International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12156-12161


# A Novel Analytical Approach for Identifying Outliers from Web Documents

**Mrs. R.L. Raheemaa Khan, Dr. M.S. Irfan Ahmed, Mr. A. M. Riyad**

*[1]Assistant Professor, Department of Master of Computer Applications*
*[2]Director, Department of Master of Computer Applications*
*[3]Assistant Professor & Head, Department of Computer Science*
*[1,2]Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.*
*[3]EMEA College of Arts & Science, Kondotty, Malappuram District., Kerala, India.*

*Orcid: [1]0000-0003-2846-9364*

**Abstract**

In this busy world, information in the web is growing tremendously as the volume of data is updated daily on web. And most of the people use the internet search engine to find and retrieve the information. Since the data is more, retrieving the relevant document becomes difficult task. So this problem has paved a way to develop an algorithm on web content mining. In this research, Kendall's Tau correlation analysis has been employed to determine the level liaison between the documents. The duplicate documents which are termed as outliers are identified if the correlation value between the documents is 1 and thus it can be eliminated. This method enforces the term frequency computation for common terms between the documents for which the ranking is done individually. From the experimental analysis, the proposed method provides better accuracy in detecting outliers in comparison with the existing methods.

**Keyword:** Correlation, Web Content Mining, Correlation Coefficient, Ranking, Outliers

## INTRODUCTION

Now-a-days, the growth of the World Wide Web beat up all the expectations. There are several billions of documents, HTML files, images, audio, video and other multimedia files available in internet and is still growing up for each and every second. Based on the impressive variety of the web content, retrieving interesting files has become a very difficult task.

Web Data Mining is a technique to crawl through various web resources to collect required and relevant information. According to the literatures, the web mining is broadly classified into Web content mining, Web structure mining and Web usage mining [1]. Web content mining, also known as text mining is the process of scanning and mining the text, pictures and graphs in Web pages to determine the significance of the content based on the user query. There are two groups of web content mining in which the one directly mines the content in the web documents and the other mines to improve the result of the search content from tools like search engine. Web usage mining basically collects the information about the identity of web users and browsing behavior at the web site and discovers the interesting usage sequence or information to better serve the needs of the web site. This information is often gathered automatically into access logs via the Web server. Web structure mining, is a taxonomy used to identify the relationship between Web pages linked by information or direct link connection. Based on web structural details, web structure mining is generally distributed into two types 1) extracting only the pages or documents from web and 2) examining the web page by constructing the tree.

This paper focuses on the outlier mining on the web document content. Commonly, outliers are the data or record that deviate so much or detached from other records which might have been engendered using a unlike mechanism or the observation that are unreliable when compared to the rest of other observations.

### Related Works

The web has several unique characteristics such as huge size, diverse in nature, constitutes different data types with heterogeneous information and significantly it is dynamic [2]. Due to these characteristics, the data mining techniques cannot be directly applied and so it becomes one of the important field in research. Kosala & Blockeel (2009) introduced the different categories of web mining and presented several issues on web content mining [2]. Agyemang et al. introduced a term web content outlier mining for the outliers that are found in web [3]. The taxonomy for web outliers along with the general framework for mining the web content outlier and few applications of web outlier mining is presented in [4]. Particularly the method employs full word matching for comparing the words in the document with that of domain dictionary. The Hybrid approach that pulls the power of n-gram technique and word based system without the existence of domain dictionary is introduced [5]

The n-gram method has been suggested that uses the existence of domain term dictionary along with the use of HTML structure of the web page for identifying the outliers in the web [6]. The author reduces the processing time by using the optimized algorithm that employs only the data caught in <Meta> and <Title> tags. The author also proposed an algorithm *WCOND-Mine* for mining web content outliers using n-grams however without the existence of term dictionary [7]. They employed the Vector space model for computing dissimilarity. The authors enhance the work by introducing the WCO algorithm for mining outlier in web which results in improved efficiency [8]. A traditional weighting technique TF .IDF that uses Term Frequency (TF) and Inverse Document Frequency (IDF) from Information Retrieval [9] which is commonly used in text mining has been used Zulkifeli et al. for identifying outliers in web documents [10].

An algorithm has been proposed based on clone detection and similarity measures to detect duplicate pages in the web sites and applications using HTML and ASP technology [11] for structured documents. An innovative approach having multilayer framework for detecting duplicated web pages using two similarity text paragraphs detection algorithms based on Edit distance and bootstrap method is recommended however it cannot find duplicates among multiple web pages [12].

Poonkuzhali et al. offered a few mathematical approach for finding outliers from the set of web documents. They employed set theory that uses intersection and union operations for removing web outliers [13] and signed approach that uses domain term dictionary for mining outliers in web content [14]. The authors introduced a method to remove the duplicate web documents through rectangular approach along with correlation analysis [15]. The authors also employed chi square test in their research for extracting required information which increases the quality of search engine result thereby increasing the efficiency [16]. However all the existing method uses domain term dictionary for identifying outliers., compiling the dictionary for each domain is a time consuming process also the methods do not provide weightage to keywords in the document.

An approach has been suggested for web content mining using Genetic Algorithm which extract relevant and required information from the web [17]. Schematic Outlier Elimination Scheme for detecting and removing has been proposed to remove unwanted redundant data into the server end [18]. However, the method works for the web server log files. The new method has been introduced without compiling the dictionary by using ranked correlation analysis that ranks the terms based on their number of occurrences in the document to mine related web content to eliminate redundancy [19].

Several Outlier detection methods are suggested to detect outliers [20], however many of the methods are applicable for numerical data. Thus for mining outliers in web content, the content data has to be transformed into numerical value for employing mathematical concepts. This paper introduces the use of Kendall's Tau correlation for detecting outliers especially duplicates in web documents.

**Outline of Paper**

The following section, Section 4 presents architectural design of the proposed system along with the explanation. Section 5 presents the algorithm and explanation for web content outlier mining. Section 6 explains the experimental results and the performance evaluation. Section 7 presents conclusions and future work.

**Architecture of the Proposed Work**

The Architecure of the proposed method has been depicted in the figure1. Basically it has five steps.

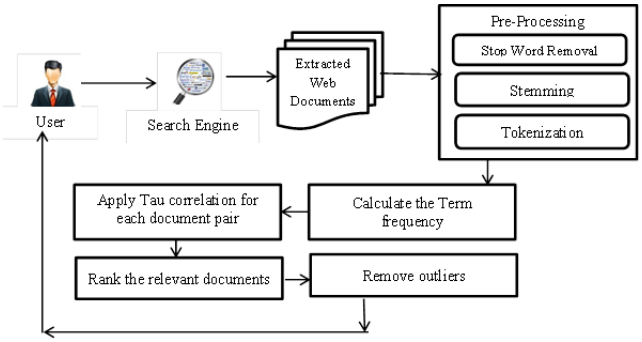


**Figure 1:** Architecture of the proposed method

The following are the steps to be performed for finding outliers.

1. User requests through search query
2. Pre-processing of extracted documents
3. Term frequency calculation
4. Comparison of term frequencies of common word between the documents
5. Relevance computation.

In the first step, the user request through the search query and is processed through the search engine. The results of the search engine are extracted. The documents from the result set might not be relevant to the user request. Therefore the set of documents are preprocessed**.** Pre-processing is an important step which reduces the size of the documents through stop word removal, Stemming, Filtering and Tokenization. Stop words are frequently used words that has less meaning and are less important than keywords (a, an, the, on, etc.,). Stop word

removal filters stop words from the document by removing every term from the built-in stop word list. Stemming is replacement of word suffixes to its root form (removed, removing, removal to the root word remove). Filtering is applied to limit the words with minimum and maximum length. Generally after stemming, the length of the word will become minimum as two or three. These words can be removed since it is not significant. In the proposed work, the minimum length is set as 3 and the maximum length is set as 15. Tokenization is the process of splitting the text of a document into a sequence of words, phrases, symbols, or other meaningful elements. For further processing these tokens are used.

The third step is the term frequency calculation. The words in the document are analyzed and the occurrence of each word is calculated. Each documents is compared with all the other documents. Thus in the next step, the set of common words for each document pair is extracted and terms are ranked based on the frequency values. The term having highest frequency will be ranked 1 and in the same way for other terms also.

The last step is relevance calculation using the statistical correlation method. In this proposed system, Kendall's Tau rank correlation coefficient is applied to find out the redundant document owing to its advantages. The distribution of Kendall's Tau has better statistical properties. Also it is very insensitive to errors due to which the correlation value will be accurate with smaller size. Since the correlation is applied only for the common terms between the documents, the Kendall's Tau correlation has been employed in this work. The rank of each term is compared in the document pair based on which the concordant and discordant pairs count is calculated. The number of larger ranks below a certain rank is concordant pair count and the number of smaller ranks below a certain rank is discordant pair count.

The Kendall τ coefficient is defined as:

$$\tau = \left| \frac{nc - nd}{n(n-1)/2} \right| \qquad (1)$$

Where nc is the number of concordant pairs and nd is the number of discordant pairs between the documents $D_i$ and $D_j$. n is the number of common terms in the documents $D_i$ and $D_j$. Only absolute correlation values are taken into account since correlation between the documents $D_i$ and $D_j$ is same as that of $D_j$ and $D_i$ and so the τ value always lies in the middle of 0 and 1. If the τ value for the documents $D_i$ and $D_j$ is 1, then $D_j$ is the duplicate of $D_i$. Also, if there is no common terms between the documents, then the documents are not related to each other and so the value of τ will be 0.

**Proposed Algorithm for Web Content Outlier Mining**

**Input:** Set of web documents.

**Method:** Kendall τ correlation coefficient

**Output:** Mining of unique web document after eliminating outliers.

**Step 1**: Input the user query Q to the Search Engine.

**Step 2**: Extract the Web Documents $D_i$ related to the given query where $1 \le i \le r$, r is the number of input documents.

**Step 3**: Pre-process the complete set of input documents by removing stop words, stemming, filtering and tokenization.

***For all the document pair $D_i$ and $D_j$ do the following steps.***

**Step 5:** Identify the common terms for the documents $D_i$ and $D_j$. Calculate the term frequency $TF(W_k)$ for all the common words $W_k$ where $1 \le k \le n$, n is the number of common words in documents $D_i$ and $D_j$ along with the list of keywords.

**Step 6:** Allocate the term frequency ranking $TFR(W_k)$ to each words $W_k$ in the document Di and $D_j$ individually where $1 \le k \le n$. n is the number of common words in document $D_i$ and $D_j$.

**Step 7:** Determine the number of concordant pairs (nc) and discordant pairs (nd) for each common terms.

**Step 8:** Compute Kendall's Tau correlation value based on the Equation (1) for each document pair. If the τ value is 1, $D_j$ is duplicate document, else $D_j$ is a unique document.

**Step 9:** Perform correlation for all the possible document pairs.

**Step 10:** Calculate the Total Correlation Value for each document by summing the calculated correlation value and assign the ranks accordingly.

**EXPLANATION**

The proposed method has been explained using an illustration. The sample data in table 1 shows the term frequencies for 6 documents denoted D1, D2, D3, D4, D5 and D6. Term Frequency (TF) for the terms Database, Knowledge, Structure, Technique and Mining from each document D1, D2, D3, D4, D5 and D6 is computed.

**Table 1:** Sample TF values in each document

| Terms / Doc | D1 | D2 | D3 | D4 | D5 | D6 |
|---|---|---|---|---|---|---|
| **Database** | 5 | 1 | 1 | 4 | 1 | 1 |
| **Knowledge** | 3 | 3 | 1 | 2 | 4 | 3 |
| **Structure** | 17 | 8 | 8 | 7 | 11 | 8 |
| **Technique** | 10 | 6 | 11 | 15 | 4 | 6 |
| **Mining** | 8 | 5 | 8 | 23 | 1 | 5 |

The terms are ranked based on their term frequency for all the documents individually. Then each document is compared

with all the other documents, for which the number of concordant pairs (nc) and the number of discordant pairs (nd) will be computed. Finally, the Kendall's Tau correlation given in equation (1) is applied for all the document pair and the correlation value is calculated. The τ value for all the document pairs based on the term frequency given in Table 1 is shown in Table 2.

**Table 2:** Correlation value (τ) for the document pairs

| Doc | D1 | D2 | D3 | D4 | D5 | D6 | TCV |
|---|---|---|---|---|---|---|---|
| **D1** | - | 0.800 | 0.671 | 0.400 | 0.447 | 0.800 | 3.118 |
| **D2** | 0.800 | - | 0.671 | 0.200 | 0.671 | **1.000** | **3.342** |
| **D3** | 0.671 | 0.671 | - | 0.671 | 0.250 | 0.671 | 2.934 |
| **D4** | 0.400 | 0.200 | 0.671 | - | 0.224 | 0.200 | 1.695 |
| **D5** | 0.447 | 0.671 | 0.250 | 0.224 | - | 0.671 | 2.263 |
| **D6** | 0.800 | **1.000** | 0.671 | 0.200 | 0.671 | - | **3.342** |

Since the τ value of D2 and D6 is 1, the document D6 is considered as a redundant document and therefore it can be eliminated. The documents are ranked based on their total correlation value calculated by summing its correlation value with all the other documents. The ranked documents based on their total correlation value is given in Table 3.

**Table 3:** Ranking for Relevant document

| Document | TCV | Rank |
|---|---|---|
| **D2** | 3.342 | 1 |
| **D1** | 3.118 | 2 |
| **D3** | 2.933 | 3 |
| **D5** | 1.815 | 4 |
| **D4** | 1.247 | 5 |

From table 3, D2 is ranked first since the Total correlation value (TCV) 3.342 is high. Then D1 is ranked next followed by D3, D5 and D4 based on their TCV values.

## EXPERIMENTAL RESULT

An experimental analysis have been made with the manually created dataset that consists of 100 web documents in which 80 web documents are relevant and 20 web documents are duplicate or irrelevant possibly outliers. These input documents are pre-processed and duplicate documents are identified based on the proposed method. The correlation value is computed for all the documents. The duplicate documents are carefully eliminated if the document having their coefficient value as 1.

The results of the proposed method is analyzed based on the various measures, such as Precision, False rate, Accuracy. The experiment is conducted by varying the number of documents including relevant documents and outliers. The experiment is analyzed by calculating the precision, recall and accuracy of the method in detecting outliers. The table 4 shows the precision, recall and accuracy of the proposed system for each trial.

**Table 4:** Calculation of Precision, False Rate & Accuracy

| Trials | No. of Documents | No. of Relevant Documents | No. of Outlier Documents | No. of Outlier Deducted | Precision Rate (%) | False Rate (%) | Accuracy Rate (%) |
|---|---|---|---|---|---|---|---|
| 1 | 10 | 8 | 2 | 1 | 50 | 50 | 90 |
| 2 | 20 | 16 | 4 | 3 | 75 | 25 | 95 |
| 3 | 30 | 24 | 6 | 4 | 66.67 | 33.33 | 93.33 |
| 4 | 40 | 32 | 8 | 6 | 75 | 25 | 95 |
| 5 | 50 | 40 | 10 | 7 | 70 | 30 | 94 |
| 6 | 60 | 48 | 12 | 8 | 66.67 | 33.33 | 93.33 |
| 7 | 70 | 56 | 14 | 9 | 64.29 | 35.71 | 92.86 |
| 8 | 80 | 64 | 16 | 10 | 62.50 | 37.5 | 92.5 |
| 9 | 90 | 72 | 18 | 11 | 61.11 | 38.89 | 92.22 |
| 10 | 100 | 80 | 20 | 13 | 65 | 35 | 93 |

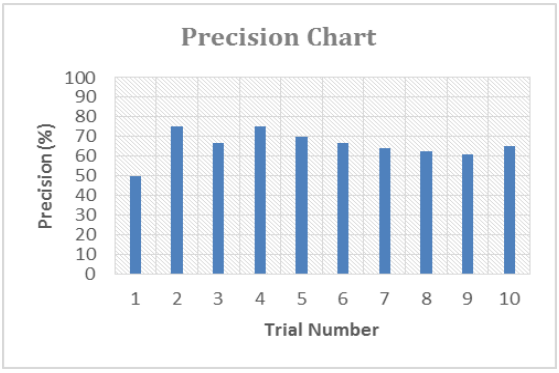


**Figure 2:** Precision values at various trials

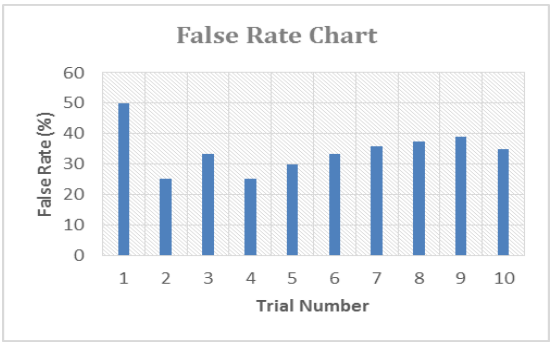


**Figure 3:** False Rate values at various trials

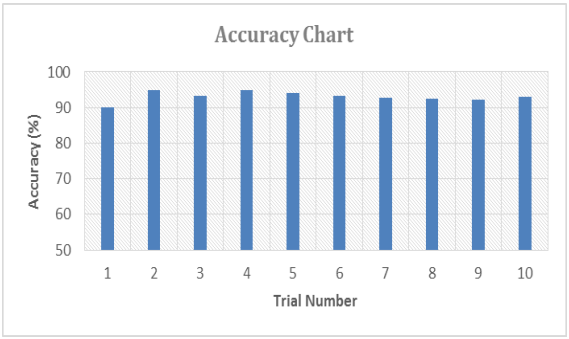


**Figure 4:** Accuracy at various trials

Based on the experimental analysis the Precision and accuracy of the proposed system is high enough. Also the false rate is low for the proposed method. The efficiency of the proposed method is compared with the other existing methods such as Ranked Correlation analysis, Weighted Approach and n-gram method. Precision and False rate is calculated by varying the number of documents having outliers and the result is compared with existing algorithms. The table 5 shows the precision and false rate at each trial for the proposed and existing methods.

**Table 5:** Comparison of Precision & False Rate with existing methods

| Doc Size | KCA (proposed) | | RCA | | Weighted Approach | | N-Gram | |
|---|---|---|---|---|---|---|---|---|
| | Precision | False Rate | Precision | False Rate | Precision | False Rate | Precision | False Rate |
| **10** | 50 | 50 | 0 | 100 | 0 | 100 | 0 | 100 |
| **20** | 75 | 25 | 50 | 50 | 50 | 50 | 25 | 75 |
| **30** | 66.67 | 33.33 | 50 | 50 | 50 | 50 | 33.33 | 66.67 |
| **40** | 75 | 25 | 62.5 | 37.5 | 50 | 50 | 37.5 | 62.5 |
| **50** | 70 | 30 | 60 | 40 | 50 | 50 | 40 | 60 |
| **60** | 66.67 | 33.33 | 50 | 50 | 50 | 50 | 33.33 | 66.67 |
| **70** | 64.29 | 35.71 | 50 | 50 | 42.86 | 57.14 | 35.71 | 64.29 |
| **80** | 62.5 | 37.5 | 50 | 50 | 43.75 | 56.25 | 37.5 | 62.5 |
| **90** | 61.11 | 38.89 | 44.44 | 55.56 | 44.44 | 55.56 | 33.33 | 66.67 |
| **100** | 65 | 35 | 50 | 50 | 45 | 55 | 35 | 65 |

The results of the precision and false rate comparison is depicted as a bar chart in Figure 5 and Figure 6 respectively.

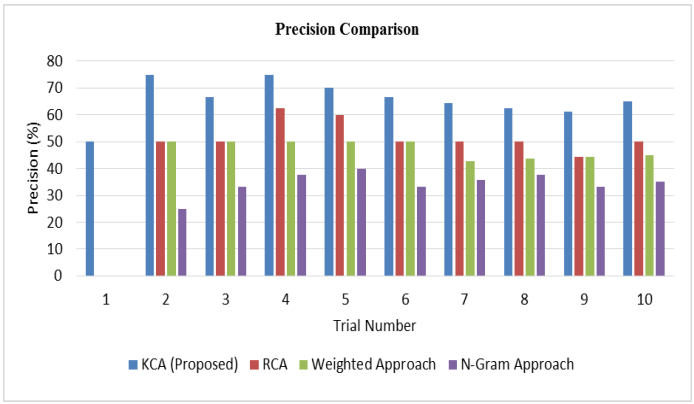


**Figure 5:** Precision Comparison Chart

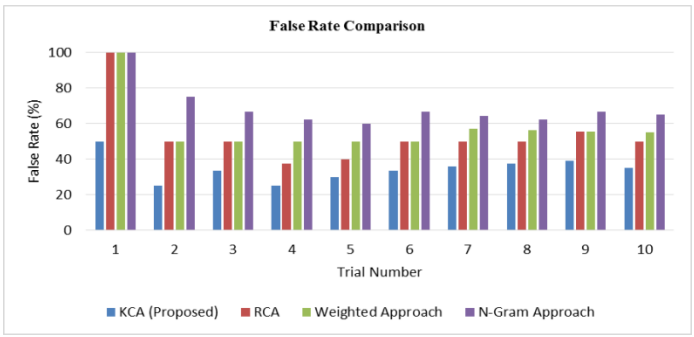


**Figure 6:** False Rate Comparison Chart

Thus from the analysis, it is clear that the precision for the proposed method is high when compared to the existing methods. Also, the proposed method reduces the false rate thereby increasing the efficiency in detecting and removing outliers.

## CONCLUSION

Web content mining is an emerging research area in the field of mining data as all the discipline rely on web. Due to the

incredible development of information found on the web, several retrieval tool have been introduced for the web extracting relevant information without duplicates**.** This paper introduces an analytical methodology based on correlation analysis to distinguish the irrelevant and duplicate outlier documents from the web documents and to remove them. Based on the experimental results, it is proved that the proposed algorithm is efficient in finding outliers when compared with the existing algorithms. The future work aims at performing experimental evaluation on the real dataset and also to employ other mathematical concepts in mining the web for the betterment of the outlier detection.

## REFERENCES

[1] Kosala R. and Blockeel H., 2000, "Web Mining Research: A Survey," ACM SIGKDD, vol. 2, no. 1, pp. 1-15.

[2] Liu B. and Chang K., 2004 , "Editorial: Special issue on Web Content Mining," SIGKDD Explorations, vol. 6, no. 2, pp. 1-4.

[3] Agyemang M., Barker K., and Alhajj R., 2006, "A Comprehensive Survey of Numeric and Symbolic Outlier Mining Techniques," Intelligent Data Analysis, vol. 10, no. 6, pp. 521-538.

[4] Agyemang M., Barker K., and Alhajj R., 2004, "Framework for Mining Web Content Outliers," in Proceedings of the 2004 ACM Symposium on Applied Computing, Cyprus, pp. 590-594.

[5] Agyemang M., Barker K., and Alhajj R., 2005, "Hybrid Approach to Web Content Outlier Mining without Query Vector," in Proceedings of 7th International Conference Data Warehousing and Knowledge Discovery, Denmark, pp. 285-294.

[6] Agyemang M., Barker K. and Alhajj R., 2005, "Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams," in Proceedings of ACM Symposium on Applied Computing, New Mexico, pp. 482-487.

[7] Agyemang M., Barker K., and Alhajj R., 2005, "WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents," IEEE Symposium on Computers and Communication, Spain.

[8] Agyemang M., Barker K., and Alhajj R., "Web Outlier Mining: Discovering Outliers from Web Datasets," Intelligent Data Analysis, vol. 9, no. 5, pp. 473-486, 2005.

[9] G. Salton, 1988, "Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer," Addison-Wesley Editors

[10] Zulkifeli WWR, Mustapha N, Mustapha A. 2012, "Classic term weighting technique for mining web content outliers". International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2012); Malaysia.

[11] Di Lucca GA, Massimiliano, Fasolina AR. 2002, "An approach to identify duplicated web pages". Proceedings of the 28th Annual International Computer Software and Applications Conference; IEEE computer Society press; p. 481.

[12] Han Z, Mo Q, Jianzhi L. 2009, "Effectively and efficiently detect web page duplication". IEEE Fourth International Conference on Digital Information Management (ICDIM); p. 1–6.

[13] 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

[14] G Poonkuzhali, K Thiagarajan, K Sarukesi and G V Uma, Signed approach for mining web content outliers. Proceedings of World Academy of Science, Engineering and Technology, Volume 56, pp -820-824.

[15] G. Poonkuzhali ,R. Kishore kumar, R. kripa keshav , P. Sudhakar and K. Sarukesi, Correlation Based Method to Detect and Remove Redundant Web Document, Advanced Materials Research, Vols. 171-172 ,2011, pp 543-546

[16] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, 2011, "Improving the quality of search results by eliminating web outliers using chisquare", published in Lecture notes in CCIS - Springer, Vol. 202, pp. 557-565.

[17] Johnson, F, and Santosh Kumar, 2013, "Web content mining using genetic algorithm." In Advances in Computing, Communication, and Control, pp. 82-93. Springer, Berlin, Heidelberg.

[18] Vasuki, S., and K. Subramanian, 2016 "An Innovative Outlier Detection Scheme to Identify the Web Page Usage Strategies." International Journal 4, no. 5.

[19] S. Sathya Bama, M. S. Ifran Ahmed and A. Saravanan, 2015, " A Mathematical Approach for Mining Web Content Outliers using Term Frequency Ranking", Indian Journal of Science and Technology, Vol 8 (14)

[20] Ali H., Imon A., and Werner M., 2009 "Detection of outliers Overview," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 1, no.1, pp. 57-70.