Websites:
http://www.sciencepub.net/newyork
http://www.sciencepub.net

Emails:
newyorksci@gmail.com
editor@sciencepub.net

New York Science Journal

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

A Mathematical Approach with Term Frequency Ranking for Mining Web Content Outliers

S. Sathya Bama [1], M.S. Irfan Ahmed [2], A. Saravanan [1]

[1.] Department of MCA, Sri Krishna College of Technology, Coimbatore, Tamil Nadu 641042, India
[2.] Department of MCA, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India
ssathya21@gmail.com

Abstract: The Internet is a massive collection of information that makes it extremely difficult to search and retrieve the required and valuable information. So, Search engine became an important tool for searching various data from the web. The primary evaluation of search engine is effectiveness and efficiency. While searching for information through search engines, always users retrieve redundant and irrelevant information. This replicated and uninteresting information affects both the effectiveness and efficiency of search engine by wasting users' time by browsing the uninterested documents and its accessing time. Web content outlier mining plays a significant role in identifying and removing these redundant document (outliers) which is an important issue among the information retrieval and web mining research communities since most of the people rely on search engines to get the required information. Most existing algorithms for web content outlier mining focuses attention on applying weightage only to the common terms in the documents whereas in this research work, a mathematical approach based on term frequency ranking to identify the duplicates, and uses the domain dictionary to check for relevant document has been carried out to improve the effectiveness and efficiency of the search engine.

[Sathya Bama S, Irfan Ahmed MS, Saravanan A. **A Mathematical Approach with Term Frequency Ranking for Mining Web Content Outlier**. *N Y Sci J* 2021;14(7):49-55]. ISSN 1554-0200(print); ISSN 2375-723X (online). http://www.sciencepub.net/newyork. 6. doi:10.7537/marsnys140721.06.

Keywords: Correlation Coefficient, Search Engines, Term Frequency Ranking, Web content Mining, Web Content Outliers

1. Introduction

In this Internet era, World Wide Web is the interactive and intermediate medium for accessing large amount of required information. Search engine act as an intermediate tool to access these information where efficiency and effectiveness is a major concern. Effectiveness, measures the ability of the search engine to find the right and required information and efficiency measures the time taken by the search engine to retrieve those documents (Bruce et al., 2009). Retrieving required information without any redundancy from the web has become complicated and challenging task for those search engines due to increase in the amount of information stored in the web. To answer these issues web mining has become an important research area.

Most of the web search engines employ conventional information retrieval and data mining techniques to discover useful information from web content effectively and efficiently. Applying data mining techniques for mining the web content is termed as web content mining. Web content mining is the process of mining, extraction and integration of useful data, information and knowledge from Web page contents. Recently, there is a rapid development of research activities in the Web Content Mining area. 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 (Kosla & Blockeel, 2000), (Campos et al., 2006).

However, the existing web mining algorithms concentrates on web contents and do not consider the redundancy in the retrieved documents (outliers) (Poonkuzhali et al., 2010), (Poonkuzhali et al., 2009a). Web content outliers mining give attention on finding outliers such as noise, irrelevant and redundant pages from the web documents. By removing these outliers unique patterns can be retrieved by eliminating unrelated patterns obtained by mining the Web Content (Poonkuzhali et al., 2009b), (Poonkuzhali et al., 2009c).

Web content outlier mining not only is helpful to detect outliers when a web portal is hacked but also may lead to the discovery of emerging business patterns and trends (Agyemang et al., 2005). Unlike traditional outlier mining algorithms designed solely for numeric data sets, web outlier mining algorithms should be applicable for varying types of data such as text, hypertext, video, audio, image and HTML tags (Agyemang et al., 2004)).

Existing web content outliers mining algorithm focus only on applying weightage to terms that are common to the document. The proposed work provides a mathematical approach based on ranking the terms using frequencies to mine related web content without duplication.

The proposed algorithm discovers the advantages of full word matching using organized domain dictionary for checking relevancy. Here the indexing is done based on the length of the word (Poonkuzhali et al., 2009b). First, the input web document is preprocessed and separated into white spaced words. Each word in the document is compared with the domain dictionary. If the word is found in the dictionary, then count is incremented by one else count is incremented by one. This process is carried out for all words in that web page. Finally, count is negative, then that page is irrelevant, otherwise it is considered as more relevant.

2. Related Work

Internet search engines are special sites on the Web that are designed to help people find information stored on other sites. There are differences in the ways various search engines work, but they all perform three basic tasks:

- They search the Internet based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

Elizabeth Liddy explained about how the search engine works (Liddy, 2001).

Web mining is an emerging research area that focuses on resolving problems while accessing and managing information on the web. In general, web mining tasks can be classified into three major categories, web structure mining, web usage mining and web content mining (Kosla & Blockeel, 2000). Web Content Mining aims to extract useful information from the web pages based on their contents (Liu et al., 2004), (Wang et al., 2008), (li, Wu & Ji, 2008), (Pokorny & Smizansky, 2005). Agent based systems has been introduced for mining the hyperlinks on a web page to find a quality web page (Gopalan & Akilandeswari, 2005). The 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 has been introduced in (Agyemang et al., 2004). A framework was designed for mining web content outliers using full word matching assuming the existence of domain dictionary with n-gram techniques for partial matching of strings with domain dictionary [9]. A Hybrid algorithm that draws from the power of n-gram based and word based system in (Agyemang et al., 2005a). A WCOND-Mine algorithm for mining web content outliers using n-grams without a domain dictionary is proposed in (Agyemang et al., 2005b), (Agyemang et al., 2006), (Huosong et al., 2010), (Brian & Page, 1998) where Vector space model is used for dissimilarity computation.

The mathematical approach based on set theoretical and signed approach for mining web content outliers presented in (Poonkuzhali et al., 2009a) and (Poonkuzhali et al., 2009b). A fuzzy clustering technique called C-Means algorithm to mine usage profiles from web log data has been introduced in (Castellano et al., 2006). A better understanding of Arabic text classification techniques is achieved in (Zubi, 2009).

The importance of using suitable measures and methods to evaluate the performance of Web document classification (Po, 2008). Ioan Dzitac et al. proposes the structure into two sections. The proposes the new structure with two sections, first one briefly discusses the different web mining tasks and the second one is focusing on advanced Artificial Intelligence (AI) methods for information retrieval and web search, link analysis, opinion mining and web usage mining (Dzitac & Moisil, 2008).

An algorithm based on clone detection and similarity metrics to detect duplicate pages in web sites and application for structured web documents has been introduced in (Lucca et al., 2002). A web-page de-duplication method by which the information from websites and web titles are extracted to eliminate duplicated web pages based on feature codes using URL hashing has been introduced for which the extraction of feature codes takes much time (Wang & Liu, 2009). Copy Detection Algorithm (COPS) scheme that aims to protect intelligent property of the document owner by detecting overlap among documents has been suggested where the semantic keyword alone is considered as terms to compute relevant measure and the cost for building the inverted index of the semantic keywords is expensive (Weng, Li & Zhong, 2008). A novel multilayer framework for detecting duplicated web pages through two similarity text paragraphs detection algorithms based on Edit distance and bootstrap method is proposed in (Han et al., 2009) but still it cannot find duplicates among multiple web pages. A traditional weighting technique TF.IDF from Information Retrieval which is commonly used in text mining is used in (Wan Zulkifeli et al., 2012), (Salton, 1988). The Linear Correlation based Method to Detect and Remove Redundant Web Document in (Poonkuzhali et al., 2011). The experiment analysis has been made and it shows the TF.IDF technique from Information Retrieval is not only compatible to use in detecting web outliers, it even returns better results than the previous works. After retrieving and removing redundancy in the web pages, Rank should be made for each rank before presenting the pages to the user. Page Rank is a numeric value that represents how important a page is on the web (Pokorny & Smizansky, 2005), (Brian & Page, 1998).

All the above works on web content mining, lack simplicity of concept and computation. In this proposed work, the statistical correlation has been applied to remove the duplicates from the retrieved web pages which is more efficient than existing algorithm in time and space.

4. Architecture of Proposed System

The Architecture of the proposed system has depicted in Figure 1. First the user gives the input query. Based on that query the documents are retrieved by the search Engine from web servers. Most of the documents retrieved from the search engine may or may not be relevant to the user query.

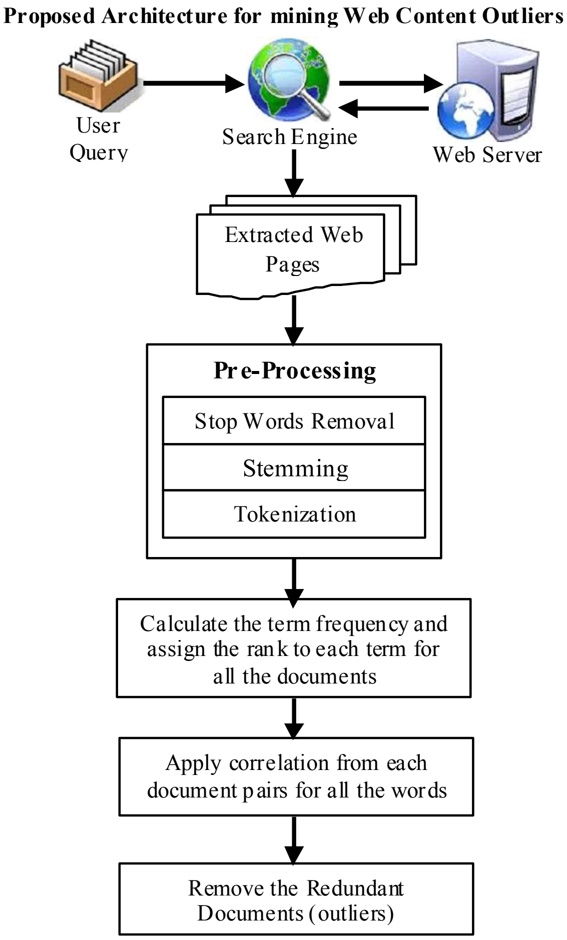
Proposed Architecture for mining Web Content Outliers
User Query
Search Engine
Web Server
Extracted Web Pages
Pre-Processing
Stop Words Removal
Stemming
Tokenization
Calculate the term frequency and assign the rank to each term for all the documents
Apply correlation from each document pairs for all the words
Remove the Redundant Documents (outliers)

Figure 1. The Architecture of the proposed system

It also assumes the existence of domain dictionary. The extracted documents undergo the preprocessing step which consists of stop words removal, stemming and tokenization. Preprocessing is necessary to make the entire document in the same format. Stop words are common words that carry less important meaning than keywords. Stemming is the process for reducing derived words to their stem or root form – generally a written word form. Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing. Next step is the term frequency calculation. Frequency of all the terms in the documents is calculated. Then the scoring or ranking should be made for each term based on the number of times it occurred in the document. The term having highest frequency should be ranked 1, similarly for other terms. Next step is to compare all the document pairs to check for redundancy. Find the terms T in the documents D_i and D_j by making union for the words in the documents along with the rank.

Next step is to compare all the document pairs to check for redundancy. Find the terms T in the documents Di and D_j by making union for the words in the documents along with the rank. If the word Wk is present in document Di and not in D_j then the rank of the word Wk for the document D_j will be zero. Apply the Spearman's rank correlation coefficient which is given by the equation, Eq(1).

$$\rho = \left|1-\frac{6\sum d^2}{n(n^2-1)}\right| \qquad (1)$$

Where d is given by $(x_i - y_i)$ where x_i and y_i are frequency of the term i in document D_1 and D_2 respectively. n is total number of words in document D_1 and D_2. The value lies between 0 and 1. If the value is 1 for document d_i and d_j then d_j is the duplicate of d_i. If there is no common word between the two documents d_i and d_j then the value will be 0.

5. Proposed Algorithm

Input : Web document, Domain Dictionary.
Method: Statistical Method
Output: Extraction of unique web document.
Step 1: Input the query Q to search Engine
Step 2: Extract set of Web Documents D_i related to the given query where $1 \le i \le r$, r is the number of retrieved documents.
Step 3: Pre-process the entire extracted document by removing stop words, stemming, and tokenization.
Step 4: Initialize the count to 0.

Step 5: For each document, Check for the availability of all the words in the document with domain dictionary.
Step 6: For each word, if the word is available then increment the count, else decrement the count.
Step 7: If the count value is positive then, the document is irrelevant else the document is irrelevant.
Step 9: Find the term frequency TF(W_{ik}) for all the words W_k in the document D_i where $1 \leq k \leq m$, m is the number of words in document D_i.
Step 10: Assign the term frequency ranking TFR(W_{ik}) to each words W_k in the document D_i where $1 \leq k \leq m$. m is the number of words in document D_j.
Step 11: Initialize i=1 and j=i+1
Step 12: Find the terms T in the documents D_i & D_j by making union for words in the documents along with rank.
Step 13: Perform the Spearman's rank correlation coefficient in Eq(1).
Step 14: If the ρ value is 1 then D_j is duplicate document, else D_j is not a duplicate.
Step 15: Increment j, and repeat from step 11 to step 14 until $j \leq r$.
Step 16: Increment i, and repeat from step 11 to step 15 until i<r.

Explanation

Consider the table of term frequencies for 3 documents denoted D_1, D_2, D_3. Compute the Term Frequency (TF) weights for the terms car, auto, insurance, best, form each document D_1, D_2, D_3. The sample TF value is given in Table 1.

Table 1. The Sample TF Value

Terms	Documents			
	D1	D2	D3	D4
Car	27	4	0	27
Auto	0	33	24	0
Insurance	0	33	12	0
best	14	0	0	14

Compute ranking for each term in each document D_1, D_2, D_3. Identical values (rank ties or value duplicates) are assigned a rank equal to the average of their positions in the ascending order of the values. The Term Frequency Ranking is given in Table 2.

Table 2. The Term Frequency Ranking

Terms	Documents			
	D1	D2	D3	D4
Car	1	3	0	1
Auto	0	1.5	1	0
Insurance	0	1.5	2	0
best	2	0	0	2

By using the formula in Eq (1), the Correlation Coefficient for each document pair is calculated and is listed in the Table 3.

Table 3. The Correlation Coefficient of all Document Pairs

	D1	D2	D3	D4
D1	-	0.25	0	1
D2	-	-	0.05	0.25
D3	-	-	-	0
D4	-	-	-	-

Since the ρ value of D_1 and D_4 is 1, the document D_4 is a redundant document and therefore it can be removed.

6. Experiment Result

An experimental analysis has been done with the dataset that consist of 35 web pages from the web in the topic of Web Content Mining. These documents are pre-processed and redundancy computation based on statistical method is done only for the retrieved documents. The correlation coefficient has been calculated for all document pairs. Finally the document having coefficient value 1 should be removed since it is redundant document. The comparison has been made with the performance of n-gram method, TF.IDF, Linear Correlation and Ranking Correlation which is shown in the Figure 2.

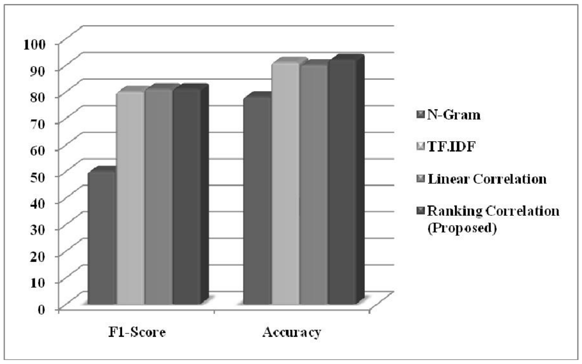

Figure 2. Comparison with Existing Method

Precision

It is the percentage of retrieved documents that are in fact relevant to the query

$$\Pr ecision = \frac{\mathrm{Re}\,lavant \cap \mathrm{Re}\,trived}{\mathrm{Re}\,trived}$$

Recall

It is the percentage of documents that are relevant to the query and were, in fact, retrieved

$$\mathrm{Re}\,call = \frac{\mathrm{Re}\,lavant \cap \mathrm{Re}\,trived}{\mathrm{Re}\,lavant}$$

F1-Score

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

$$F1 - Score = \frac{2 * \Pr ecision * \mathrm{Re}\,call}{\Pr ecision + \mathrm{Re}\,call}$$

Accuracy

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

$$Accuracy = \frac{\mathrm{Re}\,lavant}{Total\ \ Documents}$$

7. Conclusion

The huge growth of information sources available on the World Wide Web has forced the web mining researchers to develop new and effective algorithms and tools to identify relevant information without duplicates. In this paper, a mathematical approach based on correlation method is applied to detect and eliminate uninteresting document. The strength of this algorithm and key feature is that the results obtained are more accurate than other existing algorithm. Future work aims at experimental evaluation of web content mining in terms of reliability and to explore other mathematical concepts for web content mining.

Corresponding Author:
Ms. S. Sathya Bama
Department of MCA
Sri Krishna College of Technology
Coimbatore, Tamil Nadu 641042, India
E-mail: ssathya21@gmail.com

References

[1]. Bruce Croft W, Donald Metzler and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison-Wesley, 2009.

[2]. Kosla, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations. 2000. Vol. 2, Issue 1, pp: 1-15.

[3]. Ricardo Campos, Gael Dias and Celia Nunes. WISE : Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques, International conference on Web Intelligence, IEEE/WIC/ACM, 2006.

[4]. G. Poonkuzhali, G.V.Uma and K.Sarukesi. Detection and Removal of Redundant Web Content through Rectangular and Signed Approach. International Journal of Engineering Science and Technology, 2010, Vol. 2(9), 4026-4032.

[5]. G.Poonkuzhali, K.Thiagarajan and K.Sarukesi. Set theoretical approach for mining web content through outliers detection. International Journal on Research and Industrial Applications, 2009, Vol. 2, pp. 131-138.

[6]. 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, 2009, Vol. 56, pp. 820- 824.

[7]. G. Poonkuzhali, K.Thiagarajan and K.Sarukesi. Elimination of Redundant Links in Web Pages - Mathematical Approach. World Academy of Science, Engineering and Technology,2009, 52.

[8]. M. Agyemang, K. Barker and R.S. Alhajj. Mining web content outliers using structure oriented weighting techniques and n-grams. Proceedings of ACM SAC. New Mexico, 2005.

[9]. M. Agyemang, K. Barker and R.S. Alhajj. Framework for Mining Web Content Outliers. ACM Symposium on Applied Computing, 2004, pp. 590-594.

[10]. Elizabeth Liddy. How a Search Engine Works. Searcher: The Magazine for Database Professionals, 2001. Volume 9, Number 5, pg.38.

[11]. Bing Liu, Kevin Chen- Chuan Chang. Editorial: Special issue on Web Content Mining. SIGKDD Explorations, 2004, Volume 6, Issue 2.

[12]. Cheng Wang, Ying Liu, Liheng Jian and Peng Zhang. A Utility based Web Content Sensitivity Mining Approach. International Conference on Web Intelligent and Intelligent Agent Technology (WIIAT). IEEE/WIC/ACM, 2008.

[13]. Hongqi li, Zhuang Wu, Xiaogang Ji. Research on the techniques for Effectively Searching and Retrieving Information from Internet, International Symposium on Electronic Commerce and Security, 2008, IEEE.

[14]. Jaroslav Pokorny, Jozef Smizansky. Page Content Rank: An approach to the Web Content Mining. Proceedings of the IADIS International Conference on Applied Computing, 2005, February 22-25, 2 Volumes.

[15]. N.P. Gopalan, J. Akilandeswari. Distributed, Fault-tolerant Multi-agent Web Mining System for Scalable Web Search. 5th WSEAS International conference on Applied Informatics and Communications. Malta, September 15-17, 2005, pp. 384-390.

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

[17]. 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.

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

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

[20]. Brian, S. and Page, L. The anatomy of a large-scale hyper textual Web search engine, Computer Networks 30 (1-7), 1998, pp. 107-117.

[21]. G. Castellano, A. M. Fanelli and M. A. Torsello. Mining usage profiles from access data using fuzzy clustering. 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal, September 22-24, 2006.

[22]. Zakaria Suliman Zubi. Using Some Web Content Mining Techniques for Arabic Text Classification. Proceedings of the 8th WSEAS international conference on Data networks, communications, computers, 2009. pp. 73-84.

[23]. Ioan Pop. Web Document Classification and its Performance Evaluation. 9th WSEAS International Conference on Evolutionary Computing (EC’08). Sofia, Bulgaria, May 2-4, 2008.

[24]. Ioan Dzitac, Ioana Moisil. Advanced AI Techniques for Web Mining. Proceedings of the 10th WSEAS international conference on Mathematical methods, computational techniques and intelligent Systems, 2008, pp. 343-346.

[25]. Giuseppe Antoio Di Lucca, Massimiliano, Anna Rita Fasolina. An Approach to identify Duplicated web pages. Proceedings of the 28th Annual International Computer Software and Applications Conference, 2002, IEEE computer Society press.

[26]. Min-yan Wang, 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.

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

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

[29]. W.R. Wan Zulkifeli, N. Mustapha and A. Mustapha. Classic Term Weighting Technique for Mining Web Content Outliers. International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2012). Penang, Malaysia, 2012.

[30]. G. Salton. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley Editors, 1988.

[31]. 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, 2011, Vols. 171-172, pp 543-546.

7/22/2021