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

S. Sathyam Bama*, M. S. Ifran Ahmed and A. Saravanan

Department of MCA, Sri Krishna College of Technology, Coimbatore - 641042, Tamil Nadu, India;
ssathyam21@gmail.com, msirfan@gmail.com, a.saravanan21@gmail.com

Abstract

The Internet is a considerable collection of information that makes it extremely difficult to search and retrieve the required and valuable information. The primary objective of this paper is to provide the user with efficient and effective result through search engine, as the result set contains irrelevant and redundant data called outliers. In this research work, a mathematical approach called Spearman's rank correlation coefficient has been used to calculate the correlation between the document pairs. If the correlation value is 1, then the document is redundant which can be removed. This method depends on the term frequency of common words between the document pairs that is ranked based on the frequency value. This method improves the effectiveness, efficiency and reliability of the search engine. The comparison has been made with the performance of n-gram method, TF.IDF, Linear Correlation and Ranking Correlation. Thus the experimental result shows that the proposed method improves the precision, recall, f-score and accuracy. Thus the result produced by this method improves accuracy.

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

1. Introduction

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. Web Content Mining aims to extract useful information from the web pages based on their contents. Agent based systems has been introduced for mining the hyperlinks on a web page to find a quality web page. 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. 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. A Hybrid algorithm that draws from the power of n-gram based and word based system is introduced. A WCOND-Mine algorithm for mining web content outliers using n-grams without a domain dictionary is proposed in where Vector space model is used for dissimilarity computation.

The mathematical approach based on set theoretical and signed approach for mining web content outliers is presented in. A fuzzy clustering technique called C-Means algorithm to mine usage profiles from web log data has been introduced in. A better understanding of Arabic text classification techniques is achieved in. The importance of using suitable measures and methods to evaluate the performance of Web document classification is explained in. IoanDzitac et al. proposes the structure into two sections where, 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.
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 19. 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 20. 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 21. 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 22 but still it cannot find duplicates among multiple web pages. A traditional weighting technique TF.IDF (Term Frequency ‘Inverse Document Frequency) from Information Retrieval which is commonly used in text mining is explained in 23,24. The Linear Correlation based Method to Detect and Remove Redundant Web Document is suggested in 25. 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 page before presenting the page to the user. Page Rank is a numeric value that represents how important a page is on the web 26,27.

The existing web mining algorithms concentrates on web contents and do not consider the redundancy in the retrieved documents (outliers) 28,29. Web content outlier mining gives attention in 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 30.

Web content outlier mining not only helps to detect outliers when a web portal is hacked but also leads to the discovery of emerging business patterns and trends 31. 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.

Existing web content outliers mining algorithm focus only on applying weight age 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.

2. Internet Search Engines

There are differences in the ways various search engines work, but they all perform three basic tasks:

- Search the Internet based on important words.
- Keep an index of the words they find and where they find them.
- Allow users to look for words or combinations of words found in that index.

Elizabeth Liddy 32 explained about how the search engine works. Accordingly the document is pre-processed in three steps:

- Removal of stop words.
- Stemming.
- Tokenization.

The First pre-processing step is to delete stop words. A stop word list typically consists of those word classes known to convey little substantive meaning, such as articles (a, the), conjunctions (and, but), interjections (oh, but), prepositions (in, over), pronouns (he, it) and forms of the “to” be verb (is, are). The next step in pre-processing is Stemming. Stemming removes word suffixes which reduces the number of unique words in the index, which in turn reduces the storage space required for the index and speeds up the search process. For example, if a user asks for analyze, they may also want documents which contain analysis, analyzing, analyzer, analyzes and analyzed. Therefore, the document processor stems document terms to analy- so that documents which include various forms of analy- will have equal likely hood of being retrieved. The third step is Tokenization which 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. Then the weight may be assigned for each term in the document.

Next the input query will be parsed followed by stop word removal and stemming. Then the query will be expanded since the information they need may be expressed using synonyms, rather than the exact query terms for broader search which may end up with duplicates.
3. Architecture of Proposed System

The Architecture of the proposed system is depicted in Figure 1. First the user gives the input query.

**Figure 1.** Proposed architecture for mining web content outliers.

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. Then the extracted documents undergo the pre-processing step which consists of stop words removal, stemming, and tokenization. Pre-processing 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 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. If the term $W_k$ is present in document $D_i$ and not in $D_j$ then the rank of the term $W_k$ for the document $D_j$ will be zero. Next step is to compare all the document pairs to check for redundancy. The mathematical concept called correlation coefficient has been applied in this work to find out the redundant document. Spearman’s rank correlation coefficient which is given by the equation, equation (1).

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

Where $\rho$ is the correlation value, $d$ is given by $(x_i - y_i)$ where $x$ and $y$ are frequency of the term $i$ in document $D_p$ and $D_q$ respectively has been used. $n$ is total number of words in document $D_p$ and $D_q$. Always the $\rho$ value lies between 0 and 1. If the $\rho$ value is 1 for document $D_p$ and $D_q$ then $D_q$ is the duplicate of $D_p$. If there is no common word between the two documents $D_p$ and $D_q$ then the $\rho$ value will be 0.

4. Proposed Algorithm

**Input:** Web document.

**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 $1 \leq i \leq r$, $r$ is the number of retrieved documents.

Step 3 Pre-process the entire extracted document by removing stop words, stemming, and tokenization.

Step 4 Find the term frequency TF($W_{ik}$) for all the words $W_k$ in the given query for each document $D_i$ where $1 \leq k \leq m$, $m$ is the number of words in document $D_i$.

Step 5 Form a $n \times m$ matrix where $n$ is the number of words in given query and $m$ is the number of retrieved documents.

Step 6 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 $D_i$.

Step 7 For each document pair, perform the Spearman’s rank correlation coefficient given in Equation (1)

Step 8 If the $\rho$ value is 1 then $D_j$ is duplicate document, else $D_j$ is not a duplicate.
4.1 Explanation

Consider the table of term frequencies for 3 documents denoted \( D_1, D_2, D_3 \) and \( D_4 \). Compute the Term Frequency (TF) weights for the terms car, auto, and insurance, Charges, from each document \( D_1, D_2, D_3 \) and \( D_4 \). The sample TF value is given in Table 1.

Table 1. Sample TF values in each document

| Terms/Docs | \( D_1 \) | \( D_2 \) | \( D_3 \) | \( D_4 \) |
|------------|-----------|-----------|-----------|-----------|
| Car        | 27        | 4         | 0         | 27        |
| Auto       | 0         | 33        | 24        | 0         |
| Insurance  | 0         | 33        | 12        | 0         |
| charges    | 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. Term frequency ranking

| Terms/Docs | \( D_1 \) | \( D_2 \) | \( D_3 \) | \( D_4 \) |
|------------|-----------|-----------|-----------|-----------|
| Car        | 1         | 3         | 0         | 1         |
| Auto       | 0         | 1.5       | 1         | 0         |
| Insurance  | 0         | 1.5       | 2         | 0         |
| charges    | 2         | 0         | 0         | 2         |

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

Table 3. Correlation coefficient for all document pairs

| Terms/Docs | \( D_1 \) | \( D_2 \) | \( D_3 \) | \( D_4 \) |
|------------|-----------|-----------|-----------|-----------|
| \( D_1 \)  | -         | 0.25      | 0         | 1         |
| \( D_2 \)  | -         | -         | 0.05      | 0.25      |
| \( D_3 \)  | -         | -         | -         | 0         |
| \( D_4 \)  | -         | -         | -         | -         |

Since the \( \rho \) value of \( D_1 \) and \( D_4 \) is 1, the document \( D_4 \) is a redundant document and therefore it can be removed.

5. Experimental Result

An experimental analysis has been made 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.

Also the accuracy of the documents can be examined using measures like precision, recall, F Score and Accuracy.

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

\[
\text{Precision} = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved}}
\]

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

\[
\text{Recall} = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Relevant}}
\]

F-Score: F-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.

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

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

\[
\text{Accuracy} = \frac{\text{Relevant}}{\text{Total Documents}}
\]

ErrorRate = \frac{\text{Irrelevant}}{\text{Total Documents}}

6. Conclusion

A mathematical approach based on correlation method is applied to detect and eliminate redundant document. The strength of this algorithm and key feature is that the results
obtained are accurate. Future work aims at experimental evaluation of web content mining in terms of reliability and relevancy and to explore other mathematical concepts for web content mining.

7. References

1. Kosla R, Blockeel H. Web mining research: a survey. ACM SIGKDD Explorations. 2000 Jun; 2(1):1–15.
2. Liu B, Chang KC-C. Editorial: special issue on web content mining. SIGKDD Explorations. 2004; 6 (2):1–4.
3. Wang C, Liu Y, Jian L, Zhang P. A utility based web content sensitivity mining approach. International Conference on Web Intelligent and Intelligent Agent Technology (WIAT), IEEE/WIC/ACM; 2008. p. 428–31.
4. li H, Wu Z, Ji X. Research on the techniques for effectively searching and retrieving information from internet. IEEE International Symposium on Electronic Commerce and Security; 2008. p. 99–102.
5. Pokorny J, Smizansky J. Page content rank: an approach to the web content mining. Proceedings of the IADIS International Conference on Applied Computing; 2005.
6. Gopalan NP, Akillandeswari J. Distributed, fault-tolerant multi-agent web mining system for scalable web search. 5th WSEAS International conference on Applied Informatics and Communications; 2005 Sep 15–7; Malta. p. 384–90.
7. Agyemang M, Barker K, Alhajj RS. Framework for mining web content outliers. ACM Symposium on Applied Computing; 2004. p. 590–94.
8. Agyemang M, Barker K, Alhajj RS. Hybrid approach to web content outlier mining without query vector. Data Warehousing and Knowledge Discovery. 2005; 3589:285–94.
9. Agyemang M, Barker K, Alhajj RS. WCOND– mine: algorithm for detecting web content outliers from web documents. IEEE Symposium on Computers and Communications; 2005 Jun 27–30. p. 885–90.
10. Agyemang M, Barker K, Alhajj RS. A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis. 2006 Dec; 10(6):521–38.
11. Huosong X, Zhaoyan F, Liuyan P. Chinese web text outlier mining based on domain knowledge. Second WRI Global Congress on Intelligent Systems. 2010; 2:73–77.
12. Brian S, Page L. The anatomy of a large-scale hyper textual web search engine. Computer Networks. 1998; 107–17.
13. Poonkuzhali G, Thiagarajan K, Sarukesi K. Set theoretical approach for mining web content through outliers detection. International Journal on Research and Industrial Applications. 2009; 2(1):131–8.
14. Poonkuzhali G, Thiagarajan K, Sarukesi K, Uma GV. Signed approach for mining web content outliers. Proceedings of World Academy of Science, Engineering and Technology. 2009; 56:820–24.
15. Castellano G, Fanelli AM, Torsello MA. Mining usage profiles from access data using fuzzy clustering. 6th WSEAS International Conference on Simulation, Modelling and Optimization; 2006 Sep 22–24; Lisbon, Portugal. p. 157–160.
16. Zubi ZS. Using some web content mining techniques for arabic text classification. Proceedings of the 8th WSEAS International Conference on Data Networks, Communications, Computers; 2009. p. 73–84.
17. Pop I. Web document classification and its performance evaluation. 9th WSEAS International Conference on Mathematical Methods, Computational (EC08); Bulgaria; 2008.
18. Dzitac I, Moisil I. Advanced AI techniques for web mining. Proceedings of the 10th WSEAS International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems; 2008. p. 343–6.
19. Di Luca GA, Massimiliano, Fasolina AR. An approach to identify duplicated web pages. Proceedings of the 28th Annual International Computer Software and Applications Conference; IEEE computer Society press; 2002. p. 481–6.
20. Wang M-Y, Liu D-S. The research of web page de-duplication based on web pages re-shipment statement. First International Workshop on Database Technology and Applications; 2009. p. 271–74.
21. Weng Y, Li L, Zhong Y. Semantic keywords-based duplicated web pages removing. International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE ’08); 2008 Oct 19–22; Beijing. p. 1–7.
22. Han Z, Mo Q, Jianzhi L. Effectively and efficiently detect web page duplication. IEEE Fourth International Conference on Digital Information Management (ICDIM); 2009 Nov 1–4; Ann Arbor, MI. p. 1–6.
23. Zulkifeli WWR, Mustapha N, Mustapha A. Classic term weighting technique for mining web content outliers. International Conference on Computational Techniques and Artificial Intelligence (ICCTAI’2012); Malaysia. 2012.
24. Salton G. Automatic text processing: the transformation, analysis and retrieval of information by computer. Addison-Wesley Editors; 1988.
25. Poonkuzhali G, Kishore Kumar M, Kripakeshav R, Sudhakar P, Sarukesi K. Correlation based method to detect and remove redundant web document. Advanced Materials Research. 2011; 543–6.
26. Poonkuzhali G, Uma GV, Sarukesi K. Detection and removal of redundant web content through rectangular and signed approach. International Journal of Engineering Science and Technology. 2010; 2(9):4026–32.
27. Poonkuzhali G, Thiagarajan K, Sarukesi K. Elimination of redundant links in web pages-mathematical approach. World Academy of Science, Engineering and Technology. 2009; 3(4):441–4.
28. Agyemang M, Barker K, Alhajj RS. Mining web content outliers using structure oriented weighting techniques and N-grams. Proceedings of ACM SAC; New Mexico. 2005. p. 482–7.
29. Liddy E. How a search engine works. Searcher: The Magazine for Database Professionals. 2001; 9(5):38.
30. Sathyam Bama S, Ahmed MS, Saravanan A. A Mathematical Approach for improving the performance of the search engine through web content mining. Journal of Theoretical and Applied Information Technology. 2014 Feb; 60(2):343–0.