content through outliers detection International journal on research and industrial applications, Vol.2, 2009, pp. 131-138
17. G Poonkuzhali, K Thiagarajan, K Sarukesi andG V Uma, Signed approach for mining web content outliers. Proceedings of World Academy of Science, Engineering and Technology, Volume 56, pp - 820-824.
18. G. Poonkuzhali ,R. Kishore kumar, R. kripakeshav , P. Sudhakar and K. Sarukesi , Correlation Based Method to Detect and Remove Redundant Web Document, Advanced Materials Research, Vols. 171-172 ,2011, pp543-546
19. G Poonkuzhali , K Sarukesi and G V Uma,Detection and Removal of Redundant Web Document through Rectangular and Signed Approach, International Journal of Engineering, Science and Technology, Vol. 2 (9)-2010,pp4126-4132
20. M. Agyemang, K. Barker, A.S. Alhajj, "A comprehensive survey of numeric and symbolic outlier mining techniques", Intelligent DataAnalysis, Vol. 10, No (6), 2006, pp. 521-538.
21. M. Manikandan, "Improving efficiency of textual static web content mining using clustering techniques", Journal of Theoretical and Applied Information Technology, Vol. 33, No.2., 2011.
22. G. Poonkuzhali, K. Thiagarajan, K. Sarukesi,"Set theoretical approach for mining web content through outliers detection", International Journal on Research and Industrial Applications, Vol. 2, 2009, pp. 131-138.

Retrieval Number:F10970476S519/19©BEIESP

Published By:
Blue Eyes *Intelligence Engineering*
& Sciences Pssublication