A Novel Algorithm for Enhancing Search Results by Detecting Dissimilar Patterns Based on Correlation Method

Poonkuzhali Sugumaran[1], Kishore Kumar Ravi[1], and Thirumurugan Shanmugam[2]
[1]Department of Information Technology, Rajalakshmi Engineering College, India
[2]Department of Information Technology, College of Applied Science-Sohar, Oman

Abstract: *The dynamic collection and voluminous growth of information on the web poses great challenges for retrieving relevant information. Though most of the researchers focused their research work in the areas of information retrieval and web mining, still their focus is only on retrieving similar patterns leaving dissimilar patterns which are likely to contain the outlying data. So this paper concentrates on mining web content outliers which extract the dissimilar web documents taken from the group of documents of same domain. Mining web content outliers indirectly help in promoting business activities and improving the quality of the search results. Existing algorithms for web content outliers mining are developed for structured documents, whereas, World Wide Web (WWW) contains mostly unstructured and semi structured documents. Therefore, there is need to develop a technique to mine outliers for unstructured and semi structured document types. In this research work, a novel statistical approach based on correlation method is developed for retrieving relevant web document through outlier detection technique. In addition, this method also identifies the redundant web documents. Removal of both redundant and outlaid documents improves the quality of search results catering to the user needs. Evaluation of the correlation method using Normalized Discounted Cumulative Gain method (NDCG) gives search results above 90%. The experimental results proved that this methodology gives better results in terms of accuracy, recall and specificity than the existing methodologies.*

Keywords: *Correlation, dissimilar patterns, outliers, redundant, relevance, term frequency, web content outliers.*

Received January 19, 2014; accepted May 21, 2014

1. Introduction

Due to the epidemic growth of information on the web, updating current data as well as retrieving relevant information becomes a tedious task. Existing web search engines employ conventional information retrieval and data mining techniques to discover interesting web content automatically. In addition, as most of the data in the web is semi structured and unstructured which contain a mix of text, video audio and image; there is a need to mine information catering to the specific needs of the user. The aforementioned problems result in the development of web content mining which uses the principles of data mining and knowledge discovery to screen more specific data.

Web content mining refers to the discovery of useful information from web contents, including text, image, audio, video, metadata and hyperlinks etc. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages. Two groups of web content mining are those that directly mine the content of documents and those that improve on the content search of other tools like search engine [13].

1.1. Major Issues in Web Content Mining

1.1.1. Extraction of Structured Data from the Web

Data extraction is the act or process of retrieving data out of data sources for further data processing or data storage. Extraction of structured data allows one to provide value added services. Using wrapper generation and wrapper induction techniques structured data can be easily extracted.

1.1.2. Extraction of Unstructured Text from the Web

Typically unstructured data sources include web pages, email, documents, PDF, scanned text, mainframe report, text documents, spool files and etc. Currently unstructured texts are extracted by the techniques which involve machine learning and natural language processing.

1.1.3. Web Information Integration

As the data on the web is heterogeneous, each web site may use different syntaxes to represent same information. Therefore, in order to provide value added services from multiple sites, web information integration has to be done by semantically integrating information from multiple sources.

1.1.4. Knowledge Synthesis

Knowledge synthesis is done to organize whole web to give the user a coherent picture of the domain topic.

1.1.5. Segmentation and Detection of Noisy Data

Web pages are segmented to extract the main content by removing noisy blocks such as advertisement, navigation links, chat rooms and copyright notices. The removal of noisy content leads to better results.

1.1.6. Mining Web Opinion Sources to Promote Business

Mining opinions through customer reviewers of products, forums, blogs and discussion groups help to promote business by analyzing marketing intelligence, competitors' strength and weakness, customers' interest in that product etc.

Most of the existing web content mining algorithms have concentrated on finding frequent patterns while neglecting the less frequent ones that are likely to contain the outlying data such as noise, irrelevant and redundant data. This research focuses on segmentation and detection of noise issue, which implies outliers mining, whose observations deviate so much from other observations to arouse suspicions that they might have been generated using a different mechanism. They may also reflect the true properties of data from rare and interesting events which may contain more valuable information than normal data. Outlier mining is dedicated to finding data objects which differ significantly from the rest of data that have been extensively studied in statistics and recently in data mining [10].

1.2. Traditional Outlier Detection Methods

1.2.1. Distribution Based Methods

Distribution based method is mostly conducted by statistics community by which outliers are determined through discordance test based on presumed model which requires the data set parameters, distribution parameters and the expected number of outliers in advance [10].

1.2.2. Depth Based Method

In depth based method, objects are organized in layers in the data space, with the expectation that shallow layers are more likely to contain outlying data objects than the deep layers [24].

1.2.3. Deviation Based Method

In deviation based method, outliers are identified by eliminating the main characteristics of objects in a group. Objects that deviate from this description are considered as outlier [9].

1.2.4. Distance Based Method

In distance based method, a rank is assigned to all points, using distance of point from k^{th} nearest neighbour and orders points by this rank. The top n points in ranked list are declared as outliers. Alternative approaches compute the outlier factor as sum of distances from k nearest neighbours [17].

1.2.5. Density Based Method

This method relies on Local Outlier Factor (LOF) of each point, which depends on local density of neighbourhood. Points with high factor are indicated as outliers [11, 28].

1.2.6. Clustering Based Method

In this method, the isolated points or vertices or documents that are unable to be clustered are identified as outliers [16].

1.2.7. Rule Based Method

Rule based method first learn rules that capture the normal behaviour of the system. A test instance that is not covered by any such rule is considered as an outlier [14].

1.2.8. Support Vector Based

Support Vector Novelty Detector (SVND) was recently developed. The SVND estimates a sphere to contain all the normal data patterns with the smallest radius; the outliers can be identified from the normal data patterns by calculating the distance of a data pattern to the centre of the sphere [26].

1.2.9. Neutral Network Based

The Replicator Neutral Network (RNN) is employed to detect outliers based on the observation that the trained neutral network will reconstruct some small number of individuals poorly, and these individuals can be considered as outliers. The outlier factor for ranking data is measured according to the magnitude of the reconstruction error [15].

Traditional outlier mining has received a tremendous attention on finding rare and exceptional patterns from numeric datasets. However, web outlier mining targeting web datasets has received very little attention in the mining community. Web outliers are web data that show significantly different characteristics than other web data taken from the same category. Web pages that have different contents from the category in which they were taken constitute web content outliers. Web content outliers mining concentrates on finding outliers such as noise, irrelevant and redundant pages from the web documents. Also, web content outliers mining can be used to determine pages with entirely different contents from their parent web sites. Web content

outliers mining play a crucial role in identifying competitors in online business, detecting criminals, frauds and threats in private and government bodies, network intrusion detection, stock market exchange and in improving the performance of the search engines.

Previous algorithms for web content outlier mining is focused only for structured documents, whereas, World Wide Web (WWW) contains mostly unstructured and semi structured documents. Therefore, there is need to develop a technique to mine outliers present in all types of documents with more precision. Moreover, the false positive rate of existing algorithms for mining web content outlier is more than 30%. The above mentioned issues, creates the need for developing a technique to mine web outliers from all types of documents including semi-structured and unstructured documents with less than 10% false positive rate.

Discounted Cumulative Gain (DCG) is a measure of effectiveness of a web search engine algorithm, often used in information retrieval. Using a graded relevance scale of documents in a search engine result set, *DCG* measures the usefulness, or gain, of a document based on its position in the result list. Search result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using *DCG* alone, so the cumulative gain at each position for a chosen *p*-value should be normalized across queries.

In this work, multiple correlation with the value ranging from 0 to 1 is used for detecting web content outliers. It works in a bivariate method, i.e., the correlation between two documents is found iteratively. Then, correlation coefficient of each document is summed to get the total correlation coefficient. Finally, the documents are ranked in descending order based on this total correlation coefficient to detect the top '*n*' relevant documents.

1.2.10. Outline of the Paper

Section 2 presents the related works done on this area. Section 3 depicts framework for mining web content outliers. Section 4 presents the experimental results and section 5 gives evaluation of correlation method. Section 6 delivers performance measure. Section 7 provides conclusion. Finally, section 8 presents future enhancement.

2. Related Works

An overview of the major developments in the area of detection of outliers in numerical datasets is presented [7]. These include projection pursuit approaches as well as Mahalanobis distance-based procedures. They also discuss principal component-based methods, which is applicable for high dimensional data. The major algorithms within each category are briefly discussed, together with current challenges and possible directions of future research in the area of detecting numerical outliers. A new method is introduced for finding outliers in a multidimensional dataset through density based approach which uses a LOF for each object in the dataset, indicating its degree of outlier-ness. The outlier factor is local in the sense that only a restricted neighbourhood of each object is taken into account. This approach is efficient for datasets where the nearest neighbour queries are supported by index structures and still practical for very large datasets [11]. Ramaswamy *et al.* [23] presented new definition for outliers and propose a novel formulation for distance-based outliers that is based on the distance of a point from its k^{th} nearest neighbour and developed a highly efficient partition-based algorithm for mining outliers. Ramaswamy *et al.* [23] devised algorithms based on distance and density based approaches for detecting outliers present only on numeric data sets.

The characteristics of web and research areas in web mining [19], different categories of web mining and issues on web content mining [18] are presented. The presence of outliers on the web and some practical applications and motivation behind web outlier mining is discussed [1, 2, 6]. They provide taxonomy for web outliers and continue with the description of the different types of outliers present on the web. In addition, a general framework for mining web content outliers using domain dictionary is also presented. An n-gram based algorithm is proposed using domain dictionary for mining web content outliers, which explores the advantages of n-gram techniques as well as HTML structure of web documents [4]. Agyemang *et al.* [5] developed a Web Content Outliers using N-grams without a Domain dictionary (WCOND)-Mine algorithm for mining web content outliers using n-grams without a domain dictionary. Here, weights are assigned to n-grams in documents based on which HTML tags enclosed their root words. Vector space model is used for dissimilarity computation. The experimental results show finding outliers with high order n-grams (5-grams) perform better than lower order n-grams. The Hybrid Approach to Web Content Outlier Mining without Query Vector (HYCOQ) algorithm [3] extracts the power of n-gram and word based systems. This algorithm mines web content outliers using hybrid data without a domain dictionary. The algorithm uses Information Retrieval (IR) techniques to extract useful features from web documents and then applies dissimilarity algorithms to determine outlying documents based on nearest dissimilarity density. The documents with high nearest dissimilarity densities are more likely to be outlying than those with low nearest dissimilarity densities. Even though HYCOQ algorithm produces more accurate results than WCOND Mine algorithm, the overall response time is more in HYCOQ. However, the entire

algorithm stated by above authors for mining web content outliers works only for structured documents. Also, n-gram computation leads to more processing time and memory usage.

A novel word semantic similarity measurement method based on web search engines is proposed, which exploits the information, including page count and snippets [12]. Xia *et al.* [27] presents a framework and algorithm for mining Chinese web text outliers based on improved Vector Space Model (VSM) and n-gram combined with domain knowledge. A new perspective using mathematical approach is developed based on set theoretical for mining web content outliers considering different web outliers namely irrelevant, redundant and inconsistent web content [21]. The same authors developed an algorithm based on signed approach for mining web content outliers using organized domain dictionary [22]. They have also applied statistical approach for retrieving relevant information from both structured and unstructured documents [20]. Still the precision of this approach is not more than 75%. In order to improve the precision and accuracy, the authors proposed a new statistical approach based on correlation for detecting web content outliers from both structured and unstructured documents. [25] evaluates a corpus based method to find similar terms in web sites using feature selection method. Saed and omar [8] developed software to improve the efficiency of multimedia search engines by eliminating repeated occurrences.

3. Framework of the Mining Web Content Outliers

The proposed algorithm explores the advantages of full word matching and correlation method using domain dictionary and its framework is shown in Figure 1. Initially the input web documents D_i and D_j (where i=1 and j=i+1) are taken and pre-processed. The pre-processing includes, stemming, stop words removal and tokenization. Stemming is the process of comparing the root forms of the searched terms to the documents in its database. Stop words elimination is the process of not considering certain words which will not affect the final result. Tokenization is defined as splitting of the words into small meaning full constituents. After pre-processing the full word profile for the document is generated and stored in hash table. Following the above process, term frequency for all the words is computed. Followed by that correlation co-efficient is computed between these two documents. If the correlation value is 1, then the above documents are exactly redundant, therefore, store D_j to the redundant set. Similarly correlation co-efficient is computed till i≤N. Then, total correlation co-efficient of document D_i is computed. The same process is carried out for all the extracted web documents. Finally, the total correlation co-efficient R_{ij} is ranked in ascending order. The top '*n*' documents are declared as an outlaid web document.

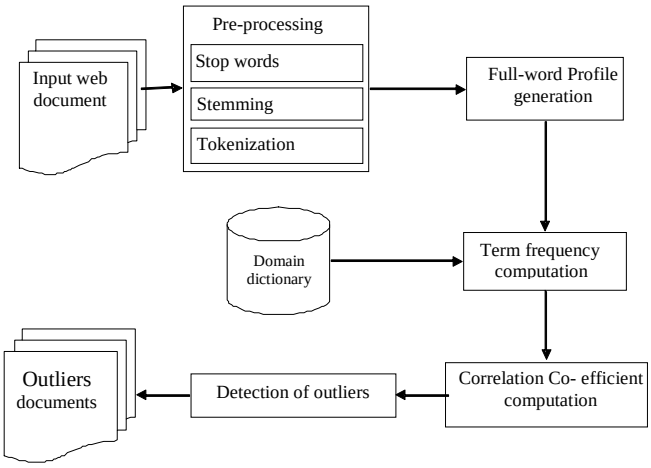

Figure 1. Architectural design of web content mining.

3.1. Pre-Processing

The pre-processing phase transforms the extracted data into a structure form that will be more easily and effectively processed for further steps. The pre-processing is the step that processes its input data to produce output which makes the rest of the process less complicated. Before performing pre-processing, except text, audio, video, image etc., are eliminated. Next, all digital numbers, punctuations like comma, full-stop, quotation mark, and special symbols are removed. The pre-processing step involves: Removal of stop words, stemming and tokenization.

3.1.1. Stop Words Removal

Stop words are words which contain least significance to be used in search queries. Usually stop words like a, an, the, is, was etc., are filtered out from search queries because they return vast amount of unnecessary information. Stop words can also be defined as "Words that do not appear in the index in a particular database because they are either insignificant or so common that the results would be higher than the system can handle". Stop words is controlled by human input and not automated.

3.1.2. Stemming

Stemming is a process of removing the common morphological and inflexional ending words to their root form. Search engines that use stemming compare the root forms of the search terms to the documents in its database. For example, if the user enters "producer" as the query, the search engine reduces the word to its root "produce" and returns all documents containing the words like produce, producer, producing, produced etc. The IR systems and search engines perform stemming for improving the performance and search speed. Thus, the key terms of a query or document are represented by stems rather

than by the original words. It also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents.

3.1.3. Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. A token is a string of characters, categorized according to the rules as a symbol. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful in the form of segmentation of text in linguistics as well as it forms part of lexical analysis in computer science.

3.2. Full Word Profile Generation

Full word profile is generated for the tokenized words and stored in a hash table, so that it can be compared easily with the words in the domain dictionary.

3.3. Term Frequency Calculation

The term frequency calculation is a weight often used in information retrieval and text mining. The term frequency calculation is taking the word count for the words present in the documents. In other words, it is finding out how many times the word has occurred in the document.

Algorithm 1: Correlation method for mining web content.

Outliers
 # Input: *Web document D*={D_1, D_2, D_3, …, D_n}
 # Output: *Outlier documents.*
D = input document set.
X_k = *Term frequency for the words in* D_i.
Y_k = *Term frequency for the words in* D_j.
R_{ij} = Correlation *coefficient.*
TCC=Total Correlation Coefficient.

Extract the input web document D_i *where* $1 \le i \le N$
Pre-process the entire extracted document.
Initialize redundant document set RD = {[ ]}.
for (i=1; $i \le n$; i++)
{
for (j=1; $j \le n$; j++)
{
if(i==j)
R_{ij}=0;
else
{
// Let m be the number of common words between D_i *and* D_j *that matches with domain dictionary.*
$X_k = TF(W_k)_i$;
$Y_k = TF(W_k)_j$; *where* $1 \le k \le m$

$$R_1 = \sum Y_K^2 - \frac{\sum Y_K^2}{|T|} \quad (1)$$

$$R_2 = \sum X_K^2 - \frac{\sum X_K^2}{|T|} \quad (2)$$

$$R_3 = \sum X_K Y_K - \frac{\sum X_K Y_K}{|T|} \quad (3)$$

$$R_{ij} = \frac{R_3}{\sqrt{R_1} * \sqrt{R_2}} \quad (4)$$

}
if (R_{ij}==1)
{
Display D_i *and* D_j *are redundant;*
$RD = RD' \bigcup D_j$; *where* $1 < i \le n$.
}
else
D_i *and* D_j *are not redundant;*
}
$TCC = \sum R_{ij}$ *where* j=1 *to* n.
}
Sort TCC in ascending order
Remove redundant data set {*RD*} *from* {*D*}
Display the top 'n*' outliers.*

3.4. Detection of Outliers

Dissimilarity measure is computed for determining the differences among documents within the same category. In this method, the dissimilarity measure is computed based on the total correlation coefficient of each document. Based on this dissimilarity measure, all the documents are sorted in ascending order. The document at the top will have less similarity measure which deviates more from the category of interest and these documents are referred as outlier documents. Similarly, the documents at the bottom will have a greater similarity measure which is more relevant to the category of interest. The top '*n*' documents that have less similarity measure are declared as outliers based on cut-off score or threshold value as specified by user. Then, the redundant documents are removed from the original document set. Finally, the top most '*n*' documents are displayed as outliers.

4. Experimental Results

This section focuses on test and experimental results of the mathematical algorithm based on correlation for mining web content outliers. Two experiments were conducted against two different datasets. These datasets were constructed by retrieving documents from the Google search engine. The first sets of experiments are conducted with distinct data samples in which outlier documents are completely unrelated. The second sets of experiments were conducted with data samples consisting of outliers taken from similar domain.

The first set of experiment is conducted with 5000 documents extracted from search engine (Google) in which 500 random documents are considered for constructing domain dictionary. From the document set, 4500 documents falls under computing domain and 500 documents falls under medical domain. Five

different trials were conducted with document set ranging from 100 to 5000. The results of the first experiment are projected in Table 1.

Table 1. Detection of outliers from different domain.

No. of Documents	Actual Outlier	Outlier Detected Through Correlation Algorithm	Outlier Detection Rate %	No of Documents
100	10	9	90	100
500	50	45	90	500
1000	100	92	92	1000
2000	200	186	93	2000
5000	500	475	95	5000

The second set of experiment is conducted with document set which slightly deviates from other documents. This experiment is conducted to measure the accuracy of detecting outliers within the same domain to prove the deficiency in which the trials range from 10 to 200. Results of the second dataset are projected in Table 2.

Table 2. Detection of outliers from same domain.

No. of Documents	Actual Outlier	Outlier Detected Through Correlation Algorithm	Outlier Detection Rate %
10	2	2	100
20	4	4	100
50	10	8	80
100	20	17	85
200	40	34	85

Initially pre-processing is done for all the documents taken in input datasets. Followed by that the term frequency for the common words between document D_i and D_j (j=i+1) is computed. Then, the correlation coefficient is computed for these documents. If the correlation coefficient value is equal to 1, then D_j is stored in Redundant Document (RD) set. The same process is repeated for the remaining documents in the input dataset. Then the total correlation coefficient is computed for each document. Finally, the documents are stored in ascending order based on the total correlation coefficient to detect top '*n*' outliers. The lowest value of total correlation coefficient indicates the least relativity (dissimilarity) of that document set whereas the maximum correlation coefficient value indicates the more relativity of the document set.

The outlier detection rate of different trials obtained from Table 1 indicates the accuracy escalating from 90 to 95 which is represented in Figure 2.

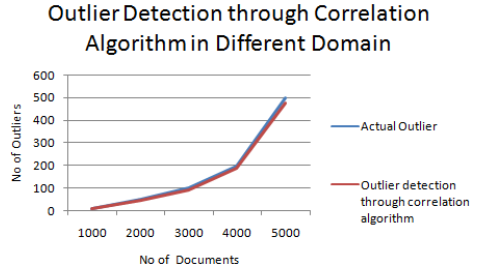

Figure 2. No. of outliers detected through correlation in different domain.

The outlier detection rate of different trials obtained from Table 2 against similar documents set indicates that the accuracy ranging from 80 to 85 which is represented in Figure 3.

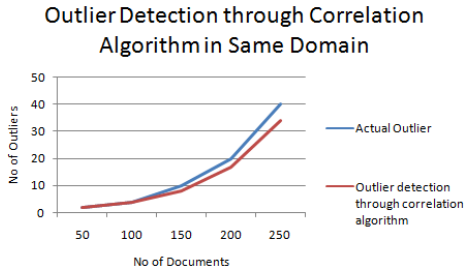

Figure 3. No. of outliers detected through correlation in same domain.

In addition, correlation method retrieves unique documents by eliminating redundant documents.

4.1. Comparative Study With Existing Method

This section presents the comparative results of correlation method with existing method based on n-gram technique [4]. The experiment is conducted with 225 resume pages and 25 recruiter pages retrieved from the dmoz open directory project (www.dmoz.org). The input dataset used by Agyemang *et al.* [4] for detecting web content outliers was also selected from www.dmoz.org. The domain dictionary is constructed using 50 resume pages taken randomly. The remaining 175 resume pages and 25 recruiter pages are taken as test data. The top '*n*' outliers detected by the correlation method and existing method based on n-gram is shown in Table 3 and in Figure 4.

Table 3. Comparative results of outlier detection through n-gram based algorithm and correlation method.

No. of Documents	Top 'n' Outliers	N-Gram Based Algorithm	Correlation Method
200	5	4	4
	10	7	8
	15	10	12
	20	12	17
	25	14	22

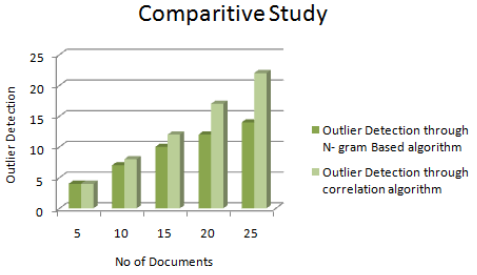

Figure 4. Comparative study with existing method.

5. Evaluation of Correlation Method

The *DCG* is defined as the sum of the products of an Outlier Score (OS) and its position weight which is logarithmically proportional to the position of the returned documents. *DCG* can be used to evaluate the search results by the ranking algorithms. *DCG* method

ranks the quality of returned documents by finding the cumulative relevance gain of all the outlier documents returned by ranking algorithms. The positional parameter *C* represents the ranking score assigned by the correlation method. The positional parameter *I* represent the ideal ranking score assigned by the human experts.

The *DCG* accumulated at a particular rank position *C* through correlation ranking is defined as:

$$DCG_C = \sum_{C=1}^{n} \left(\frac{os_c}{log_2(C+1)} \right) \quad (5)$$

The *DCG* accumulated at a particular rank position *I* through Ideal ranking is defined as:

$$DCG_I = \sum_{I=1}^{n} \left(\frac{os_I}{log_2(I+1)} \right) \quad (6)$$

The Normalized DCG (NDCG) is defined as searched result lists which vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using *DCG* alone, so the cumulative gain at each position for a chosen value of '*p*' should be normalized across queries through ideal ranking. It is also defined as ratio of *DCG* of the ranked documents derived from the correlation method to the ideal ranking given by human experts. *NDCG* is computed as:

$$NDCG = \frac{DCG_C}{DCG_I} \quad (7)$$

The cumulative gain is normalized for the ideal result list and its value vary from 0 to 1. The results obtained through correlation method is listed in Table 2 are validated by *NDCG* method using the above formula and its outcome is projected in Figure 5. Here the ranking of each document was judged by the human experts (called Ideal ranking) and an outlier score was assigned to each document by 5 points graded scale, where 5 indicated the most outlier document and 1 indicated the least outlier document. All the documents are sorted by the outlier scores.

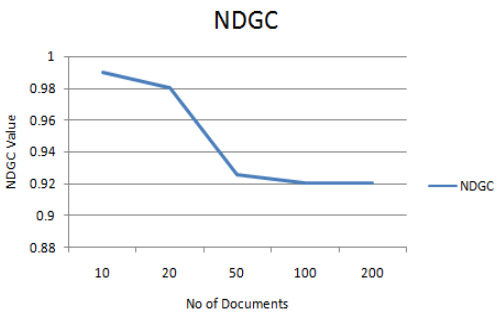

Figure 5. Evaluation of correlation algorithm using *NDCG* method.

6. Performance Measure

The performance analysis for the correlation method with n-gram based method is evaluated for the metrics such as accuracy, recall and specificity using quantitative calculations and tabulated in Tables 4 and 5.

Table 4. Performance measure for existing (n-gram based) method.

Top 'n' Outliers	Specificity	Recall	Accuracy
5	0.99	0.8	0.99
10	0.98	0.7	0.97
15	0.97	0.6	0.95
20	0.95	0.6	0.92
25	0.93	0.56	0.89

Table 5. Performance measure for correlation method.

Top 'n' Outliers	Specificity	Recall	Accuracy
5	0.99	0.8	0.99
10	0.98	0.8	0.98
15	0.98	0.8	0.97
20	0.98	0.85	0.97
25	0.98	0.85	0.97

Accuracy is defined as closeness of agreement between a measured quantity value and a true quantity value of measured [7]. Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes the condition. The accuracy for the correlation method and existing method for the data taken from both Tables 4 and 5 are projected in Figure 6.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

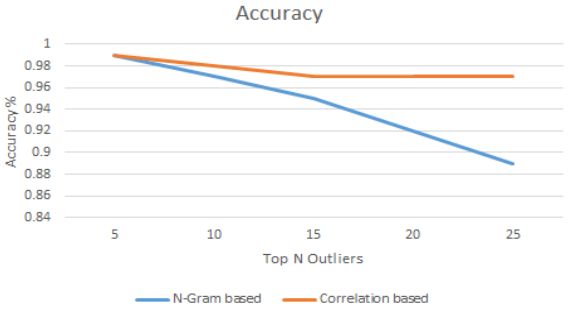

Figure 6. Comparative study of accuracy with n-gram based method to correlation method.

Recall in information retrieval is defined as the fractions of documents that are relevant to the query are successfully retrieved. The recall for the correlation method and existing method for the data taken from both Tables 4 and 5 are given in Figure 7.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

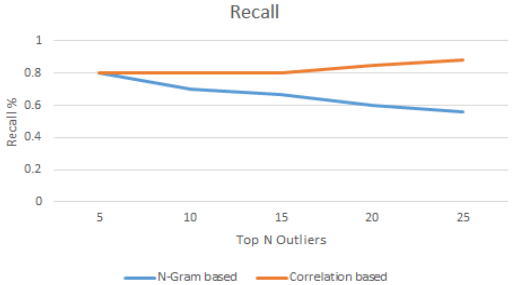

Figure 7. Comparative study of recall with n-gram based method to correlation method.

Specificity measures the ability of a test to be negative when the condition is actually not present, or how many of the negative test examples are excluded. The specificity for the correlation method and the existing method for the data taken from both Tables 4 and 5 are projected in Figure 8.

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

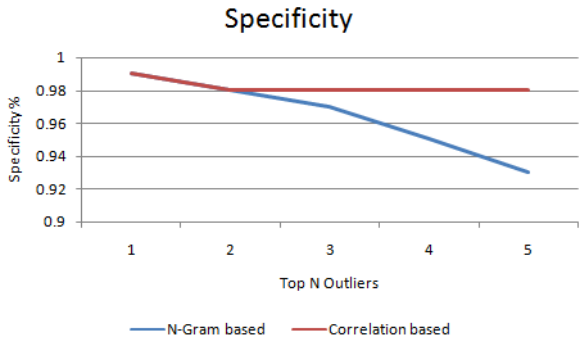

Figure 8. Comparative study of specificity with n-gram based method to correlation method.

Table 6. Nomenclature.

TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

7. Conclusions

The popularity of WWW has received a tremendous attention by majority of people to find and retrieve relevant information for various purposes. Therefore, most of the researchers pay attention to web content mining for extracting similar patterns. To shift this paradigm, this work mainly focuses on extracting dissimilar patterns called web outliers which have tremendous applications like search engines for improving the quality of search results, pattern detection, trend analysis, plagiarism detection and spam filtering. This research work addresses the web content outliers mining based on mathematical method through correlation method. Research in the above mentioned field has led to several new ideas and innovations. The following are the outcomes of the research:

- Survey on the evolution of outliers, types of outliers and various outlier mining techniques, especially web content outliers.
- Construction of domain dictionary.
- Design and implementation of web content outlier mining through correlation method.
- Analysis on the impact of results obtained with different cases of datasets for the correlation method.
- Performance evaluation for correlation ranking algorithm is done based on *NDCG* method.
- Performance analysis is done in terms of recall and accuracy of the correlation method with an existing method based on n-grams and structure oriented weighting technique.

From the above results obtained, it is observed that the proposed algorithm will be able to detect outliers present in all types of web datasets (structured, semi-structured and unstructured) and the false rate of correlation method is lower than the existing techniques.

8. Future Enhancement

As such this research has brought out the effectiveness of web content outlier mining through mathematical approach for all types of web documents. Further research in this area could be:

- Mining of outliers present for heterogeneous web documents containing hypertext, image, audio and video data.
- The reason for uncovered outlier documents (failed to detect) can be analysed.
- Other mathematical tools can be explored to improve further results.
- A benchmark dataset for comparing web content outlier algorithms can be accomplished.

References

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

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

[3] Agyemang M., Barker K., and Alhajj R., "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, 2005.

[4] Agyemang M., Barker K. and Alhajj R., "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, 2005.

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

[6] 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.

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

[8] Alqaraleh S. and Ramadan O., "Elimination of Repeated Occurrences in Multimedia Search Engines," *in the International Arab Journal of Information Technology*, vol. 11, no. 2, pp. 134-139, 2014.

[9] Arning A., Agrawal R., and Raghavan P., "A Linear Method for Deviation Detection in Large Databases," *in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Oregon, pp. 164-169, 1996.

[10] Barnett V. and Lewis T., *Outliers in Statistical Data*, Willey, 1994.

[11] Breunig M., Kriegel H., Ng R., and Sander J., "LOF: Identifying Density-Based Local Outliers," *in Proceedings of 2000 ACM SIGMOD International Conference Management of Data*, Dallas, pp. 93-104, 2000.

[12] Brin S. and Page L., "The Anatomy of a Large-Scale Hyper Textual Web Search Engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117, 1998.

[13] Chakrabarti S., *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann, 2002.

[14] Furnkranz J., "Separate-and-Conquer Rule Learning," *Artificial Intelligence Review*, vol. 13, no. 1, pp. 3-54, 1999.

[15] Hawkins S., He H., Willams G., and Baster R., "Outlier Detection using Replicator Neural Networks," *in Proceeding of the DaWaK02*, France, pp. 170-180, 2002.

[16] Jiang M., Tseng S., and Su C., "Two Phase Clustering Process for Outlier Detection," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 691-700, 2001.

[17] Knorr E. and Ng R., "Algorithms for Mining Distance-Based Outliers in Large Dataset," *in Proceeding of 24th VLDB Conference*, New York, pp. 392-403, 1998.

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

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

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

[21] Poonkuzhali G., Thiagarajan K., and Sarukesi K., "Set Theoretical Approach for Mining Web Content through Outliers Detection," *International Journal on Research and Industrial Applications*, vol. 2, no. 1, pp. 131-138, 2009.

[22] Poonkuzhali G., Thiagarajan K., Sarukesi K., and Uma V., "Signed Approach for Mining web Content Outliers," *in Proceeding of World Academy of Science, Engineering and Technology*, pp. 820-824, 2009.

[23] Ramaswamy S., Rastogi R., and Shim K., "Efficient Algorithm for Mining Outliers from Large Data Sets," *in Proceeding of the 2000 ACM SIGMOD International Conference on Management of Data*, Texas, pp. 427-438, 2000.

[24] Ruts I. and Rousseuw P., "Computing depth Contours of Bivariate Points Cloud," *Computational Statistics and Data Analysis*, vol. 23, no. 1, pp. 153-168, 1996.

[25] Siddiqui M., Fayoumi M., and Yusuf N., "A Corpus Based Approach to Find Similar Keywords for Search Engine Marketing," *International Arab Journal of Information Technology*, vol. 10, no. 5, pp. 460-466, 2013.

[26] Tax D. and Duin R., "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45-66, 2004.

[27] Xia H., Fan Z., and Peng L., "Web Text Outlier Mining Based on Domain Knowledge," *in Proceeding of the 2010 Second WRI Global Congress on Intelligent Systems*, Washington, pp. 73-77, 2010.

[28] Yang P. and Huang B., "An Efficient Outlier Mining Algorithm for Large Dataset," *in Proceedings of International Conference on Information Management, Innovation Management and Industrial Engineering*, Taipei, pp. 199-202, 2008.

Poonkuzahli Sugumaran has a very distinguished career span of nearly 15 year, currently professor and head of Information technology in Rajalakshmi Engineering College, Chennai. She obtained PhD in Computer Science from Anna University. Her areas of specialization are Web Mining, Outlier mining and Information Retrieval.

Kishore Kumar Ravi currently working as Assistant Professor in Department of Information Technology in Rajalakshmi Engineering College, Chennai. He obtained M.E in Computer Science from Anna University. His areas of specialization are Web Mining, Information Retrieval and Service Oriented Computing.

Thirumurugan Shanmugam currently working as Assistant Professor in Department of Information Technology in College of Applied Science-Sohar, Oman. He obtained PhD in Computer Science from Anna University. His areas of specialization are Network, Applied Mathematics and Software Reliability Engineering.