Advanced Materials Research *Online: 2010-12-06*
ISSN: 1662-8985, Vols. 171-172, pp 543-546
doi:10.4028/www.scientific.net/AMR.171-172.543
© 2011 Trans Tech Publications, Switzerland

Correlation Based Method to Detect and Remove Redundant Web Document

G.Poonkuzhali[1,*] , R.Kishore Kumar[2], R.Kripa Keshav[3],
P.Sudhakar[4] and K.Sarukesi[5,+]

[1,2,3] Department of Computer Science and Engineering, Rajalakshmi Engineering College,
Chennai,Tamil-Nadu, India.[4] Vernalis Systems, Chennai, Tamil Nadu, India ,
[5] Hindustan Institute of Technology and Science-Chennai,Tamil Nadu, India
[*] poonkuzhali.s@rajalakshmi.edu.in , [+].profsaru@gmail.com

Keywords: Duplicates, Linear Correlation, Term Frequency, Web Document

Abstract. The enrichment of internet has resulted in the flooding of abundant information on WWW with more replicas. As the duplicated web pages increase the indexing space and time complexity, finding and removing these pages becomes significant for search engines and other likely system which will improve on accuracy of search results as well as search speed. Web content mining plays a vital role in resolving these aspects. Existing algorithm for web content mining focus attention on applying weightage to structured documents whereas in this research work, a mathematical approach based on linear correlation is developed to detect and remove the duplicates present in both structured and unstructured web document. In the proposed work, linear correlation between two web documents is found out. If the correlated value is 1 then the documents are said to be exactly redundant and it should be eliminated otherwise not redundant.

Introduction

In today's world search engine has been a critical tool for searching information. Accuracy of the search engine results become bottle neck for the users due to redundancy. Most of the web search engines typically employ conventional information retrieval and data mining techniques to discover useful information from the web content. Web content mining is defined as the automatic search of information resource available online, and involves mining web data content. R. Kosala et al[4]summarized the research works done for unstructured data and semi-structured data from information retrieval view. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation.

Major issue involved in information retrieval view is redundancy. The problems caused by redundancy are wastage of memory space and increase of retrieval time which in turn reduces the efficiency of search results. In this approach linear correlation is applied to find the redundant web documents. Correlation is a statistical measurement of the relationship between two variables. A possible correlation ranges from +1 to –1. A zero correlation indicates that there is no relationship between the variables. A correlation of –1 indicates a perfect negative correlation. A correlation of +1 indicates a perfect positive correlation [8]. The proposed work provides, a mathematical approach based on linear correlation method to mine unique web content for both structured and unstructured web documents. The results revealed that, this method can achieve satisfactory results.

Related Works

G Poonkuzhali et al. [1]. Proposed a Set theoretical approach used for detecting and eliminating redundant links in web documents. The temporal link-analysis algorithm works based on last modification time returned by the HTTP response based on timestamp of nodes and links to rank the new pages, was presented by Shiguang Ju et al.[5]. Min-yan Wang et al.[3] developed web pages reshipment process where original websites and web titles are extracted to eliminate duplicate web pages based on feature codes. Di Lucca et al. [2] devised an approach, based on similarity metrics for the detection of duplicated pages in Web sites and applications. The analysis of numerous Web sites and Web applications is performed to evaluate this approach. The experiments illustrated that, the proposed methods detected clones among static Web pages and the efficiency of the method was proved by a manual verification. The method produced comparable results, but different computational costs were involved. Similarity metrics such as Levenshtein distance technique and frequency based technique to detect duplicated clients, server etc.,. Yunhe Weng et al. [6] used semantic keywords combined with sentence overlapping for removing duplicated web pages. Zhongming Han et al. [7] proposed multilayer framework to detect duplicated web pages based on tag statistic and text similarity.

Architectural Design

In this framework, web documents are extracted from search engines based on user query. Then pre-processing task such as stop words elimination, removal of image, video etc., other than text, stemming and tokenization is done for the extracted web documents. Initially, first two documents 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. 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.

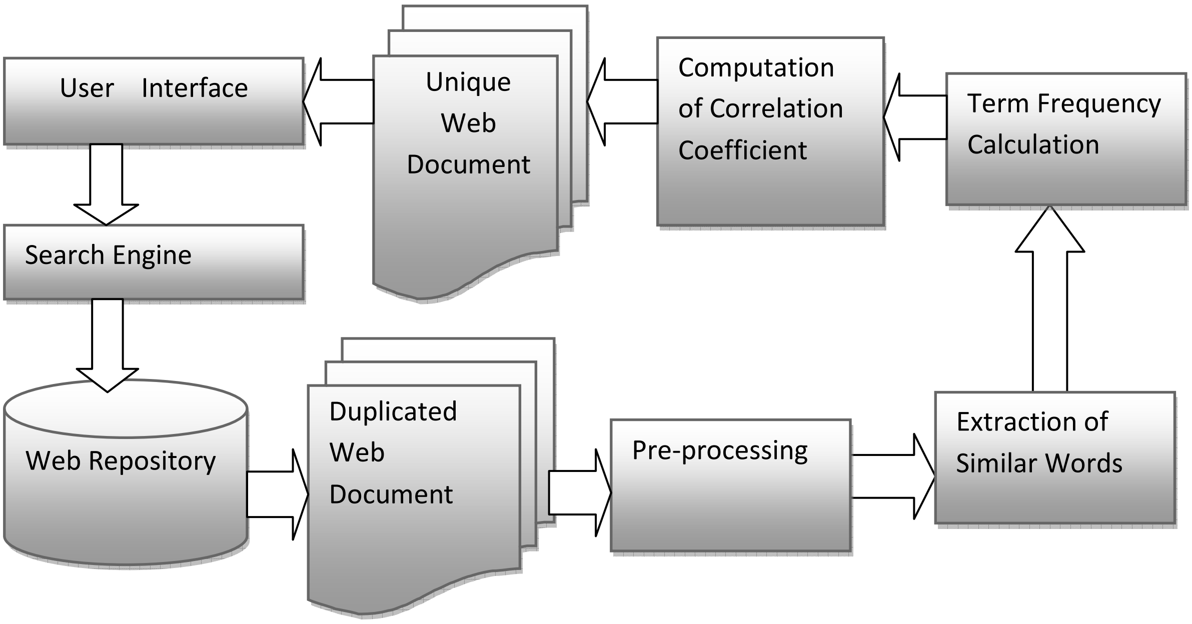

Fig. 1 Architectural Design of the Proposed Work

Algorithm for the proposed system

```
Input: Web document.
Method: Linear Correlation method.
Output: Identification and elimination of redundant web document.
Step1: Extract the input web document $D_i$ where $1\le i\le N$.
Step2: Pre-process the entire extracted document.
Step3: Initialize i=1.
Step 4: Initialize j=i+1.
Step 5: Consider the document $D_i$ and $D_j$.
Step 6: Extract the common words present in $D_i$ and $D_j$ and denoted by T.
Step 7: Find the term frequency TF ($W_k$) for the common words in Di with Dj where $1\le k\le m$.
Step 8: Perform the correlation between $D_i$ and $D_j$.
        i)   Determine $X_i$ of $W_i$ from $D_i$ where $X_i$ is the term frequency similarly, $Y_j$ of $W_j$
             from $D_j$ where $Y_j$ is the term frequency.
        ii)  Calculate : $\sum X_i$ , $\sum X_i^2$ , $\sum Y_j$ , $\sum Y_j^2$, $\sum X_iY_j$
        iii) Compute : R1 , R2 and R3 Where $R1 =\sum X_i^2 -((\sum X_i)^2/T)$ , $R2 =\sum Y_j^2 - ((\sum Y_j)^2/T)$,
             $R3 =\sum X_iY_j -((\sum X_i \sum Y_j)/T)$
        iv)  Perform Rxy Where $Rxy= R3/(\sqrt{R1} * \sqrt{R2})$

Step 9: If the $R_{xy}$ is equal to 1 then $D_i$ and $D_j$ are redundant, eliminate $D_j$ from set of documents.
        Else $D_i$ and $D_j$ are not redundant.
Step 10: Increment j, and repeat from step 5 to step 9 until $j\le N$.
Step 11: Increment i, and repeat from step 4 to step 10 until $i<N$.
```

Experimental Results

Here the URL of 5 web documents is taken as test data. Then the content of these URLs are fetched for further processing. Followed by that pre-processing is done for these entire document. Then term frequency for all the words is calculated. Finally, correlation coefficient is computed for all these documents..Table 2 contains correlated value of the web documents. From the obtained values, redundancy between the documents is computed.

Table 1. URL of the input documents

Document Number	URL
D_1	www.waset.org/journals/waset/v56/v56-150.pdf
D_2	www.computer.org/portal/web/csdl/doi/10.../COMPSAC.2009.110
D_3	www.tedescoanalytics.com/content/.../pdf/Neural%20Analysis.pdf
D_4	inkinghub.elsevier.com/retrieve/pii/S0167865509001962
D_5	portal.acm.org/citation.cfm?id=968022
D_6	www.seminarprojects.com/Thread-signed-approach-for-mining-web-content-outliers

Table 2. Experimental Results

URL	D1	D2	D3	D4	D5	D6
D2	0.459979552	*	*	*	*	*
D3	0.029999089	0.625712664	*	*	*	*
D4	0.655694746	0.147125696	0.417428165	*	*	*
D5	0.152078405	0.305756507	0.642773455	0.063789834	*	*
D6	1.00000007	0.459979552	0.029999089	0.655694746	0.152078405	*

From Table 2 it is clear that D_1 and D_6 are redundant, others are not redundant. Hence, out of the set of 6 documents taken, it is sufficient to focus on 5 documents after removing redundancy.

Conclusion

Web is a highly dynamic information source which serves for diversity of user communities. Web contains duplicate pages in abundance. Therefore, efficient identification and removal of these duplicated web pages becomes a vital issue that has arisen from the escalating amount of data. In this paper, a mathematical approach based on Linear Correlation Method is applied to detect and eliminate redundant document, thereby, improves the quality of search results.

Acknowledgment

The authors would like to thank Dr. Ponnammal Natarajan worked as Former Director – Research , Anna University- Chennai, India and currently an Advisor, (Research and Development), Rajalakshmi Engineering College and Dr. K..Ravi, Associate Professor, Department of Mathematics, Sacred Heart College-Tirupattur, India for their intuitive ideas and fruitful discussions with respect to the paper's contribution.

References

[1] G.Poonkuzhali, K.Thiagarajan, K.Sarukesi: Elimination of redundant links in web pages – Mathematical Approach,World Academy of Science, Engineering and Technology,V52, (2009) pp.562-565.

[2] Giuseppe Antoio Di Lucca, Massimiliano: 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). pp: 481- 486

[3] Min-yan Wang, Dong-Sheng Liux: The Research of web page De-duplication based on web pages Re-shipment Statement, First Interrnational Workshop on Database Technology and Applicationsv(2000), pp.271-274

[4] Raymond Kosala: Web Mining Research:A Survey,IEEE(2000).

[5] Shiguang Ju, Zheng Wang, Xia Lv: Improvement of page ranking algorithm based on timestamp and link, International Symposium on Information Processing(2008), pp. 36-40.

[6] Yunhe Weng, Lei Li, Yixin Zhong: Semantic keywords-based duplicated web pages removing, IEEE(2008).

[7] Zhongming Han, Qian Mo, Liu, Jianzhi: Effectively and Efficiently Detect Web Page Duplication, IEEE(2009).

[8] Robert Johnson in: Elementary Statistics, sixth edition, Duxbury press, Belmount California.