Identifying Outliers from Web Documents Using Reflective Weighted Correlation

Raheemaa Khan, Mohammed SaleemIrfan Ahmed, HusniHamadAlmistarihi

Abstract*: Due to the enormous use of the Web, the size of the web is getting increased rapidly at every second. Web mining is an important research field where the web documents are mined to extract useful knowledge related to web content and its usage. Web content mining is one of the categories of web mining, where the web pages and web documents are mined to eliminate web outliers. Generally, due to enormous usage of Internet, the contents in the web is becoming redundant as the same data is stored at several web servers by several users. Thus, accessing the relevant web pages is becoming very challenging task for the search engine. This paper focuses on web content mining wherein the set of web documents extracted by the search engine are examined to mine the interesting documents required for the user by removing the redundant and irrelevant documents. The proposed method employs a powerful mathematical concept called correlation analysis. In specific, reflective weighted correlation analysis has been used along with the term frequencies to identify the outliers thereby removing them improves the quality of results. Also, the score for the documents is computed and based on which the irrelevant documents are removed and the significant documents are extracted for the user. The method is highly useful for identifying and removing outliers. The proposed method is evaluated based on the experimental analysis and the results show that the proposed method has better accuracy of above 90% in predicting and differentiating the outliers from significant documents. The results are also compared with the other existing methods with accuracy and execution time as parameters.

***Index Terms*: Web Content Mining, Reflective Weighted Correlation Coefficient, Ranking, Outliers, Term Frequency.**

I. INTRODUCTION

Internet and Web are the two most powerful tool used by humans throughout the world. The web provides a large space for storing and retrieving the content from anywhere by anybody and at any time. The service provided by the web for humans is incredible. Though practically the web seems to be simple, technically the web is a more complicated system due to its unique characteristics. There are two main challenges exist in the web such as the size and diverse nature. The web is dynamic and the size increases rapidly at every second as several contents are uploaded by its users every day. Due to the increase in the usage of the Internet for each and every activity, the sizes of the web becomes unmanageable. As the web is useful for all the fields of work, not only the size increases, it also leads to content diversity. The web stores and accesses the information in the forms of structured, semi-structured and unstructured web pages. The web is mainly used for searching the information through a search engine. The user provides a search query to the search engine based on their interestingness. The web employs web crawler to fetch the information related to the search query from various web servers. As the users can download the documents and upload them in their web sites, several web servers may have the same content or document and are rigorously fetched by the crawlers. Thus, the search engine produces several irrelevant and duplicate documents which lead to decrease in the performance, as the results produced by the search engine may not be useful and even the irrelevant documents may distract the users' focus. Presenting the information that is significant for the web user is the most challenging task. Thus, mining the web has become the most important research field that helps in accessing and managing the web. The data mining techniques and the information retrieval concepts along with other fields such as machine learning and statistics are used in mining web due to its complex nature.

Web Mining is the process of discovering and extracting useful and interesting information from the World Wide Web. It has an ability to extract the various type of data stored in the web such as text, image, video, audio, multimedia and more [1]. Based on knowledge extraction, web mining is classified into three categories. They are web usage mining, web structure mining and web content mining [2]. Due to the massive amount of technological development, it is also possible to extract the users' interest which will be highly helpful in predicting the trends. The web usage mining aims at discovering the usage and browsing patterns from the web log files. This process is helpful in maintaining customer relationship, personalize and in managing the web sites [3]. Web mining is the process of analyzing the stricture of the web in terms of the graph structure. It provides the relationship between the web pages. It also analyses the link structure of the web pages using nodes and hyperlinks. Ultimately it constructs the summary of a website in terms of graph structure [4]. Web content mining is the process of extracting useful information from the web in terms of web pages and Web documents [5].

Outliers are the non-interesting records that deviate from other interesting records. Generally, the web content includes not only informative content but also some non-informative contents that disturbs the users' interest. In web pages, these non-informative

Revised Manuscript Received on June 05, 2019

Mrs. R. L. Raheemaa Khan, Department of Computer Science, Bharathiar University. Tamil Nadu, South India.

Dr. M. S. IrfanAhmed,Department of specialization in Trusted Networks from.Taibah University - Saudi Arabia

Dr. HusniHamadAlmistarihi, Department of specialization in Grid Computing and Distributed Systems, Taibah University - Saudi Arabia .

contents are named as noises and for search results having this less or non-informative contents are termed as web content outlier mining [6]. Several outlier algorithms are available for the numeric and structured datasets. However, the efficiency of the algorithm for this web contents are very poor [7].

This paper focuses on web content mining that eliminates the noises such as irrelevant content and duplicate contents and mines the interesting documents from the set of web documents extracted by the search engine. The proposed method uses reflective weighted correlation for finding the dependencies between the documents that are extracted. The performance of the system has been analyzed using various measures such as normalized discounted cumulative gain (NDCG), precision and accuracy. The proposed method shows a better accuracy rate of above 90% in extracting the interesting documents that are related to the search query.

The outline of the paper is as follows. Section 2 list outs some of the existing method in web content mining from literature. Section 3 introduces the architectural framework of the proposed reflective weighted correlation analysis along with the algorithm. Section 4analyzes the performance of the proposed method and results produced by the experiments and the results are compared with the existing methods. Section 5 concludes the paper with the accomplishment of the proposed method.

II. RELATED WORK

Web mining research focuses on introducing and developing new techniques for extracting useful knowledge from the web efficiently and effectively [8]. The authors presented several challenging characteristics of the web such as heterogeneous data types and information present in the web, dynamic nature of the web as the data can be uploaded at any time and so the size is not static and it changes at every second, redundancy of data and noise are also enormous. Web content mining is much more important due to a large number of contents getting increased day by day. Though there are several applications exist in mining the web contents, some of the notable applications are resource findings, information selection and pre-processing, generalization, analysis and visualization [9].

A set of procedures was proposed in the literature for mining the URL rules based on which the duplication can be detected without fetching the entire contents from the server. A machine learning techniques were employed to mine crawler logs and rule generalization. The performance of the method was explicitly analyzed and proved to be efficient [10]. Web content mining techniques were employed to transform the high categorical to low categorical information. The method takes advantage of statistics from mathematics and clustering from data mining [11]. An automatic learning method was suggested by King et al., (2007) to train an ontology with the entire knowledge in a three-level taxonomy. The ontologies are mined to identify the significant rules for classification based on which an extensive analysis can be made on the contents extracted from the search engine. Instead of a set of terms, the documents are represented as a set of subjects that lessens the synonymy [12]. Another method was introduced in detecting near duplicates. The method uses charikar's fingerprinting for detecting f-bit fingerprints that differatleast k-bit positions from other documents. The advantage of the method is that it is useful for online search queries and batch queries [13].

The benefit of the HTML tag structure of web pages and the n-gram method for limited matching of strings to implement n-gram based procedure for mining web content outliers was suggested by Agyemang et al. (2005a). To decrease the processing interval, the proposed method employs only data enclosed in <Meta> and <Title> tags [7]. The authors also proposed and documented the general model that supports the improvement of content-oriented procedures for mining web outliers [14]. The authors introduced WCOND-mine algorithm for detecting web content outliers based on n-grams without the existence of a domain dictionary [15]. A novel algorithm called HyCOQ was also suggested by them using a hybrid process that employs the power of n-gram oriented and word-oriented systems [16]. Poonkuzhali et al. (2009) focused on mining web content outliers by signifying a signed approach and full word matching with the organized domain dictionary [6]. The authors suggested several techniques such as set theory [17], weighted approach [18], statistical methods [19], correlation Analysis [20], signed with weighed techniques [21] for detecting outliers and to extract the core contents from the web documents.

A novel weight based pattern approach was introduced which is based on full word matching and their frequencies are used to mine the relevant documents and to eliminate the redundant and irrelevant documents [22]. A mathematical that employs the advantage of Spearman's rank correlation coefficient was suggested to eliminate the noises present in the input web documents sets. It uses the term frequencies to rank the documents which increase the effectiveness of the search engine [23]. The authors extended their work which uses an enhanced weighted approach that provides various weights to the keywords, content words and headings terms present in the documents [24].

All the above works identified from the literature provides good results however, the efficiency of removing the redundant and irrelevant documents are still can be increased. The execution time and accuracy of the systems can be increased still further to provide accurate results with a minimum time is the main objective of the research. Thus, the proposed framework employs the reflective weighted correlation in removing duplicates and irrelevant documents from the set of web documents.

III. PROPOSED WEB CONTENT OUTLIER MINING FRAMEWORK

The proposed framework of eliminating the irrelevant documents, duplicates and near duplicates employs the mathematical concept of correlation. The variation of Pearson correlation called reflective correlation coefficient with weights where the data values are not centered around their mean values is introduced in this framework. The proposed framework is depicted in Fig.1. The web user is the beneficiary for the proposed system, who provides the search query to the search engine. The search engine employs web crawler for accessing

the information related to the user requirements given in the form of a search query. The crawlers extract the information related to the given query from various web servers and the search engine displays the result to the user. As the results produced by the search engine are filthy, the proposed framework processes the given input documents to categorize the irrelevant and redundant documents from the core content and to eliminate them for improving the results produced by the search engine. The extracted documents are given as an input for the proposed framework where the documents and the query phraseundergo pre-processing.

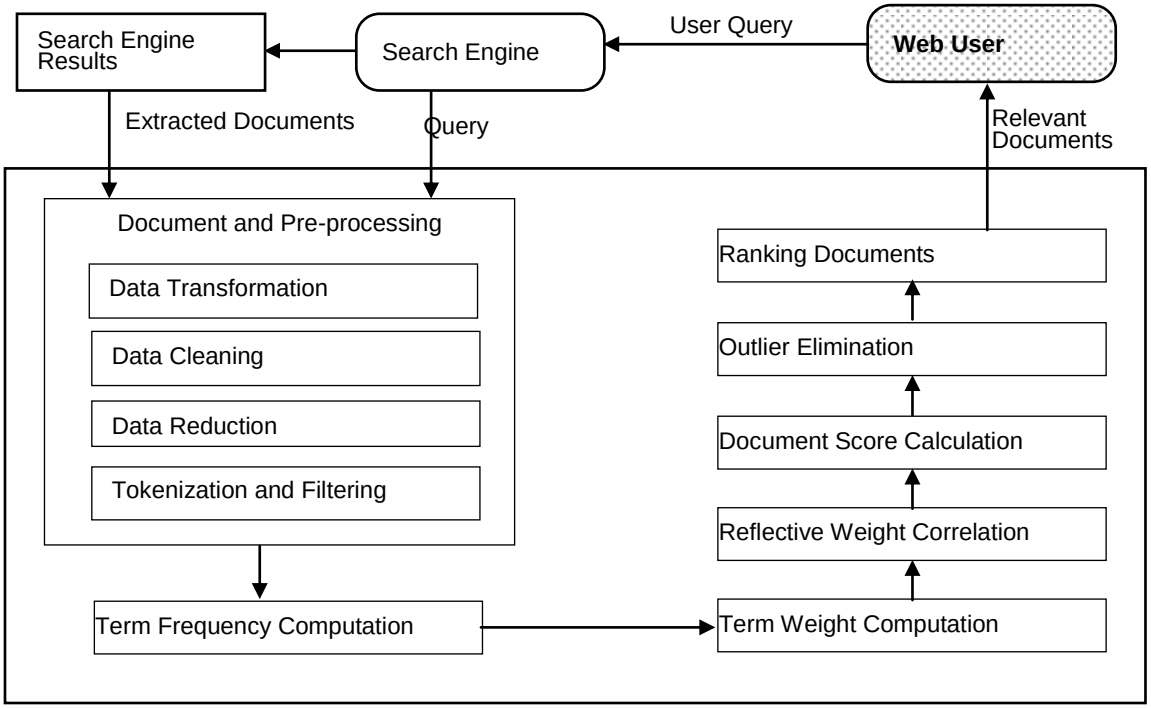

Figure 1. Proposed Web Content Outlier Mining Framework

The headings and the non-heading terms are processed separately. This pre-processing step is essential as it transforms the content to the common format and reduces the size by applying the data cleaning and data reduction processes. The first step in pre-processing is data transformation in which the words represented as terms are converted to the common format ie., the capital letters are converted to small case letters. This is essential for comparing the two terms. The second step in pre-processing is data cleaning in which the terms that constitute less meaning referred to as stop words are eliminated. Stop words are the words that are most commonly and frequently used in any languages with less meaning. The terms such as is, an, a, the, on, etc., are used in English have less meaning though they are used frequently.

The next step in this phase is data reduction. Data reduction employs stemming which is the process of reducing the terms to its root forms. For example, terms such as *expelling*, *expelled*, *expulsion* is reduced to its root word *expel*. This is the powerful step as it reduces the size of the words which is easy to process. There are several algorithms exist for the stemming process. The proposed framework makes use of porter stemming algorithm which is used universally. The next step is tokenization in which the set of words are broken into individual words which are highly essential in identifying the term weights and their frequencies. Finally, the pre-processing step ends with filtering. Using the filtering process, the words that are having a length less than 3 are eliminated and the words having a length greater than 15 are also eliminated.

Once the extracted documents and the query phrase are pre-processed, the next step is term frequency computation. The term frequency is one parameter that is used to identify the significance of the term in a document. Each document is processed and the occurrence of all the unique term is counted. The weight for the terms is assigned based on the appearance of the terms. If the word has appeared in the heading section, then the weight will be high than the normal terms. Each unique word in the document is compared with the key terms of the search query. If the term in the heading is a key term, then the weight is given as 1 else the weight of the term is assigned as 0.5. Each normal content term is compared with the key terms and if the match occurs, then the weight of the term is given as 0.75 else the weight of the term is assigned as 0.25.

The correlation between the documents for heading terms and the content terms are processed individually and the sum of the results provide the correlation between the documents. The reflective weighted correlation coefficient is computed between the documents by fetching the common heading terms and common terms in the normal terms list. The average of these two coefficients will give the final correlation value between the documents. The formula to compute the reflective weighted correlation coefficients (RWCC) is given in Eq. (1).

$$RWCC(x,y) = \frac{RWCC(h,x,y) + RWCC(n,x,y)}{2} \quad (1)$$

Where $RWCC(x,y)$ is the computed correlation between the document x and

document y,$RWCC(h,x,y)$ refers to the reflective weighted correlation coefficient for the heading terms between the documents x and y and $RWCC(n,x,y)$ refers to the reflective weighted correlation coefficient for the normal terms between the documents x and y. The formula to compute $RWCC(h,x,y)$ is given in Eq. (2).

$$RWCC(h,x,y) = \frac{\sum_{i=1}^{n} w_{h_i} x_{h_i} y_{h_i}}{\sqrt{\sum_{i=1}^{n}(w_{h_i} x_{h_i}^2)\sum_{i=1}^{n}(w_{h_i} y_{h_i}^2)}} \quad (2)$$

where $w_{h_i}$is the weight of the common heading term *i*, x_{h_i} is the frequency of the common heading term *i* in document x and $y_{h_i}$is the frequency of the common heading term *i* in document y. *n* is the number of common heading terms between the documents x and y. The formula to compute $RWCC(n,x,y)$ is given in Eq. (3).

$$RWCC(n,x,y) = \frac{\sum_{j=1}^{m} w_j x_j y_j}{\sqrt{\sum_{j=1}^{m}(w_j x_j^2)\sum_{j=1}^{m}(w_j y_j^2)}} \quad (3)$$

wherew_j is the weight of the common non-heading term *j*, $x_j$is the frequency of the common non-heading term *j* in document x and y_j is the frequency of the common non-heading term *j* in document y. *m* is the number of common normal terms between the documents x and y. The correlation is computed using the formula in Eq. (1). If the coefficient value is 1, then the two documents are considered as similar. Then one document can be deleted as it is a duplicate document. The score of a document can be computed by summing the correlation of a document with all the other documents. The documents are arranged as per the scores in descending order and the documents having the score less than the user threshold can be eliminated. The threshold can be computed by computing the ratio between the number of unique documents and the total number of input documents. The formula is given in Eq. (4).

$$Threshold\ Value = \frac{d}{N} \quad (4)$$

where, d is the number of unique documents after eliminating the duplicates and N is the total number of input documents. Thus, finally, the web documents are mined by eliminating the noises and extracting the significant documents.

A. Proposed Web Content Outlier Mining Algorithm

Input: Input web documents
Method: Statistical Method
Output: Extraction of the unique web documents.

Step 1: Input the user query to the search engine.

Step 2: Extract the documents D_i where *i*is the number of retrieved documents.

Step 3: The terms in the documents are grouped into two parts where one is the heading and the other is the normal content.

Step 4: The retrieved documents and query phrase are pre-processed by removing stop words, applying stemming, tokenization and filtering process.

Step 5: For each document, heading terms and the non-heading or normal terms are stored in a separate list.

Step 6: Compute the frequency of each heading terms and non-heading terms in the document

Step 7: For each document pair D_i and D_j, compare the heading term list and the normal term list of the documents and fetch the common terms from heading and non-heading lists separately.

Step 8: Compute the weights for the terms by comparing the key terms and based on the content type. If the terms in the heading list is a keyword then assign the weight as 1 else for all other non-keyword heading terms, assign the weight as 0.5. Similarly, if the content term is a key term then assign the weight as 0.75 else assign the term weight for non-heading terms as 0.25.

Step 9: Compute reflective weighted correlation coefficient for the heading terms as in Eq. (2)

Step 10: Compute the reflective weighted correlation coefficient for the non-heading terms as in Eq. (3)

Step 11: If the calculated coefficient value of two documents D_i and D_j is 1, then D_j is a redundant document and the document can be removed.

Step 12: Continue the steps from 6 to 10 for all the document pairs.

Step 13: The total correlation value or document score is calculated by summing the coefficient of the document with all the other documents.

Step 14: The irrelevant documents can be removed by fixing the threshold value as in Eq. (4) and removing the documents having a total score less than the threshold value provides the required result.

Step 15: Finally, the ranks are assigned based on the total correlation value and the extracted significant documents can be presented to the user.

IV. EXPERIMENTAL ANALYSIS

Experimental analysis has been made for the proposedweighted reflective correlation coefficient based web content outlier mining framework. The initial experiment is accomplished by extracting the set of documents for the given user query "Recent Research in Web Content Outlier Mining" against a Google search engine. The top 6 documents are extracted and the document are served as an input for the initial experiment and the details are listed in Table I. The documents undergone pre-processing phase in which the transformation of the text to the common format has been performed. The stop words are removed, the stemming procedure has been carried out, the tokenization has been performed and finally, the filter has been applied to eliminate the terms that are lesser than a length 3and greater than a length 15.

Table I. List of input documents

Doc ID	Document URL
D1	http://shodhganga.inflibnet.ac.in/bitstream/10603/26342/8/08_chapter%203.pdf
D2	http://www.yildiz.edu.tr/~aktas/courses/CE-0114890/g1-p3.pdf
D3	https://www.ijert.org/research/comparative-study-of-web-mining-algorithms-IJERTV3IS040350.pdf
D4	https://acadpubl.eu/jsi/2018-118-18/articles/18b/68.pdf
D5	http://staffwww.itn.liu.se/~aidvi/courses/06/dm/papers/Web%20Mining/WebMining.pdf
D6	https://www.worldwidejournals.com/paripex/recent_issues_pdf/2013/March/March_2013_1363610659_829fa_37.pdf

The term frequency is computed for all the terms in documents both for heading and the other non-heading terms. The key terms form the query is extracted. For computing the weights, the heading terms and non-heading terms are compared with the query key terms. If the common heading term is a key term, then the weight is assigned as 1 and for non-query terms, the weight is assigned as 0.5.

Similarly, for normal content terms, the query terms are assigned the weight as 0.75 and for non-query terms, the weights are assigned as 0.25. Finally, the reflective weighted correlation coefficient is computed using the formula given in Eq. (1), Eq. (2) and Eq. (3). The correlation value and the final score of the documents are given in Table II.

Table II. Correlation between the documents

	D1	D2	D3	D4	D5	D6	Score
D1	0.00	0.94	0.84	0.92	0.88	0.46	**4.03**
D2	0.94	0.00	0.83	0.93	0.80	0.52	**4.02**
D3	0.84	0.83	0.00	0.88	0.92	0.58	**4.04**
D4	0.92	0.93	0.81	0.00	0.91	0.59	**4.17**
D5	0.88	0.80	0.91	0.91	0.00	0.90	**4.39**
D6	0.46	0.52	0.58	0.59	0.90	0.00	**3.04**

The document score for D5 is having 4.39 and so it is ranked as 1 followed by D4 as rank 2, D3 as rank 3, D1 as rank 4, D2 as rank 5. As there are no duplicate documents in the input set and thus the threshold value will be 1. And so the threshold cannot be used. However, the document D6 having a very low score when compared with the other documents are considered as the irrelevant document.

The documents are ranked manually and the ranks are compared with the one computed using the proposed method. The performance is measured using various metrics for evaluating the ranked system such as precision at each position, average precision, discounted cumulative gain and normalized discounted cumulative gain [25]. The precision at each position is computed and the values are given in Table III.Thus from Table III, the average precision is 0.836, ie., 83.6%.

Table III. Precision at each position

Doc ID	Ranking		Precision at each position	Precision in %
	Proposed	Manual		
D1	4	5	0.6	60
D2	5	4	0.75	75
D3	3	3	1	100
D4	2	2	1	100
D5	1	1	1	100
D6	6	6	0.67	67

Another measure for calculating the quality of ranking and measuring the usefulness normally termed as a gain of a document at its retrieved position is discounted cumulative gain. The gain is gathered from the top position to the bottom position. The main idea behind this measure is that the relevant documents appearing at the lower positions must be penalized as the graded relevance (GR) value is reduced logarithmically proportional to the position of the result [26]. The formula to compute the DCG at position p is given in Eq. (5).

$$DCG_p = GR_1 + \sum_{i=2}^{p} \frac{GR_i}{log_2 i} \quad (5)$$

The DCG computation at each position is given in Table IV.

Table IV. DCG computation at each position

Doc ID	Rank (i)	Gain (GR_i)	$log_2 i$	$GR_i / log_2 i$	DCG_i
D5	1	4.39	-		4.39
D4	2	4.17	1.00	4.17	8.56
D3	3	4.04	1.58	2.56	11.12
D2	4	4.02	2.00	2.01	13.13
D1	5	4.03	2.32	1.74	14.86

To compute the normalized discounted cumulative gain, the formula is given in Eq. (6).

$$NDCG_p = \frac{DCG_p}{IDCG_p} \quad (6)$$

To normalize DCG values, ideal DCG must be computed. This is calculated by ordering the gain values in descending order and then computing the DCG value for the last position. The DCG value at position 5 is computed by ordering the gain as 4.39, 4.17, 4.04, 4.03 and 4.02. Thus, the DCG for this ideal ordering at positon 5 which is represented as $IDCG_5$= 14.892. The DCG_5 at position 5 is 14.864. Thus the normalized DCG is 0.9981.Thus the proposed method provides better performance in removing the redundant and irrelevant documents and in scoring as well as ranking the documents in a better way.

As there is no benchmark dataset for web content outlier mining, another analysis has been done by creating a dataset. 80 web documents have been extracted in which outliers are planted such that it contains 80 core content documents and 20 outlier documents for the particular key phrases. These documents are pre-processed by removing stop words, applying stemming, tokenization and filtering processes.

The frequency of the terms is computed. Before applying the proposed reflective weighted correlation coefficient, the common words are extracted and the weights are assigned for key terms and non-key terms based on the location of the term in the document (heading or normal terms). Then the RWCC is applied between each document pair and the correlation between the documents are computed. The resultant correlation value is 1 if the documents are similar. The lower value indicates that there is no dependency or

similarity between the documents. Finally, the scores for each document is computed by summing the correlation value of a document between all the other documents. For further analysis, the proposed method is analyzed by varying the number of input documents thereby the number of relevant documents and the number of outlier documents is also varied. Several performance metrics such as precision, false rate and accuracy of the proposed framework is analysed based on the outlier detection. The details about the trials along with the number of relevant and outlier documents at each trial are given in Table V.The details about the precision, false rate and accuracy of the proposed system at various trials are given in Table VI.

Table V. Trial Details

Trial ID	Total Input Documents	No. of Relevant Documents	No. of Outlier Documents
T1	10	8	2
T2	20	16	4
T3	30	24	6
T4	40	32	8
T5	50	40	10
T6	60	48	12
T7	70	56	14
T8	80	64	16
T9	90	72	18
T10	100	80	20

Table VI. Precision, False Rate and Accuracy

Trial ID	Outlier Deducted	Precision (%)	False Rate (%)	Accuracy (%)
T1	1	50.00	50.00	90.00
T2	3	75.00	25.00	95.00
T3	5	83.33	16.67	96.67
T4	6	75.00	25.00	95.00
T5	8	80.00	20.00	96.00
T6	10	83.33	16.67	96.67
T7	11	78.57	21.43	95.71
T8	12	75.00	25.00	95.00
T9	14	77.78	22.22	95.56
T10	16	80.00	20.00	96.00

The precision at various trials is thus depicted as a graph in Figure 2. Thus the precision value is above 75% for most of the trials.

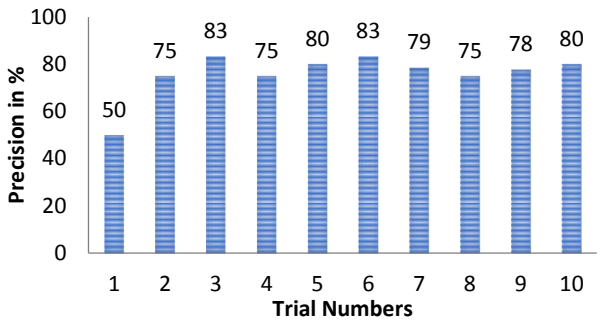

Figure 2. Precision at various trials

The false rate at various trials is depicted as a graph in Figure 3. Thus the false rate value of the proposed method is below 25 for all the trials.

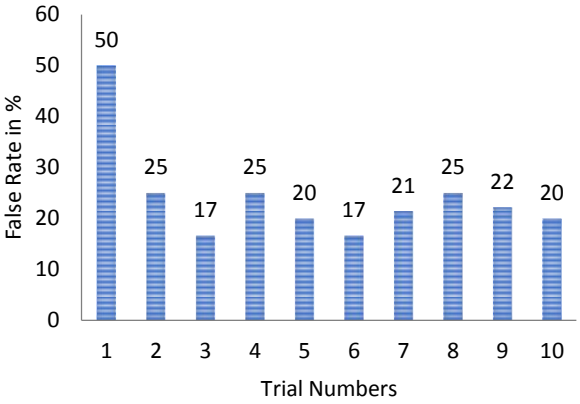

Figure 3.False rate at various trials

The accuracy of the proposed methods is measured at various trials by varying the number of input documents that includes both relevant and outlier documents. The accuracy values at various trials are depicted as a graph in Figure 4. Thus the accuracy value of the proposed method is above 90% for all the trials.

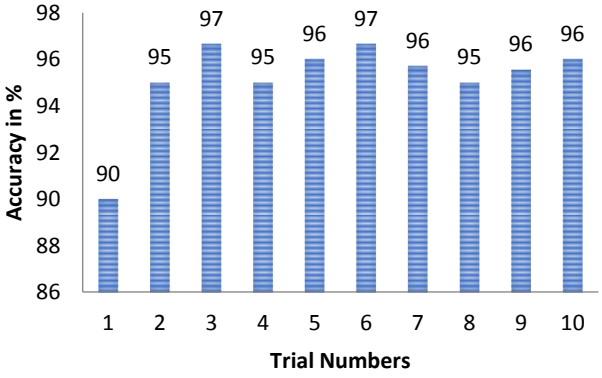

Figure 4. Accuracy at various trials

From the above analysis it is clear that the proposed method has better precision and accuracy for the given data. However, to prove that the method is efficient than any other method, the proposed method is compared with several existing methods such as N-Gram approach [7], Weighted Approach [18], WCA [24] and AWPF [22].

The experiment is performed with 100 documents with 80 relevant documents and 20 outlier documents. The true positive (TP), false positive (FP), false negative (FN) and true negative (TN) and the precision values of all the existing methods and the proposed method are shown in Table VII.The graph for the precision comparison is given in Figure 5.

Table VII. Precision Comparison

Algorithm	TP	FP	FN	TN	Precision
N-gram	62	18	9	11	77.5
Weighted Approach	65	15	6	14	81.25
WCA	69	11	5	15	86.25
AWPF	71	9	4	16	88.75
Proposed	72	8	3	17	90.00

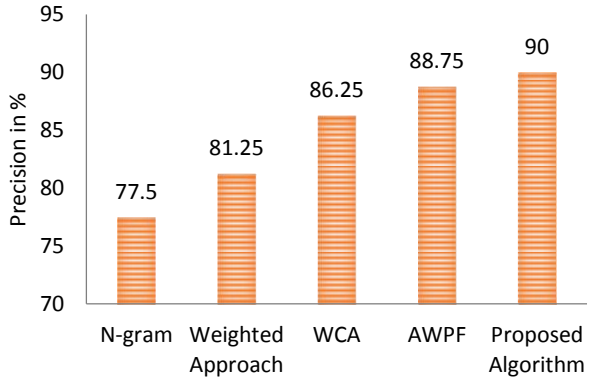

Figure 5. Precision Comparison

Similarly, the accuracy, false rate and execution time for the proposed and existing methods are given in Table VIII. The accuracy and false rate is compared and is represented as a graph in Figure 6.

Table VIII. Accuracy, False rate and Execution Time Comparison

Algorithm	Accuracy	False Rate	Execution Time
N-gram	73	27	38
Weighted Approach	79	21	25
WCA	84	16	26
AWPF	87	13	19
Proposed	89	11	18

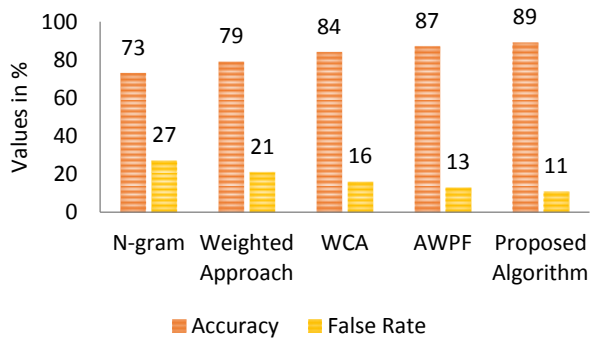

Figure 6. Accuracy and False Rate Comparison

From the above analysis the precision and accuracy of the proposed method are 90% and 89% which are higher than the existing methods. Also, the proposed method reduces the false rate with 11% which is less than any other existing methods compared in this study. The execution time of the proposed method is 18 seconds whereas for N-gram, weighted, WCA and AWPF methods are 38, 25, 26 and 19 seconds respectively. Thus the method is effective in accessing the relevant information.

V. CONCLUSION

Due to the evolving technological development, the web content becomes the most common research field. Though there are several techniques exists for extracting the relevant documents from the web, the results produced by the search engine is not still efficient. The paper introduces a novel statistical method termed reflective weighted correlation analysis for extracting the relevant documents from the web by eliminating the redundant and irrelevant documents. The term frequencies are computed for all the documents which converts the text to numbers that is suitable for applying correlation. The method computes the correlation between the document to identify the outlier documents and the scores are computed for the unique documents to extract the relevant documents. The experimental results show that the proposed method has an average of 90% for the analyses made which is higher than the other existing methods considered in the work. The future work aims at enhancing the performance of the web content mining for other types of data present in the web such as image, video and multimedia.

REFERENCES

1. K. Sharma, G. Shrivastava. And V. Kumar,. Web mining: Today and tomorrow. In 2011 3rd International Conference on Electronics Computer Technology 2011, April,Vol. 1, pp. 399-403. IEEE.
2. R. Kosala and H. Blockeel, Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 2000, pp.1-15.
3. A. Scime,. Web Mining: applications and techniques. IGI Global.ed., 2005
4. R.Kumar and A.K. Singh, . Web structure mining: Exploring hyperlinks and algorithms for information retrieval. American Journal of applied sciences, 7(6), 2010, pp.840-845.
5. F.Johnson and S.K Gupta, Web content mining techniques: a survey. International Journal of Computer Applications, 47(11).2012.
6. G Poonkuzhali, K Thiagarajan, K. Sarukesi, and G.V Uma, Signed approach for mining web content outliers. World Academy of Science, Engineering and Technology, 56(09),2009
7. M. Agyemang, K. Barker, R.S. Alhajj, "Mining web content outliers using structure oriented weighting techniques and n-grams", Proceedings of ACM SAC. New Mexico, 2005a.
8. B.Liu and K.Chen-Chuan-Chang,. Special issue on web content mining. Acm Sigkdd explorations newsletter, 6(2), 2004,pp.1-4.
9. V.Bharanipriya and V.K.Prasad,. Web content mining tools: a comparative study. International Journal of Information Technology and Knowledge Management, 4(1), 2011, pp.211-215.
10. A.Agarwal, H.S. Koppula, K.P. Leela, K.P.Chitrapura,S. Garg, P.K.GM, C. Haty, A. Roy and A.Sasturkar,. URL normalization for de-duplication of web pages. In Proceedings of the 18th ACM conference on Information and knowledge management 2009, November (pp. 1987-1990). ACM.
11. R.Anderson, L.S.Peranich, R..Dungca, J.P. Milana, X.Shao, P.C.Dulany, K.M. Hassibi, and J.C.Baker,. Detecting and measuring risk with predictive models using content mining. U.S. Patent 7,376,618.Fair Isaac Corp, 2008
12. J.D.King, , Y.Li, X. Tao, and R. Nayak, Mining world knowledge for analysis of search engine content. Web Intelligence and Agent Systems: An International Journal, 5(3), 2007,pp.233-253.
13. G.S Manku, A. Jain, and A.Das Sarma, . Detecting

near-duplicates for web crawling. In Proceedings of the 16th international conference on World Wide Web,2007, May (pp. 141-150). ACM.
14. M.Agyemang, K.Barker, &R.Alhajj, 'Web outlier mining: Discovering outliers from web datasets', Intelligent Data Analysis, 2005b,vol. 9, no. 5, ,pp. 473-486.
15. M. Agyemang, K. Barker, & R. Alhajj, 'WCOND-Mine: algorithm for detecting web content outliers from Web documents', Proceedings of Symposium on Computers and Communications, RS 2005c,pp. 885-890.
16. M. Agyemang, K. Barker, & R. Alhajj., 'Hybrid approach to web content outlier mining without query vector', Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg, 2005d,pp. 285-294.
17. G. Poonkuzhali, K.Thiagarajan, &K. Sarukesi, 'Set theoretical Approach for mining web content through Outliers detection', International journal on research and industrial applications, 2009,vol. 2, pp. 131-138.
18. G.Poonkuzhali, Web Content Outlier Mining through Mathematical Approach. Ph.D. Thesis, Anna University, Chennai. 2011a
19. G.Poonkuzhali, RK.Kumar, RK.Keshav, K.Thiagarajan, &K.Sarukesi, 'Effective Algorithms for Improving the Performance of Search Engine Results', International Journal of Applied Mathematics and Informatics, 2011b, vol. 5, no. 3, pp. 216-223.
20. G.Poonkuzhali, R.Kishore Kumar, R.Krip Keshav,P.Sudhakar, &K.Sarukesi, 'Correlation Based Method to Detect and Remove Redundant Web Document', Advanced Materials Research, 2011c,vol. 171, pp. 543-546.
21. G.Poonkuzhali, P.Sudhakar, &K.Sarukesi, 'Signed - with - Weight Technique for Mining Web Content Outliers', International Conference on Communication, Computing and Information Technology,2012, pp. 40-45.
22. S. Sathya Bama, M. S. Irfan Ahmed, A. Saravanan, 'Average Weight based Pattern frequency for Performing Outlier Mining in Web Documents', International Journal of Emerging Technology and Advanced Engineering, 2017, Volume 7, Issue 9, pp. 702-709,
23. S.S.Bama, M.I.Ahmed, and A.Saravanan, A Mathematical Approach for Mining Web Content Outliers using Term Frequency Ranking. Indian Journal of Science and Technology, 8(14), 2015.
24. S.S.Bama, M.I.Ahmed and A.Saravanan, Enhancing the Search Engine Results through Web Content Ranking. International Journal of Applied Engineering Research, 10(5), 2015, pp.13625-13635.
25. S.S.Bama, M.I.Ahmed and A.Saravanan, A survey on performance evaluation measures for information retrieval system. International Research Journal of Engineering and Technology, 2(2),2015, pp.1015-1020.
26. C. Burges, T.Shaked, E.Renshaw, A.Lazier, M. Deeds, N. Hamilton and G.N.Hullender, Learning to rank using gradient descent. In Proceedings of the 22nd International Conference on Machine learning (ICML-05) 2005 pp. 89-96.

AUTHORS PROFILE

Mrs. R. L. Raheemaa Khan, is a research scholar in the Department of Computer Science, Bharathiar University. She had done her M.Sc., MCA from BharathiarUniversity and M.Phil., in Computer Science from Madurai Kamaraj University. Her area of interest are data mining, web content mining. Now she is pursuing her research work in predicting outliers in web content. She has presented papers at conferences, published articles and papers in various journals. Also she has eight years of teaching experience in both UG and PG level.

Dr. M. S. IrfanAhmed, holds the doctoral degree from Alagappa University, Karaikudi, M.Phil from Bharathiar University, Coimbatore and MCA from Bharathidasan University, Trichy. He is currently working in Taibah University - Saudi Arabia and has 25 years of rich experience in areas of teaching, research and administration. His area of specialization in Trusted Networks. He has several research publications in well-known international Journals and conferences. He is also engaged to create linkage between industry and academia. He is associated with several Editorial Board of reputed Journals. He have chaired the sessions in the many National and International Conference. He has also won the Best Faculty Award in the year 2012.

Dr. HusniHamadAlmistarihi, holds the doctoral degree from UniversitiSains Malaysia (USM), Malaysia, Master degree in " Computer Information Systems" – Jordan ARABIC ACADEMY and B.Sc. in "Computer Information Systems" – Jordan, Amman University. He is currently working as Chairman of the board of directors in Taibah University - Saudi Arabia and has 24 years of rich experience in areas of teaching, research and administration. His area of specialization in Grid Computing and Distributed Systems. He has several research publications in well-known international Journals and conferences. He is also engaged in directing all the process, establishing academic and training standard, designing training courses curriculum and material, coordinating training and consulting with AIPS – USA. He has done many international certifications.