Mining Web Content Outliers by using Term Weighting Technique and Rank Correlation Coefficient Approach

Thinzar Tun, Khin Mo Mo Tun
University of Information Technology, Yangon, Myanmar
thinzartun@uit.edu.mm, khinmomotun@uit.edu.mm

Abstract

In the Internet area, World Wide Web (www) involves with voluminous amount of information with more redundant and irrelevant web pages. Outliers are the data that differ significantly from the rest of data. Web content mining is a subarea under web mining that mines required and useful knowledge or information from web page content. Web content outlier mining concentrates on finding outliers such as irrelevant and redundant pages from the web pages. Webs contain unstructured and semi-structured documents, so algorithms for web content mining are needed to handle both unstructured and semi structured documents. The proposed system based on big web data. The objective of proposed system is to obtain higher accurate result. In this proposal, Term Frequency Inverse Document Frequency (TF.IDF) technique based on full word matching with domain dictionary is used to remove the irrelevant documents from the unstructured web documents based on user's input query. Removal of outliers (irrelevant and redundant contents) from webs not only leads to reduction in indexing space and time complexity, but also improves the accuracy of search results. The documents that have very little similarity words from the user's input query are assumed as the web outliers. And then a mathematical approach called Spearman's rank correlation coefficient is used to remove the redundant web documents and to retrieve ranked relevant web documents.

Keywords- outliers, web content mining, term frequency, correlation coefficient

1. Introduction

With the exponential growth of information available on the internet, updating incoming data and retrieving relevant information from the web quickly and efficiently is a growing concern. Most of the web search engines typically employ conventional information retrieval and data mining techniques to discover automatically useful and previously unknown information from web. With the enormous growth on the web, users get easily lost in the rich hyper structure. In addition, as most of the data in the web is unstructured, and contains a mix of text, video, audio etc. There is a need to mine information to cater to the specific needs of the users. Web mining is an emerging research area focused on resolving these problems [1].

Web mining is the application of data mining techniques to automatically discover useful and previously unknown information from the web documents. Web Mining has adapted techniques from the field of data mining, databases and information retrieval. In general, web mining tasks can be classified into three major categories: web structure mining, web usage mining and web content mining. Web structure mining is the discovery of interesting patterns from the hyperlink structure of the web. Web usage mining mines secondary information extracted from user interactions with the web while surfing. Web content mining aims to extract useful information from the web pages based on their contents. So similar pages can be grouped together to enhance performance. web content mining aim at summarizing information on web pages to facilitate efficient and effective information retrieval. [5].

Outliers are observations that deviate so much from other observations to arouse suspicions that they might have been generated using a different mechanism. Outliers may also reflect the true properties of data from rare and interesting events which may contain more valuable information than normal data. Outlier mining is dedicated to finding data objects which differ significantly from the rest of data. Traditional outlier mining techniques can easily detect outliers that present in numeric datasets, but it becomes extremely difficult to detect outliers which are in web dataset. Web outliers are data that present in web which has different characteristics from the web data taken from the same category. Different contents of the web pages from the category in which they were taken constitute web content outliers. Web content outliers mining concentrates on discovering outliers from the web contents of a web page [3].

2. Theoretical Background

The n-gram based and word based techniques are useable in the preprocessing part of mining web content outlier. Word based systems applies different techniques

ISBN 978-99971-0-381-9 © 2017 ICAIT

than the n-gram based systems. Besides applying full word matching, the domain dictionary was indexed based on the length of word in order to enhance term searching quality. The word based technique just maintains the size of the words. Although the words are in variable length, the efficiency of word based web content outlier mining can be increased by indexing the words in two dimensional format (i, j) and indexing the domain dictionary based on length of the word. The organized domain dictionary ensured that the memory space, search time and run time for checking the relevancy of the web documents gets reduced [8].

The Term Frequency. Inverse Document Frequency (TF.IDF) is a weighting method often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word for a document in a collection or corpus. The TF.IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Variations of the TF.IDF weighting method are often used by search engines as a central tool in scoring and ranking a document's relevance given a user's input query [9].

In statistics, Spearman's rank correlation is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation between two variables will be high when observation has a similar. Spearman's coefficient is appropriate for both continuous and discrete variables including ordinal variables [10].

3. Related Works

The same authors use an n-gram method with domain dictionary and without domain dictionary in [4] and [5] to determine the similarity of strings and expand it to include pages containing similar strings. The experimental results show that finding outliers with high order n-grams (5-grams) perform better than lower order n-grams. The existing approach using WCOND Mine algorithm based on n-grams that works only for structured documents. The n-gram based systems become slowly for very large datasets because of the huge number of n-gram vectors generated during mining web content outliers. The word-based techniques just maintain the size of the words. Although the words are in variable length, the efficiency of word based web content outlier mining can be increased by indexing the words in two-dimensional format (i,j) and indexing the domain dictionary based on length of the word. The organized domain dictionary ensured that the memory space search time and run time for checking the relevancy of the web documents gets reduced. The n-gram based system takes a longer time to complete a task than the word-based systems even though the size of data is not too large. A traditional weighting technique TF.IDF (Term Frequency * Inverse Document Frequency) from information retrieval 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 [8].

The author S.Poonkuzhali proposed a signed with weight technique based on full word matching for structured and unstructured documents to retrieve relevant document and linear correlation is used to remove duplicates [2]. A mathematical approach called Spearman's correlation coefficient is used to calculate the correlation between the document pairs to remove redundant web pages. This method depends on the term frequency of common words between document pairs that is ranked based on the frequency value. This method gives better performance than linear correlation and ranking correlation [6]. In the proposed system, Term Frequency Inverse Document Frequency (TF.IDF) technique based on full word matching with domain dictionary is used to mine and remove irrelevant web pages and Spearman's correlation coefficient is applied to eliminate redundant web pages.

4. Architecture Design

4.1. Extracted web pages

The document extraction is the process of retrieving the desired pages belonging to the category of interest. The documents are retrieved by search Engine based on the user's input query. Most of retrieved documents may or may not relevant to the user query [7].

4.2. Preprocessing

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 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).

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, phrases, symbols, or other meaningful elements called tokens. A token is a string of characters,

categorized according to the rules as a symbol. The list of tokens becomes input for further processing. [7].

4.3. Generate Full Word Profile

The filtered datasets from preprocessing stage are then used to generate the full word profile. At this time, 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 4th word in 1st page). Then the jth word from ith 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. If the words exist in both sides, it will be flagged as 1, otherwise 0 will be returned. Then the word frequency will be counted. The full word profile generated by indexing all word with two-dimensional format (i, j) where 'i' represent web pages, 'j' represent words and every word attached with word frequency, word length and the binary number which mentioned either it exists in domain dictionary or not [1].

4.4. Compute Relevancy with TF.IDF

In the weighting computation, a classic term weighting technique, TF.IDF from Information Retrieval (IR) was adopted to evaluate the representativeness of terms in the web content. The dissimilarity measure computed to determine the difference among pages within the same category. The Maximum Frequency Normalization applied to Term Frequency (TF) weighting because when the document length varies, the relative frequency is preferred. Since term frequency alone may not have the discriminating power to pick up all relevant documents from other irrelevant documents, an IDF (Inverse Document Frequency) factor which takes the collection distribution into account has been proposed to help to improve the performance of IR.

The dissimilarity measure will only compute the words that exist in the dictionary because the formula returns only a binary value. Then the words that did not exist in the domain dictionary will not be computed. The reason is the word that exists in the dictionary is more relevant to the domain category and it represents the power of the document. The outliers come out with the lowest frequency of word that exists in the dictionary and there will be only a few words that exist in the domain dictionary. Therefore, the dissimilarity measures will return a higher dissimilarity value than other web pages [8].

The dissimilarity equation is below:

$$DM_i = \frac{\sum_{i,j}[(0.5 + \frac{0.5 * f(t_j, e_i)}{MaxFreq(d_i)})(log_{10}\frac{N}{k})]}{e_i}$$

where f (t_j, e_i) denotes the frequency of term t_j present in the document d_i in the domain dictionary, while MaxFreq(d_i) determine maximum frequency of a word in a document, N is the total number of documents and k is the number of documents with term t_j appears.

4.5. Determination irrelevant documents

The output from the dissimilarity measure was ranked to determine outliers or irrelevant documents. The documents at the top will have high dissimilarity measure deviates more from the category of user interest. Also, the documents at the bottom will have less dissimilarity measure which is more relevant to the category of interest. So, the top 'n' documents that have high dissimilarity measure are declared as outliers based on threshold value.

4.6. Compute redundancy with Spearman's correlation coefficient

In Spearman's correlation coefficient method, frequency of all the terms in the document 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, similarity for other terms. If the term W_k is present in document Di and not in Dj the rank of the term Wk for the document Dj 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 documents. Spearman's rank correlation coefficient equation is below:

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

Where ρ is correlation value, d is given by (x_i–y_i) where x_i and y_i 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 ρ value lies between 0 and 1. If the ρ value is 1 for document D_p and D_q then D_q is the redundant of D_p. If there is no common word between the two documents D_p and D_q then the ρ value will be 0 [6]. And then remove all redundant web documents.

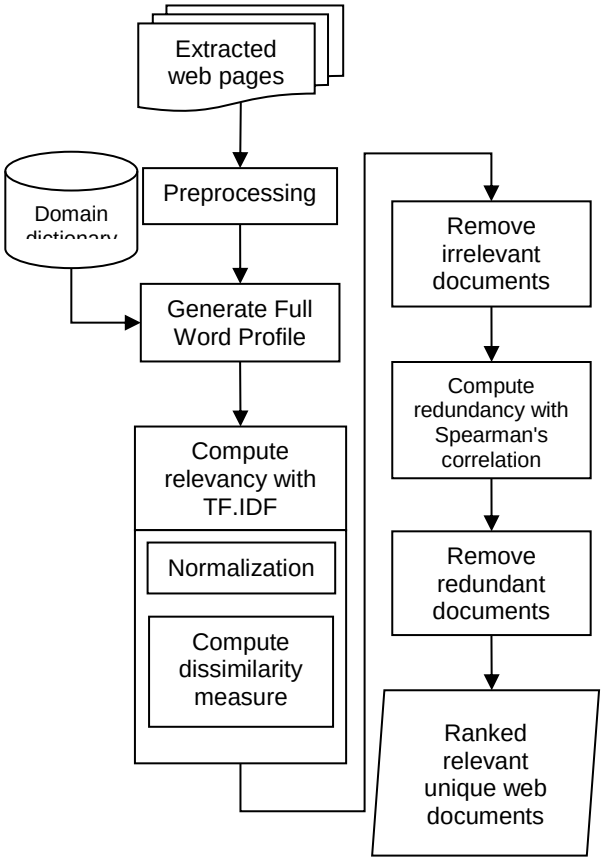

Figure 1. Architecture design of proposed system

5. Proposed Algorithm and Experimental Results

Input: Domain Dictionary and Web Documents d_i

Output: Ranked relevant unique web documents

1. Extract the set of documents
2. Preprocess the entire extracted documents by removing stop words, stemming and tokenization
3. Generate full word profile
4. Generate organized domain dictionary

//relevancy computation with TF.IDF
5. For (int i=0; i<NoOfDoc; i++) {
6. For (int j=1; j<=NoOfWords; j++) {
7. If (j exists in the domain dictionary) {
8. Compute dissimilarity measure (DM_i)

$$DM_i = \frac{\sum_{i,j}[(0.5 + \frac{0.5 * f(t_i, e_i)}{MaxFreq(d_i)})(log_{10}\frac{N}{k})]}{e_i}$$

9.}}

10. DM_i= DM_i/ number of words in the document that exist in the domain dictionary.
11. Rank the result of DM_i
12. Determine irrelevant documents and remove it

// redundant computation with Spearman correlation coefficient method

13. Find the term frequency TF (W_{ik}) for all the words W_k in the given query for each document d_i where $1 \le k \le m$, m is the number of words in document d_i.
14. Form a n x m matrix where n is the number of words in given query and m is the number of retrieved documents.
15. Assign the term frequency ranking TFR (W_{ik}) to each words W_k in document di where $1 \le k \le m$. m is the number of words in document d_j.
16. Assign the term frequency ranking TFR (W_{ik}) to each words W_k in document di where $1 \le k \le m$. m is the number of words in document d_j.
17. For each document pair, perform the Spearman's rank correlation coefficient

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

18. If the ρ value is 1 then d_j is redundant document, else d_j is not a redundant.
19. Remove redundant documents

An analysis has been made with the proposed system and the existing methods. A case study has been tested with the dataset that consists 200 web pages from Science medical folder provided by the 20 Newsgroup datasets. There is no benchmark data for testing web content outliers, so embedded motive is the only way to know if the outliers returned are actually real outliers (irrelevant and redundant contents). The outliers usually constitute less than 10% of the entire dataset [8]. So, 20 web pages from the Course folder of University Cornell, provided by World Wide Knowledge Base (WEBKB) to detect outliers or irrelevant documents. Then documents are retrieved and processed with TF.IDF method to remove irrelevant documents. The results are ranked and top 20 web documents are defined as outliers or irrelevant documents. And another web documents are declared as relevant documents.

Next, Spearman correlation coefficient method is calculated for each of document pairs from relevant retrieved documents. Finally, the document having coefficient value 1 is defined as redundant document and removed it. It shows that the proposed method generates high F-measure and accuracy compared with the existing methods. Precision is the faction of retrieved documents that are relevant to the query. Recall is the fraction of the relevant documents that are

successfully retrieved. F1-Measure is the 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 below:

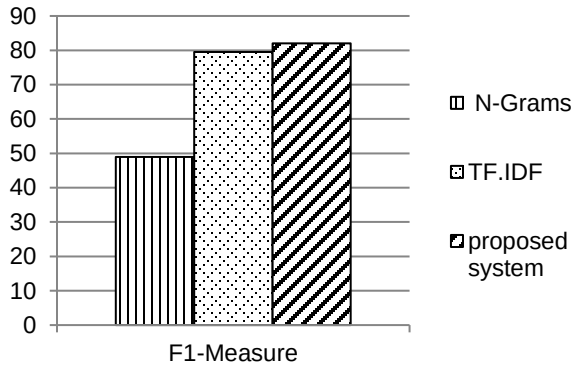

Figure 2. Results on F1-Measure

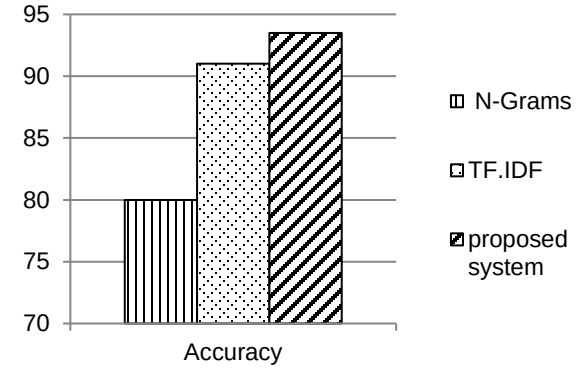

Figure 3. Results on accuracy

6. Conclusion

The massive growth of internet and World Wide Web encourages developing the automated tools to retrieve relevant sources quickly without duplicates. The key feature of the proposed system is to improve accuracy result. In the proposed system, the traditional term weighting technique TF.IDF based on full word matching with domain dictionary is used to remove irrelevant web documents. And Spearman's correlation coefficient method is used to calculate the correlation between the document pairs to eliminate redundant web documents.

7. References

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

[2] 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 (ACACOS '11), 2011.

[3] K.Sarukesi, P.Sudhakar, S. Poonkuzhali, "Signed-With-Weight Technique for Mining Web Content Outliers", Special Issue of International Journal of Computer Applications (0975 – 8887) the International Conference on Communication, Computing and Information Technology (ICCCMIT) 2012.

[4] 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.

[5] 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.

[6] S. SATHYA BAMA, M.S. IRFAN AHMED, A. SARAVANAN, "A Mathematical Approach for Mining Web Content Outliers using Term Frequency Ranking", Indian Journal of Science and Technology, Vol 8(14).

[7] 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, Vol.60, No2.

[8] 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.

[9] https://en.wikipedia.org/wiki/Tf–idf

[10] https://en.wikipedia.org/wiki/Spearman's_rank_ *correlation* coefficient