ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 154-158
doi:10.18178/wcse.2019.03.026

Mining Web Content Outliers for Improving the Quality of Search Results by using Mathematical Approaches

Thinzar Tun[1+] and Khin Mo Mo Tun[2 +]

[1] University of Information Technology, Yangon, Myanmar

[2] Faculty of Computing Department, University of Information Technology, Yangon, Myanmar

Abstract. The main task of Web mining is to provide users for retrieving relevant information from the web effectively and efficiently. The unnecessary irrelevant duplicated web pages on searching information from web affect the low quality of search results and increase indexing space and time complexity. It becomes a challenging task to provide high quality and effective search result to retrieve information. Web content outlier mining focus on mining outliers such as irrelevant and redundant pages from other the web pages under the same categories. A mathematical approach, Statistical Correlation Coefficient Approach with Term Frequency Inverse Document Frequency (TF.IDF) technique and domain dictionary is used to remove the irrelevant documents. And Kendall's Tau rank correlation coefficient is used to remove the redundant web documents and to retrieve ranked unique web documents. The results from proposed method gives F1-measures and accuracy higher than existing methods.

Keywords: web content outliers, TF.IDF, Statistical correlation coefficient, Kendall's Tau rank correlation

1. Introduction

Web mining is the application of data mining technique which is an unstructured or semi-structured data and it automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web. It is very difficult to find the relevant information from the web without containing irrelevant and redundant information accurately, quickly and easily becomes a challenge in web mining [1], [2]. Web content mining is the process of mining useful information from the contents of web pages and web documents, which are mostly text, images, audio and video. Web content mining is semi-structured and unstructured nature of the web [3]. Web content mining can directly mine the content of documents and improve on the content search of other tools like search engine [4].

Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance, and intrusion detection [5]. Traditional outlier mining has received a tremendous attention on finding rare and exceptional patterns from numeric datasets. However, web outlier mining targeting web datasets has received very little attention in the mining community. Web outliers are web data that shows significantly different characteristics than other data taken from same category. Web content outliers mining concentrates on finding outliers such as noise, irrelevant and redundant pages from web documents [6].

2. Related Work

An n-gram based method with domain dictionary and without domain dictionary in [7] and [8] to mine web content outliers. The results show that finding outliers with high order n-grams (5-grams) perform better

+ Corresponding author. Tel.: + 95943051653, +959792991473; fax: +951 – 9664254
E-mail address: [1]thinzartun@uit.edu.mm, @thinzartunaug89@gmail.com, [2]khinmomotun@uit.edu.mm

than lower order n-grams. In n-grams, the fixed lengths concept helps in memory utilization and supports partial matching of strings which is good for outlier detection. But n-gram based systems become slowly for very large datasets because of huge number of n-gram vectors generated during mining web content outliers.

In paper [9] and [10], the authors use TF.IDF with domain dictionary based on full word profile. The word-based techniques just maintain the size of the words. The organized domain dictionary ensured that the memory space search time and runtime for checking the relevancy of the web documents gets reduced. A traditional weighting technique TF.IDF is only compatible to use in detection web outliers; it even returns better results than previous works. But it cannot remove redundant web pages if they exist.

The proportionate Approach is used to mine web content outliers in [6]. Although this method can handle unstructured and structured documents, they provide less accurate results. The author G. Poonkuzhali et al. proposed linear correlation approach is used to eliminate redundant documents [11]. The author S. Sathya Bama et al. used Normalized Term Frequency with statistical correlation coefficient approach to discard irrelevant and redundant documents for improving the performance of the search engine [12]. And the Spearman correlation coefficient method is used to remove redundant documents [13]. A novel analytical approach Kendall's Tau correlation analysis is used for identifying redundant documents from web documents [14]. To enhance to improve the accuracy and F1-measure, we use a statistical correlation coefficient approach based on TF.IDF with domain dictionary to remove irrelevant documents and use Kendall's Tau correlation approach to remove duplicated documents.

3. Architecture of Web Content Outliers Mining

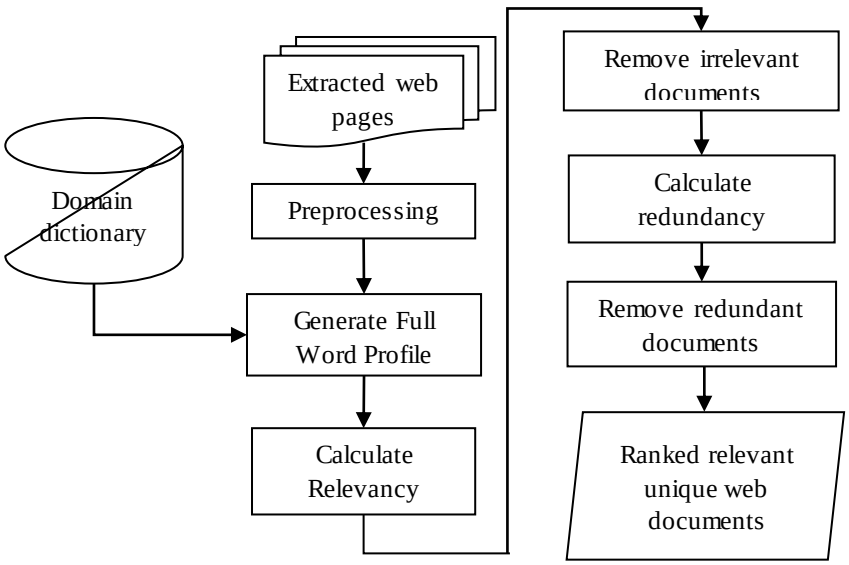

Fig. 1: System Architecture

3.1. Web Pages Extraction

Fig. 1 show the architecture of web content outlier mining. Firstly, based on the keywords extracted from user query, the search engine examines its index and extracts a list of web pages according to its category. Most of the documents retrieved from search engine might not be relevant to the query that the user enters.

3.2. Preprocessing

The preprocessing consists of stop words elimination, stemming and tokenization. Preprocessing step is used to format all input documents. Stop words list typically consists of those word classes known to convey little substantive meaning. Removing stop words reduces the indexing space and increases the efficiency and effectiveness of web search. Stemming is often used to normalize the morphological variants of the same base word. Stemming removes word suffixes which reduce the number of unique words in the index by reducing the storage space required for the index and speeds up the search process. Tokenization is the process of breaking a stream of text into words or other meaningful elements called tokens.

3.3. Full Word Profile Generation

At this stage, the domain dictionary has been indexed based on the length of the word. The full word profile for the document is generated in matrix form (i.e., $W_{1,4}$ represents 4^{th} word in 1^{st} page). Then the j^{th} word from i^{th} page is taken and its length is calculated (i.e., $| W_{ji} |$) and depending on the number of characters, the respective index on domain dictionary is searched [10].

3.4. Computation of Relevancy

The relevancy is only computed the words that exists in the dictionary. Because the word that exists in domain dictionary is more relevant to user category and it shows the power of document. Firstly, find the term frequency (TF) weighting by using Maximum Frequency Normalization for all terms in the documents as in eq. (1) because the relative frequency is suitable when the document length varies.

$$\text{TF} = \frac{TF(W_{ik})}{Max(TF(W_{ik}))} \quad \text{eq. (1)}$$

Where $TF(W_{ik})$ is the frequency of terms in the document D_i and Max $(TF(W_{ik}))$ is the maximum frequency of a word in a document D_i. The weakness in TF scheme is that it does not consider the situation where a term appears in many documents of the collection. Such a term may not be discriminative. So inverse document frequency (IDF) is calculated as in eq. (2).

$$\text{IDF} = \log \frac{N}{k} \quad \text{eq. (2)}$$

where N is the total number of documents and k is the number of documents with the word W_k appears. Next, find TF.IDF value of each word W_k in the document D_i and TF.IDF value of user given query below:

$$\text{TF.IDF} = \text{TF} * \text{IDF} \quad \text{eq. (3)}$$

The statistical correlation coefficient (CC) have to be computed for each document with given query for relevancy based on the below eq. (4).

$$\text{CC} = 1 - \frac{\sum_i D_i}{\sum_i Max(x_i, y_i)} \quad \text{eq. (4)}$$

Where D_i is given by $|x_i - y_i|$ where x_i and y_i are TF.IDF value of the term i in document D_1 and D_2 respectively. Always the CC value lies between 0 and 1. If the CC value is greater than user threshold for document di and given query, then D_i is the relevant document. Else the retrieved document is not relevant and it can be eliminated. And according to the similarity value the documents are ranked in ascending order.

3.5. Computation of Redundancy

Kendall's Tau rank correlation coefficient is applied to find out the redundant document owing to its advantages. The distribution of Kendall's Tau has better statistical properties. Also it is very insensitive to errors due to which the correlation value will be accurate with smaller size. Since the correlation is applied only for the terms between the documents, the Kendall's Tau correlation has been employed in this work. The rank of each term is compared in the document pair based on which the concordant and discordant pairs count is calculated. The number of larger ranks below a certain rank is concordant pair count and the number of smaller ranks below a certain rank is discordant pair count [14]. The Kendall τ coefficient is below:

$$\tau = \frac{(number\ of\ concordant\ pair) - (number\ of\ discordant\ pairr)}{N(N-1)/2} \quad \text{eq.(5)}$$

Where N is the number of common terms in the documents D_i and D_j. The number of concordant pairs and the number of discordant pairs between the documents D_i and D_j is computed. The τ value always lies in the middle of 0 and 1. If the τ value for the documents D_i and D_j is 1, then Dj is the redundant of D_i and remove it. Else the documents are not redundant if the value τ is 0. Finally rank the documents according to sum of Kendall's Tau correlation value for each document.

4. Web Content Outliers Mining Algorithm

Input: Web documents

Output: Ranked relevant unique web documents after removing web content outliers

1. Input the user query to Search Engine and preprocess that query
2. Retrieve web documents D_i related to that given query where 1 <= i <= r, r is number of retrieved

documents

3. Preprocess the entire extracted documents
4. Generate full word profile and organized domain dictionary

//Relevancy Computation

5. Calculate TF.IDF value for all the words W_k in the document D_i and terms in the user query given in eq. (3) where 1<= k <= n, n is the number of words in document D_i
6. Calculate Statistical Coefficient Correlation given in eq. (4) for each document with the query
7. If the value is greater than user threshold for each document D_i and query then Di is the relevant document, else it can be discarded that is irrelevant

//Redundancy Computation

8. Find term frequency TF(W_k) for all the words W_k where $1 \leq k \leq n$, n is the number of words in documents D_i and D_j
9. Find the term frequency ranking TFR(W_k) to each words W_k in the document D_i and D_j where 1<=k <=n. n is the number of common words in document Di and Dj
10. Find the number of concordant pairs (nc) and discordant pairs (nd) for each term.
11. Calculate Kendall's Tau correlation value for each document pair given in eq. (5). If the τ value is 1, Dj is duplicate document, else Dj is not redundant and a unique document
12. Computed correlation for all the possible document pairs
13. Find the sum of Kendall's Tau correlation value for each document and assign the ranks accordingly

5. Experimental Results

The case study has been made with the dataset that include 150 web pages extracted from the "Web Content Mining" domain in search engine. The statistical correlation coefficient with TF.IDF is used to eliminate irrelevant documents between set of retrieved documents and given query. If the correlation value is greater than user threshold, the documents are irrelevant and remove it. Kendall's Tau correlation analysis is calculated for each document pairs from relevant documents. the document having coefficient value 1 is defined as duplicated documents and eliminate it. It shows that the proposed method gives high F-measure and accuracy compared with the existing methods.

Precision is the percentage of retrieved documents that are in fact relevant to the desired category. Recall is the fraction of the relevant documents that are successfully retrieved. F1-Measure is a harmonic mean of precision and recall. F1-Measure reaches its best value at 1 and worst value at 0. Accuracy is the measure which matches the actual value of the quantity being measured. The F1-Measure and accuracy results are shown in Fig. 2:

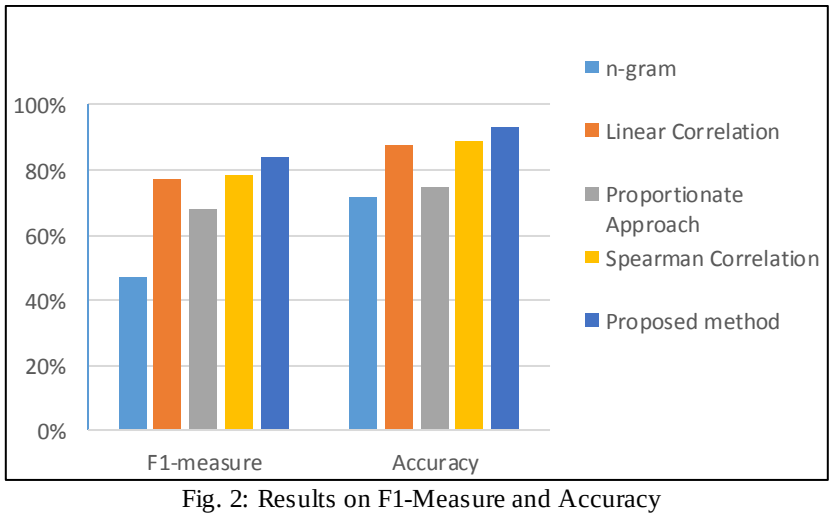

Fig. 2: Results on F1-Measure and Accuracy

6. Conclusion

The information on the web are huge, heterogeneous and unstructured. Retrieving information on web, the unnecessary and duplicated information called web content outliers may be contained. In the proposed system, Statistical Correlation Coefficient Approach with Term Frequency Inverse Document Frequency (TF.IDF) technique and domain dictionary is used to remove the unnecessary irrelevant web documents. And Kendall's Tau rank correlation is used to calculate the correlation between the each of document pairs to eliminate redundant web documents.

7. References

[1] P. Patil. Application for Data Mining and Web Data Mining Challenges, International Journal of Computer Science and Mobile Computing, March 2017.

[2] S. Vijiyarani, E. Suganya. Research issues in web mining, International Journal of Computer-Aided Technologies (IJCA) Vol.2, No.3, July 2015.

[3] A. kumar, R. K. Singh. A Study on Web Content Mining, International Journal of Engineering and Computer Science ISSN:2319-7242, 1 Jan 2017.

[4] D. Javadiya, R.Patel. Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT), Dec 2012.

[5] J. Han, M. Kamber, J. Pei. Data Mining Concepts and Techniques, Third Edition, 2012.

[6] P. Gnanasambandan, S. Poonkuzhali. Proportionate Approach for Retrieving Relevant Web Documents by using Outlier Detection Method, International Journal of Pure and Applied Mathematics, Volume 118, No.18, 2018.

[7] 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.

[8] M. Agyemang, K. Barker, and R.S. Alhajj. WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents, Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC), 2005.

[9] E. Sateesh, M. L. Prasanthi. Classic Outlier Detection from Web Clusters using Dissimilarity Measure, India Journal of Research, ISSN - 2250-1991, March 2013.

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

[11] G. Poonkuzhali, K. Sarukesi, and G.V. Uma. Web content outlier mining through mathematical approach and trust rating, 10th WSEAS International Conference on Applied Computer and Applied Computational Science, 2011.

[12] S. Sathya Bama, M.S. Irfan Ahmed. A. Saravanan. A Mathematical Approach for Improving the Performance of the Search Engine Through Web Content Mining, Journal Theoretical and Applied Information Technology, 20th February 2014.

[13] S. Vijiyarani, E. Suganya. Research issues in web mining, International Journal of Computer-Aided Technologies (IJCA) Vol.2, No.3, July 2015.

[14] R.L. Raheemaa Khan, M.S. Irfan Ahmed, A. M. Riyad. A Novel Analytical Approach for Identifying Outliers from Web Documents, International Journal of Applied Engineering Research, Volume 12, Number 22, 2017.