Web Content Outlier Mining Through Mathematical Approach and Trust Rating

G.POONKUZHALI[1] , K.SARUKESI[2] , G.V. UMA[3]

[1] Department of Computer Science and Engineering, Rajalakshmi Engineering College,
Affiliated to Anna University
[2] Hindustan Institute of Technology and Science
[3] Department of Information Science & Technology, Anna University
Chennai, Tamil Nadu, INDIA
[1]poonkuzhali.s@rajalakshmi.edu.in , [2]profsaru@gmail.com, [3] gvuma@annauniv.edu

Abstract: - In this Internet era, the WWW is flooded with voluminous amount of information with more replicated and irrelevant web pages. As the unnecessary and duplicated web pages increase the indexing space and time complexity, finding and removing these pages become a significant issue among the information retrieval and web mining research communities as most of the people rely on search engines to get the required information. Web content outlier mining plays a decisive role in covering all these aspects. Existing algorithms for web content outlier mining focuses attention on applying weightage only to structured documents whereas in this research work, a mathematical approach based on two way rectangular representations, signed approach of trust rating and correlation method is developed for retrieving right information without duplicates present in both structured and unstructured web documents.

Key-Words: - Correlation, Duplicates, Outliers, Relevance, Signed approach, Term frequency, Web content.

1 Introduction

Due to voluminous amount of information available on the web, most of the people like to perform transaction over the internet. Web content outliers mining play a powerful way to manage corporate reputation. Today, most of the people prefer to do their jobs in online, as the web provides all necessary information that promotes their business. Reputation management is rapidly becoming an important strategy for many organizations in decision making. In order to track and report on millions of blogs, search engines, review sites, news groups and online forums can be a challenging task for an individual or companies to promote their business. Existing web mining algorithms do not consider documents having varying contents within the same category called web content outliers. Web content outliers mining concentrates on finding outliers such as noise, irrelevant and redundant pages from the web documents. Web content outliers are resulted with distinct web document. From this approach, unique patterns can be retrieved by eliminating unrelated patterns obtained by mining the Web Content Outliers.

Mining Web Content Outliers may lead to identification of competitors, fraud deduction, emerging business trends in electronic commerce and improvement in the quality of results returned by search engines. Shifting through the dynamic, unstructured and ever growing web data for outliers is more challenging than finding outliers in numeric datasets. Most of the existing web mining algorithms concentrated on finding frequent patterns while neglecting less frequent ones that are likely to contain outlying data. Thus developing user friendly and automated tools for providing relevant information without redundant links accurately, quickly, and easily to retrieve desired information becomes a primary challenge among web mining research communities.

Existing algorithm for web content outliers mining focus attention only on applying weightage to structured documents. The proposed work provides, a mathematical approach based on signed, correlation and rectangular representation of trust rating to mine related web content without duplication for both structured and unstructured web documents.

2 Related Works

Malik Agyemang et al establish the presence of outliers on the web with various types of outliers present on the web and designed a framework for mining web content outliers using full word matching assuming the existence of domain dictionary. The above authors developed the work with n-gram techniques for partial matching of strings with domain dictionary[7]-[9]. Malik Agyemang et al. enhanced the same work without domain dictionary. Based on the above ideas, Malik Agyemang et al prolonged the work by presenting HyCOQ which a hybrid algorithm that draws from the power of n-gram based and word based system[10] . There is a remarkable improvement in recall with hybrid documents compared to using raw words and n-grams without a domain dictionary still it covers mining only structured web documents. G.Poonkuzhali et al presented the mathematical approach based on set theoretical and signed approach for mining web content outliers[1]-[2]. The same authors developed an algorithm for eliminating redundant web content through rectangular approach and correlation method.[3]-[4]. K.Thiagarajan et al. implemented weighted graph approach of trust reputation management through signed concept which can also be applied for retrieving the relevant content, SMS and SPAM filtering[6].

Giuseppe Antoio Di Lucca et al. proposed an algorithm based on clone detection and similarity metrics to detect duplicate pages in web sites and application implemented with HTML which works only for structured web documents[5]. Min-yan Wang et al. suggested a web page de-duplication method in which the information including original websites and web titles are extracted to eliminate duplicated web pages based on feature codes with the help of URL hashing[11]. Through this method large-scale duplicated web pages can be eliminated but extraction of feature codes takes much time. Yunhe Weng et al. come up with an idea of improved COPS (Copy Detection Algorithm) scheme which aims to protect intelligent property of the document owner by detecting overlap among documents[12]. This method performs similarity computation only for the pages that are relevant to the suspicious pages. Here, the semantic keyword alone is considered as terms to compute relevant measure. In spite of its good performance in both effectiveness and efficiency in tackling the large scale duplicated pages, the cost for building the inverted index of the semantic keywords is expensive. Zhongming Han et al. developed a novel multilayer framework for detecting duplicated web pages through two similarity text paragraphs detection algorithms based on Edit distance and bootstrap method[13]. This method achieves high performance in detecting duplicates efficiently simply by tag statistic and text comparison, still it cannot find duplicates among multiple web pages.

3 Architectural design

In the proposed system, web documents are extracted from the search engines based on user query to the web. The extracted web documents D is sliced into ‘n’ web pages and each page is divided into ‘m’ words. Then the sliced web-document is preprocessed. The 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 preprocessing the term frequency of all the words are calculated. Followed by that relevancy checking of web document is performed using signed approach .Then redundancy checking of web documents is performed using correlation and signed approach of trust rating. Finally, a mined web document is obtained which contains desired information of the end user.

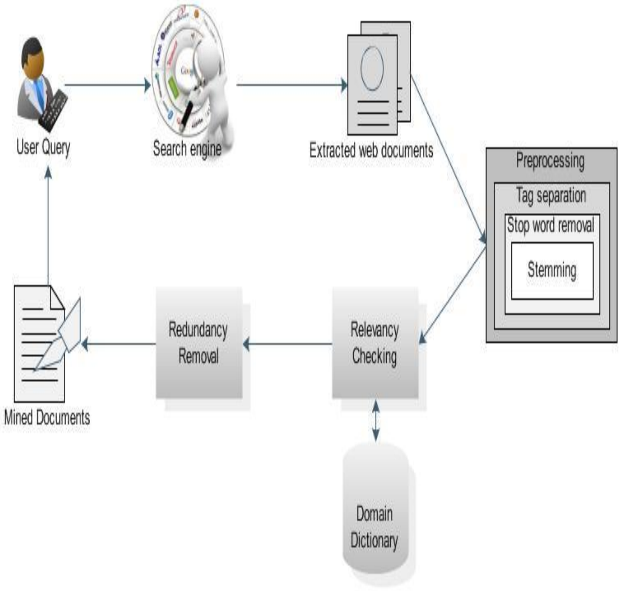

Fig. 1 System Design of the proposed work

4 Signed Approach for relevancy computation

The proposed algorithm explores the advantages of full word matching and signed approach using organized domain dictionary where the indexing is done based on the length of the word. First, the input web document is preprocessed and separated into white spaced words. The full word profile for the document is generated in matrix form (i.e., $W_{1,5}$ - represents 5th word in 1st document). Following the above process, term frequency for all the words are found out. Then the j^{th} word from i^{th} document is taken and its length is calculated (i.e., $| W_{ij} |$) and depending on the number of characters, the respective index on the domain dictionary is searched. If the word (W_{ij}) is found in the dictionary, then positive count is incremented by its corresponding term frequency else negative count is incremented by its corresponding term frequency. This process is carried out for all words in that web document. Finally, positive count is compared with the negative count to check the relevancy of that web page. If the positive count is less than the negative count, then that page is irrelevant, otherwise it is considered as more relevant.

Algorithm 1: Relevancy Computation algorithm through signed approach

```
Input  : Domain Dictionary, Web Document $D_i$
Output  : Relevant Web Document and Irrelevant Web Document.

1: extract the input web document $D_n$ 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 the document $D_i$
6: initialize $PC=0$, $NC=0$.
7: compute the term frequency $TF(W_j)$ for all
     words in $D_i$ where $1 \leq j \leq m$
8:  if $W_j$ exist in domain dictionary then
                update $PC = PC + TF(W_j)$
     else
                update $NC = NC + TF(W_j)$
9:  increment j
10: repeat step 8 and step 9 till $j \leq m$.
11: compare Positive count (PC) with Negative
     count (NC)
     if $PC < NC$ then
          $D_i$ is outlaid (irrelevant) web document.
     else
          $D_i$ is relevant web document.
12: increment i
13: repeat from step 5 till $i \leq N$.
```

4.1 Two way rectangular representation for checking relevant document

In two way representation, consider only n x m matrix for relevancy computation, where n represents total number of documents and m represents maximum number of words in taken from any of the extracted documents. Ignore all columns which consist of only zero entries

Table 1. Rectangular representation for checking relevancy in given documents

Word / Document	W_1	W_2	W_3	--	--	W_m
D_1	(1,0)	(1,0)	(0,1)	--	--	(0,0)
D_2	(0,1)	(0,1)	(1,0)	--	--	(0,1)
D_3	(1,0)	(1,0)	(0,1)	--	(1,0)	(0,0)
-	--	--	--	--	--	--
-	--	--	--	--	--	--
D_n	(1,0)	(0,1)	(1,0)	--	(0,0)	(0,0)

Table 2. Nomenclature

(D_i, W_j)	j^{th} word from i^{th} web document.
PC	Positive Count
NC	Negative Count
D_i	i^{th} web document.
TF	Term Frequency

5 Correlation method for redundancy checking of web document

The relevant document extracted from the above phase is sent to this phase for further processing. First pre-processing is done for all the documents. Then, i^{th} document and $i+1^{th}$ document are taken for redundancy computation. Common words between these documents are extracted and the term frequency for all the common words is found out. 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 remove the second document from the original document set. This process is repeated for the remaining documents.

Algorithm 2: Redundancy computation using linear correlation – Ref.[3]

Input: Web document.
Output: Identification and elimination of redundant web document.

```
1: extract the relevant web document $D_n$ where $1≤n≤N$.
2: pre-process the entire extracted document.
3: initialize $i=1$.
4: initialize $j=i+1$.
5: consider the document $D_i$ and $D_j$.
6: extract the common words present in $D_i$ and
   $D_j$. Let $T$ be the total number of common
   words.
7: compute the term frequency $TF(W_k)$ for the
   common words in $D_i$ and $D_j$ where $1≤k≤m$.
8: perform the correlation between $D_i$ and $D_j$.
   determine: $X_i$ to the term frequency for all the
   words in document $D_i$ and $Y_j$ to the term
   frequency for all the words in document $D_j$.
   calculate: $\sum X_i$, $\sum X_i^2$, $\sum Y_j$, $\sum Y_j^2$, $\sum X_i Y$
   compute: $R_1$, $R_2$ and $R_3$
      Where $R_1 = \sum X_i^2 - ((\sum X_i)^2/T)$,
            $R_2 = \sum Y_j^2 - ((Y_j)^2/T)$,
            $R_3 = \sum X_i Y_j - ((\sum X_i \sum Y_j)/T)$
   perform: $R_{xy} = R_3/(\sqrt{R_1} * \sqrt{R_2})$
9: If the $R_{xy}$ is equal to 1 then
      $D_i$ and $D_j$ are redundant, hence eliminate
         $D_j$ from set of documents.
   else
      $D_i$ and $D_j$ are not redundant, hence retain
   both the documents.
10: increment $j$, and repeat from step 5 to step 9
    until $j≤N$.
11: increment $i$, and repeat from step 4 to step
    10 until $i<N$.
```

5.1 Two way rectangular representation for checking redundant document

Two way rectangular representation for checking redundant document holds two important characteristics:

- Upper triangular matrix
- In diagonal there is path containing all vertices.

Table 3. Rectangular representation for checking redundancy in given documents

Web documents	D_2	D_3	D_4	D_5	D_6
D_1	(+,-)	(+,-)	(+,-)	(+,-)	(+,+)
D_2	(0,0)	(+,-)	(+,-)	(+,-)	(+,-)
D_3	(0,0)	(0,0)	(+,-)	(+,-)	(+,-)
D_4	(0,0)	(0,0)	(0,0)	(+,-)	(+,-)
D_5	(0,0)	(0,0)	(0,0)	(0,0)	(+,-)

Table 4. Explanation and Comparative Study with Signed Approach

Page Comparison	Signed Values	Result
(D_i, D_j)	(+ , +)	Redundant Document
(D_i, D_j)	(+, -)	Not redundant

$$F(x) = (D_i, D_j) \begin{array}{l} = (+,+) \text{ or } (+,-) \text{ if } j > i \\ = (0,0) \text{ otherwise} \end{array}$$

Note: Since the first co-ordinate is always compared with other pages, it never have the possibility of having -ve sign , therefore, (- ,-) and (- , +) is not considered.

Nomenclature: (D_i, D_j) implies ith web document is compared with jth web document for redundancy checking

6. Results and Discussion

An experimental analysis has been done with 150 documents extracted from the web related to web mining domain. These documents are first pre-processed and then the relevancy computation using signed approach is performed. Followed by that, redundancy computation based on correlation method is done only for the relevant documents. Here results obtained for 15 input documents after relevancy computation is listed in table 5 and the results of redundancy computation is projected in table 6.. Finally mined web document without redundancy in computed in table 7. The precision and recall of web documents after relevancy computation and redundancy computation is given in fig 2.

Table 5. Experimental results of Relevancy Computation

D.No	Document Name	Result
D_1	An integrated framework for WCM.pdf	Relevant
D_2	Software engineering.pdf	Irrelevant
D_3	Deep_WCM.pdf	Relevant
D_4	Copy of Deep_WCM.pdf	Relevant
D_5	Elimination of Redundant Links.pdf	Irrelevant
D_6	Framework_WCOM.pdf	Relevant
D_7	Identify duplicated content.pdf	Irrelevant
D_8	Medical Mining.pdf	Irrelevant
D_9	Neural Analysis.pdf	Irrelevant
D_{10}	Outlier_lattice.pdf	Irrelevant
D_{11}	Page content rank.pdf	Relevant
D_{12}	Signed Approach.pdf	Irrelevant
D_{13}	WCM.pdf	Relevant
D_{14}	Fuzzy approach.pdf	Irrelevant
D_{15}	Copy Page Content rank.pdf	Relevant

Table 6. Experimental results of Redundancy Computation

D.No	D_3	D_4	D_6	D_{11}	D_{13}
D_1	0.277	0.277	0.306	0.056	0.3046
D_3	*	1	0.152	0.174	0.1733
D_4	*	*	0.152	0.174	0.1733
D_6	*	*	*	0.064	0.3739
D_{11}	*	*	*	*	0.2527

Table 7. Resultant Document

D.No	Document Name
D_1	An integrated framework for WCM.pdf
D_3	Deep_WCM.pdf
D_6	Framework_WCOM.pdf
D_{11}	Page content rank.pdf
D_{13}	WCM.pdf

$$\text{Precision} = \frac{\text{Relevant} \cap \text{Retrieved originally}}{\text{Retrieved after refinement}}$$

$$\text{Recall} = \frac{\text{Relevant} \cap \text{Retrieved originally}}{\text{Relevant}}$$

Table 8. Precision of the proposed approach

Dataset Size	Relevant document through signed approach	Relevant document Manually computed	Precision
50	30	35	0.88
75	50	58	0.88
100	60	70	0.88
150	85	100	0.87
200	112	130	0.88

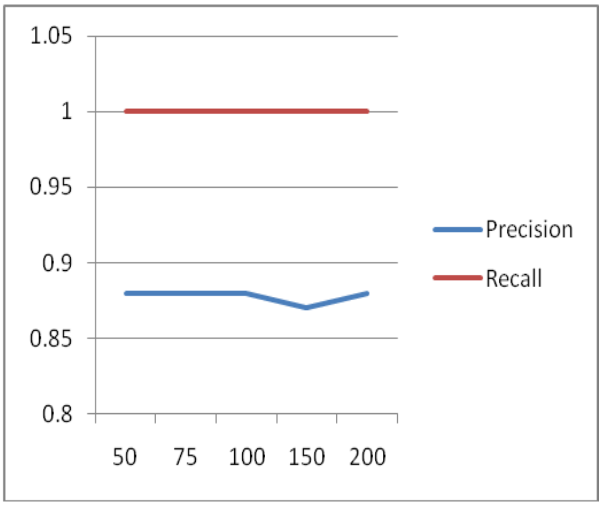

Fig 2. Precision and Recall

7 Signed approach for retrieving unique document

Signed graph approach is applied for the output of relevance analysis and redundancy computation phases. Trust rating for the signed weight +- is assigned with + sign and for the remaining signed weight (++,-+,--) trust rating is assigned with – sign. Final decision is made based on trust rating. If the trust ratings holds + sign, then it indicates mined web documents.

Table 9. Signed approach for Mined Document

Signed Format	Factors	Trust Rating
++	Relevant and Redundant	-
+-	Relevant and Not Redundant	+
-+	Not Relevant and Redundant	-
- -	Not Relevant and not Redundant	-

Relevant document without redundancy implies unique mined documents.

Conclusion

The massive growth of information sources available on the World Wide Web has forced the web mining researchers to develop the automated tools to locate relevant resources quickly without duplicates. In this paper, a mathematical approach based on signed, rectangular and correlation methods are applied to detect and eliminate irrelevant and redundant document. The essence of this algorithm is that it works for both structured and unstructured web documents. Another key feature is that the results obtained are accurate.

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 intuitive ideas and fruitful discussions with respect to the paper's contribution.

References:

[1] 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

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

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

[4] G Poonkuzhali , K Sarukesi and G V Uma, Detection and Removal of Redundant Web Document through Rectangular and Signed Approach, *International Journal of Engineering , Science and Technology,* Vol. 2 (9)-2010,pp 4126-4132

[5] Giuseppe Antoio Di Lucca, Massimiliano and Anna Rita Fasolina, An Approach to identify duplicated web pages. In: proceedings of the 28th Annual *International Computer Software and Applications Conference,* IEEE computer Society press,2002.

[6] K. Thiagarajan, A. Raghunathan, Ponnamal Natarajan, G. Poonkuzhali and Prashant Ranjan, Weighted Graph Approach for Trust Reputation Managements, *International Conference on Intelligent Systems and Technologies,* Published in Proc. Of World Academy of Science and Technology- Vol 56, 2009,pp-830-836.

[7] 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.

[8] 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.

[9] 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.

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

[11] Min-yan Wang and Dong-Sheng Liu , The Research of web page De-duplication based on web pages Re-shipment Statement. *First Interrnational Workshop on Database Technology and Applications,* 2009,pp.271-274

[12] Yunhe Weng, Lei Li and Yixin Zhong , Semantic keywords-based duplicated web pages removing, *IEEE,* 2008

[13] Zhongming Han, Qian Mo, Liu and Jianzhi , Effectively and Efficiently Detect Web Page Duplication, *IEEE* ,2009.