Signed-With-Weight Technique for Mining Web Content Outliers

S. Poonkuzhali
Rajalakshmi Engineering College, affiliated to Anna University,
Chennai, India.

P.Sudhakar
Kamaraj College of Engineering and Technology, affiliated to Anna University,
Chennai, India.

K.Sarukesi
Hindustan Institute of Technology and Science,
Chennai
India

ABSTRACT
Web outlier mining is dedicated for finding web pages which differ significantly from the rest of the web document taken from the same category. Most of the existing algorithms for web content outlier mining is developed for structured documents, whereas WWW contains mostly unstructured and semi structured documents. Moreover, the false positive rate in the existing algorithms for mining web content outlier is more than 30%. Therefore, there is need to develop a technique to mine web outliers from unstructured and semi structured document types with less false positive rate. This paper, concentrates on mining web content outliers which extracts the dissimilar web document taken from the group of documents of same domain. The proposed work implement a novel mathematical approach based on signed-with-weight technique for mining web content outliers which retrieves top n outlier web documents from both structured and unstructured web documents. The proven results show the performance measure of this approach in terms of precision and recall is more than 90%. Also, the false positive rate of this algorithm is less than 15%.

Keywords
Dissimilarity Weight, Outlier mining, Term Frequency, Weighted approach, Web content mining, Web content Outliers.

1. INTRODUCTION
In the present era of Internet, World Wide Web is an accumulated and interactive medium for accessing an enormous conglomeration of information. The information in the web consists of diverse data types such as structured data, semi structured data and lack of structure of Web data, automated discovery of targeted or unexpected knowledge/information becomes a challenging task. The eruptive growth rate of web data leads to many complications in retrieval of relevant information. In addition to this, the navigation of many links in an attempt to find desired information cause wastage of user time and makes the user annoyed. The unrestricted amount of information contains inessential and redundant information which increases the indexing space and time complexity. The above mentioned issues degrade the quality of search results which in turn deteriorate the performance of search engines. Thus, developing user-friendly and automated tools for organizing and retrieving relevant information without redundancy from the web documents has been on a higher demand. Web mining is an emerging research area focused on resolving these problems. Web Mining has adapted techniques from the field of data mining, databases and information retrieval. Web mining is categorized into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining refers to the discovery of user access patterns from search logs or other activity logs and it also helps to find patterns for a particular group of people, or for Internet users in a particular region. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. It usually involves analysis of the in-links and out-links of a web page, and it has been used to classify web pages, to create similarity measure between web pages and to rank results in search engines. Web content mining aims to extract/mine useful information or knowledge from Web page contents.

In recent years the growth of the World Wide Web has been jaw dropping. There is a huge amount of text, document, image, audio and video present in the internet world and it is still rising in an alarming rate. As there is an explosion in information, retrieving interesting information has become a herculean task. Web content mining uses the idea of combing data mining and knowledge discovery to retrieve the information based on the user query. Web content mining refers to the detection of useful information from web contents. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, information extraction from structured and unstructured web pages, detection and elimination of noisy content and mining opinion sources on the web. Web content mining is focused in two different directions: First involves mining of content from web documents directly. Second, involves indirect approach which improves the search results of other tools like search engine.

The currently available algorithms for web content mining do not concentrate on dissimilar patterns which contains outlying data such as noise, redundant and rare interesting data. Outliers are observations that deviate so much from other observations to arouse suspicions that they might have been generated using a different mechanism. Outliers 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. Traditional outlier mining algorithms which are designed for numeric datasets cannot be used on web datasets because they contain multimedia. Web outliers are data present in the web which has different characteristics from the web data taken from the same category. Different contents of the web pages from the category in which they were taken constitute web content outliers. Web content outliers mining concentrates on discovering outliers from the web contents of a web page [7],[11].

At present, very few algorithms have been developed for mining outliers over web content. Still those approaches

focused only on structured documents, but, in reality most of the documents in World Wide Web is unstructured and semi-structured. 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 false positive rate. In this research work, a mathematical approach based signed and weighted technique is developed for mining web outliers in both structured and unstructured web documents. Removal of irrelevant content not only leads to reduction in indexing space and time complexity, but also improves the quality and accuracy of search results

Outline of Paper

Section 2 presents the related works done on this area. Section 3 presents Signed-with-weight technique for web content outlier deduction. Section 4 presents the experimental results and performance evaluation. Finally, Section 5 presents conclusions and future work.

2. RELATED WORKS

Ali et al[1] presents an overview of the major developments in the area of detection of Outliers in numerical datasets. 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. Anguilli et al[2] proposes, a new definition of distance-based outlier and an algorithm, called HilOut, designed to efficiently detect the top n outliers of a large and high-dimensional data set. Breunig et al [4] introduced a new method for finding outliers in a multidimensional dataset through density based approach which uses a local outlier (LOF) for each object in the dataset, indicating its degree of outlier. Ramaswamy et al[15] 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 kth nearest neighbor and developed a highly efficient partition-based algorithm for mining outliers. The above author's devised algorithms based on distance and density based approaches for detecting outliers present only on numeric data sets.

Bing et al [3] presents characteristics of web and various issues on web content mining. Raymond Kosal et al[16] discuss about research areas in web mining and different categories of web mining . Malik et al [9]-[10] establish the presence of outliers on the web and discusses some practical applications and motivation behind web outlier mining. 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.

The above authors [11] proposed an n-gram based algorithm using domain dictionary for mining web content outliers, which explores the advantages of n-gram techniques as well as HTML structure of web documents. The same authors [12] developed a 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. Based on the above ideas, authors [13] prolonged the work by presenting HyCOQ a hybrid algorithm which extracts the power of n-gram and word based systems. 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. Xia Xuosong et al [17] present a framework and algorithm for mining Chinese web text outliers based on improved Vector Space Model (VSM) and n-gram combined with domain knowledge.

G.Poonkuzhali et al [6] worked in a new perspective using mathematical approach based on set theoretical for mining web content outliers considering different web outliers namely irrelevant, redundant and inconsistent web content The same authors developed an algorithm based on signed approach for mining web content outliers using organized domain dictionary[7]. They have also applied statistical approach for retrieving relevant information from both structured and unstructured documents [8]. In the proposed work, the authors have extended their previous work based on signed approach for improving the precision.

3. SIGNED-WITH-WEIGHT TECHNIQUE FOR WEB CONTENT OUTLIERS DETECTION

The proposed algorithm explores the advantages of full word matching, signed and weighted approach using domain dictionary. Initially the input web document Di is taken and it is pre-processed into 'm' words. Pre-process contains the following steps i.e. stemming, stop words elimination 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. Then a word (Wj) taken from document Di is searched on the domain dictionary. If the word (Wj) is found in the dictionary, then positive hit count is incremented by its corresponding term frequency else negative hit count is incremented by its corresponding term frequency. This process is carried out for all words in that web document. Then the dissimilarity weight DWi is computed for the document Di. The same process is carried out for all the extracted web documents. Finally, rank the dissimilarity weight in ascending order. The top 'n' documents are declared as an outlaid web document.

The details of each process are described in the following subsections.

3.1 Document Extraction

Input documents are extracted from the search engines or web crawlers belonging to a particular domain based on user interest. This algorithm works well for both structured and unstructured documents. In case, if structured documents are extracted, then all the tags get removed before entering into the pre-processing phase. Unstructured documents can be directly parsed for pre-processing.

3.2 Pre-Processing

The pre-processing phase transforms the extracted data into a structure form that will be more easily and effectively processed for the purpose of the user. 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

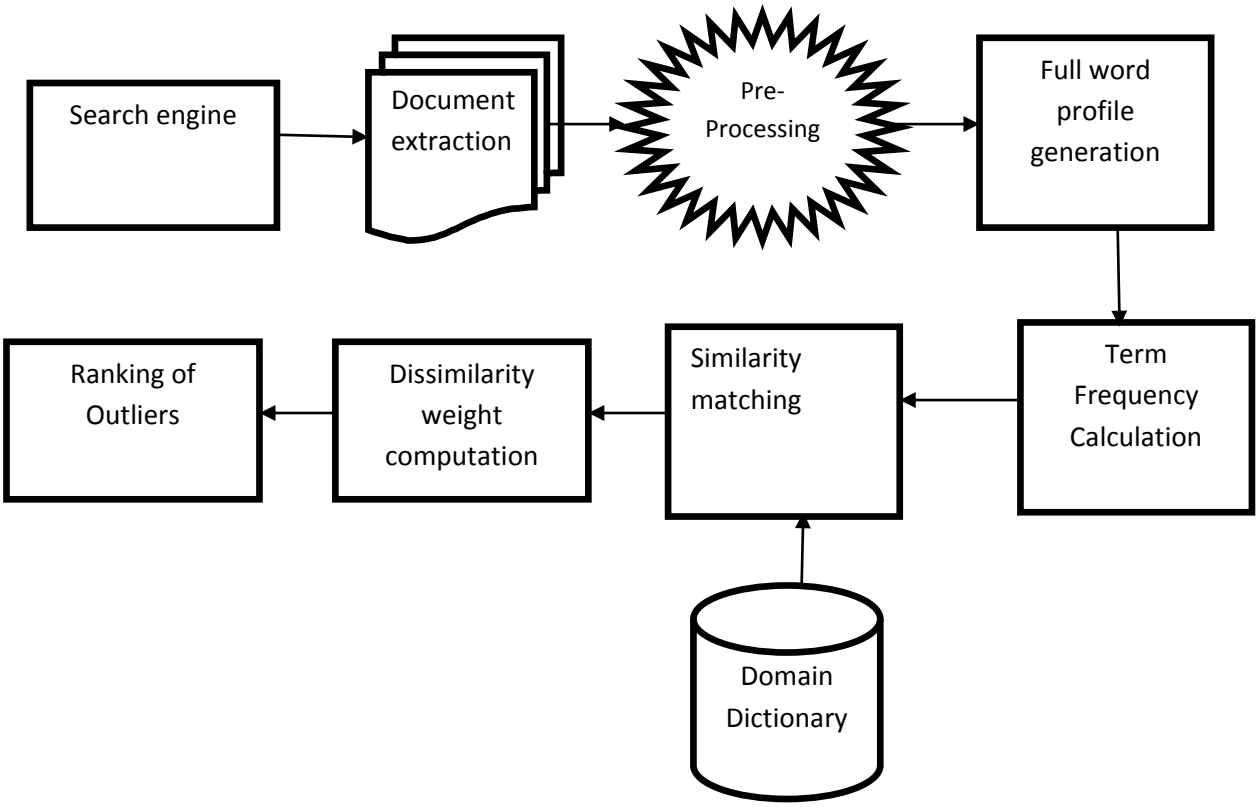

Fig 1. Architectural design of the proposed system

stop words, stemming and tokenization. Stop words are common words that carry less important meaning than keywords. Usually search engines remove stop words from a keyword phrase to return the most relevant result which in turn improves search performance. Stemming is a process for removing the commoner morphological and inflexional endings from words in English. Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens.

3.3 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.

3.4 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..

3.5 Similarity Matching

After pre-processing and term frequency calculation, all the words from the extracted documents are searched in the domain dictionary. If that word matched with domain dictionary then positive hit count is incremented by its corresponding term frequency else negative hit count is incremented by its corresponding term frequency.

3.6 Dissimilarity weight Computation

Dissimilarity weight (DW) computation is done, after calculating the total positive hits and negative hits for each document. Dissimilarity weight (DW) computation is the ratio of sum of positive hits multiplied with α- value and negative hits multiplied with β-value to the sum of positive hits and negative hits. Here the α- value should be greater than β-value as it shows more similarity with domain words.

3.7 Detection of Outliers

All the extracted documents are ranked in ascending order based on dissimilarity weight. Finally, the top n documents are declared as outliers. The algorithm for detecting web content outliers is given below:

Algorithm: Signed-with-weight Approach for Web Content Outliers Mining

Input: Web document D = {D1,D2,D3, ----- DN}

Method: Signed-with-Weight Approach

Output: Outlier documents

1: Extract the input web document Dn where

$1 \leq n \leq N$;

2: Pre-process the entire extracted document;

3: Generate the full word profile;

4: Initialize i=1;

5: Consider Di ;

6: Initialize $PositiveHits(i) = 0$; NegativeHits(i) $= 0$;

7: Compute the term frequency TF (Wj) for all words in Di where $1 \leq j \leq m$ and m is the total number of words;

8: if Wj exist in domain dictionary then

$$\text{PositiveHits(i)} = \text{PositiveHits(i)} + \text{TF (Wj)};$$

else

$$\text{NegativeHits(i)} = \text{NegativeHits(i)} + \text{TF (Wj)};$$

9: Increment j;

10: Repeat step 8 and step 9 till $j \leq m$;

11: Compute dissimilarity weight DWi as

$$\text{DWi} = \frac{(\text{PositiveHits(i)} * \alpha) + (\text{NegativeHits(i)} * \beta)}{\text{PositiveHits(i)} + \text{NegativeHits(i)}}$$

Where $\alpha > \beta$; *

12: Increment i;

13: Repeat from step 5 till $i \leq N$;

14. Sort DW in ascending order;

15. Display top' n' outlier web documents

***** Since the positive hits get multiplied by α and this corresponds to Wj of Di in domain dictionary, α > β gives high value for DWi and so we focus on ascending order. If β > α, the result will be descending order. If α = β it will not be possible to order the documents, as all the documents will have unique value.

4. EXPERIMENTAL RESULTS

Two sets of experiments have been done to evaluate the performance of the proposed approach for detecting web content outliers. First set of experiments presents the results of proposed algorithm applied over different types of test cases. As there is no real dataset for mining web content outliers present in unstructured document, a new dataset consisting of 100 web pages related to web content mining from different search engines are created for testing purpose. A dictionary is compiled using first 25 web content mining pages. The dataset consists of remaining 75 relevant documents (RD) related to web content mining domain and 75 outlier documents (OD) which is not restricted to web content mining. Datasets with three different cases are used to produce results as follows:

CASE 1: Dataset with less number of outlier documents (OD<RD).

CASE 2: Dataset with balanced outlier and relevant documents (OD=RD).

CASE 3: Dataset with oft repeated outlier documents (OD>RD).

The input documents are first pre-processed and then the term frequencies are computed for all the words. Followed by that the dissimilarity weight for each document is computed for detecting the outlaid document. The weight factor for positive hits should be greater than the weight factor for negative hits in computing dissimilarity weight measure. Table 1, 2, 3 gives the results obtained through this signed-with-weight technique for three set of test cases.

Table 1. Signed-with- weight technique results for Case 1 datasets (Outlaid document < Relevant document).

Document Size	Actual Outlier	Top 'n' outliers	Outliers detected through proposed algorithm
100	30	10	10
		15	14
		20	18
		25	22
		30	27

Table 2. Signed-with weight technique results for Case 2 datasets (Outlaid document = Relevant document).

Document Size	Actual Outlier	Top 'n' outliers	Outliers detected through proposed algorithm
100	50	10	10
		20	20
		30	29
		40	38
		50	44

Table 3. Signed-with weight technique results for Case 1 datasets (Outlaid document >Relevant document).

Document Size	Actual Outlier	Top 'n' outliers	Outliers detected through proposed algorithm
100	70	10	10
		25	25
		40	40
		55	55
		70	64

The second set of experiments presents the comparative results of proposed approaches with existing approach on N-Gram based algorithm (Malik et al 2005a). The experiment is conducted with 225 resume pages and 25 recruiter pages retrieved from the dmoz open directory project (www.dmoz.org). The same input structured dataset was used by Malik et al to evaluate web content outliers algorithm based on N-Gram approach. The domain dictionary is constructed using 50 resume pages. The remaining 175 resume pages and 25 recruiter pages are taken as test data. The top 'n' outliers detected by the proposed approaches and existing approach is shown in table 4 and in figure 2.

The precision of the proposed approach is high when compared to the existing approach and is presented in figure 3. The existing approach using WCOND Mine algorithm based on n-grams works only for structured documents. Even the processing time and memory utilization of existing approach will be more as it supports partial matching of strings. Also, the second set of experiment results listed in Table 4 implies that the false positive rate for the WCOND-Mine algorithm is more than 30% while increasing the top n outliers. But the proposed approach using signed-with-weight technique based on full word matching works for structured and unstructured documents with less than 15% false positive rate.

4.1 Precision

Precision is the number of correct results divided by the number of all returned results It can also be evaluated at a given cut-off rank, considering only the top most results returned by the system.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

4.2 Recall

Recall is the number of correct results divided by the number of results that should have been returned.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4.3 F-Score

F-Score is a measure of a test's accuracy. F-Score is the harmonic mean of precision and recall. F1 score reaches its best value at 1 and worst score at 0.

$$\text{F} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.4 Accuracy

Accuracy is the measure which matches the actual value of the quantity being measured.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Nomenclature:
TP- True Positive
TN – True Negative
FP- False Positive
FN – False Negative

Table 4.
Comparative Study with existing approach (WCOND-Mine) with proposed approach (Signed-with-weight)

Document Size	Top 'n' outliers	Outliers detected	
		WCOND-Mine Algorithm	WAMWCO Algorithm
400	10	7	10
	20	13	19
	30	18	29

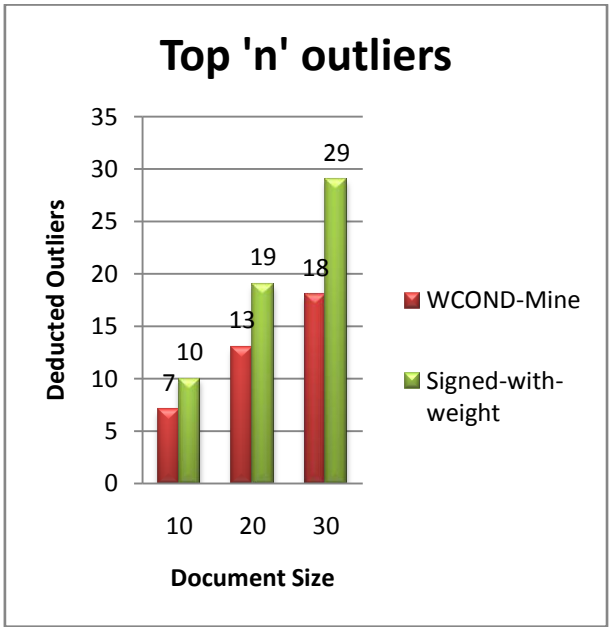

Fig 2: Comparative results of Existing approach (WCOND-Mine) with Proposed approach (Signed-with-weight) for detecting top 'n' outliers

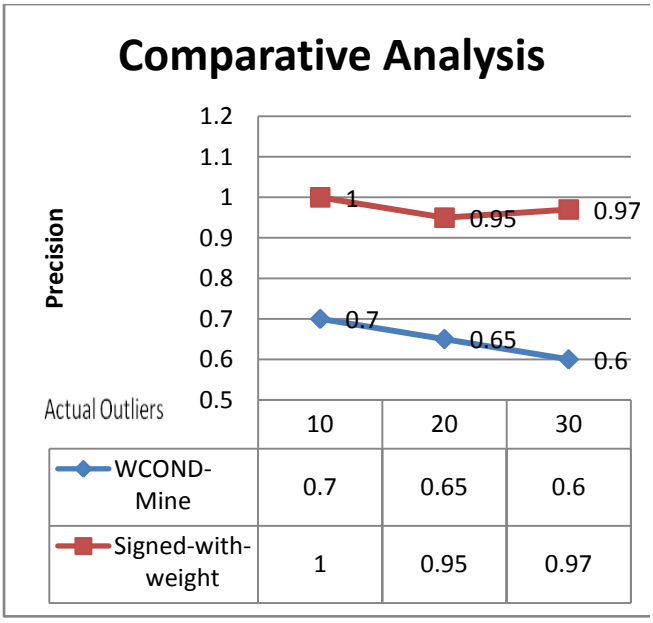

Fig . 3 Precision of Proposed approach (Signed-with-weight) with Existing Approach (WCOND – Mine)

5. CONCLUSION AND FUTURE WORK

Nowadays, 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 and plagiarism detection. The proposed work retrieves the outlier web document through signed-with-weight technique. The emphasis of this algorithm is that it works for both structured and unstructured web documents. This eminent approach gives high precision with less false positive rate than the existing approaches. Future work aims at mining web content outliers containing images.

6. ACKNOWLEDGEMENT

The authors would like to thank Prof. Ponnammal Natarajan worked as former Director – Research , Anna University-Chennai and Currently Advisor, (Research and Development), Rajalakshmi Engineering College for her constant encouragement and discussion in bringing out this research paper in an efficient manner.

7. REFERENCES

[1] Ali S. Hadi,A. H. M. Rahmatullah Imon(2009), Mark Werner, Detection of outliers Overview, Wiley Interdisciplinary Reviews: Computational Statistics, Volume 1, Issue 1, pp-57-70.

[2] Anguilli, F., and Pizzuti, C., Elomaa,T. (Eds.). Fast Outlier Detection in High Dimensional Spaces. PKDD, LNAI 2431, 2002, pp 15-27

[3] Bing Liu, Kevin Chen- Chuan Chang , Editorial: Special issue on Web Content Mining , SIGKDD Explorations, Volume 6, Issue 2.

[4] Breunig, M.M., Kriegel, H-P., Ng R.T., and Sander, J. LOF: Identifying Outliers in Large Dataset. Proc. of ACM SIGMOD 2000, Dallas, TX 2000.

[5] Barnett, V. and Lewis, T. Outliers in Statistical Data. John Willey, 1994

[6] 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

[7] 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, 2009, pp -820-824.

[8] G.Poonkuzhali, R.Kishore Kumar, R.Kripa Keshav and K.Sarukesi paper titled "Statistical Approach for Improving the Quality of Search Engine" " in the Book " RECENT RESEARCHES IN APPLIED COMPUTER AND APPLIED COMPUTATIONAL SCIENCE", included in ISI/SCI Web of Science and Web of Knowledge,Venice, Italy, 2011, pp-89-93.

[9] Malik Agyemang, Ken Barker and Rada S. Alhajj, Framework for Mining Web Content Outliersb. In: ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004, pp 590-594.

[10] Malik Agyemang, Ken Barker, Reda Alhajj, Web outlier mining: Discovering outliers from web datasets, Intelligent Data Analysis,Vol. 9, No (5)/2005, pp 473-486

[11] Malik Agyemang, Ken Barker and Rada S. Alhajj Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams' ACM Symposium on Applied Computing., Santa Fe, New Mexico,2005, pp 482-487.

[12] Malik Agyemang Ken Barker and Rada S. Alhajj WCOND –Mine : Algorithm for detecting Web Content Outliers from Web Documents. IEEE Symposium on Computers and Communication. 2005.

[13] Malik Agyemang Ken Barker and Rada S. Alhajj, Hybrid Approach to Web Content Outlier Mining without Query Vector. Springer –Berlin, 2005,Vol. 3589.

[14] Malik Agyemang, Ken Barker, Reda Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intelligent Data Analysis,Vol. 10, No (6)/2006, pp 521-538.

[15] Ramaswamy S, Rastogi R, Shim k, Efficient Algorithm for mining outliers from large data sets, proc. Of ACM SIGMOD 2000, pp 127 – 138.

[16] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD, July 2000, Vol-2, pp 1-15.

[17] Xia Huosong, Fan Zhaoyan, Peng Liuyan, "Chinese Web Text Outlier Mining Based on Domain Knowledge," Intelligent Systems, WRI Global Congress on, vol. 2, pp. 73-77, 2010 Second WRI Global Congress on Intelligent Systems, 2010