Journal of Theoretical and Applied Information Technology
20th February 2014. Vol. 60 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MATHEMATICAL APPROACH FOR IMPROVING THE PERFORMANCE OF THE SEARCH ENGINE THROUGH WEB CONTENT MINING

[1]S.SATHYA BAMA, [2]M.S.IRFAN AHMED, [3]A.SARAVANAN
[1,3] Assistant Professor, Sri Krishna College of Technology, Coimbatore, India
[2] Professor, Sri Krishna College of Engineering and Technology, Coimbatore, India
E-mail: [1]ssathya21@gmail.com , [2]msirfan@gmail.com, [3]a.saravanan21@gmail.com

ABSTRACT

Internet is a rapid growing technology that contains a vast and rich set of information stored on the web. To retrieve, share and to process all these information from the web, a tool has been created called search engine which plays an essential role for the web users. On searching the information from web servers through search engines, many irrelevant and redundant documents containing the required information will be retrieved and presented to the users. But this irrelevant and replicated information affects the performance of the search engine by wasting the user's time by surfing the uninterested documents which is inefficient to the web users. So, to make the search effective, and to improve the performance of the search, many researchers turned their attention towards Web mining since web is used in almost all areas. Web content mining is a subarea under web mining that mines required and useful knowledge or information from the web content. Most existing algorithms focus on applying weightage only to the common terms in the documents by which the accuracy gets consecutively reduced. The performance of a search engine can be improved through this proposed approach based on term frequency ranking to mine the web contents.

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

1. INTRODUCTION

In the present day, almost all disciplines use a most interactive and intermediate standard to accumulate and access huge quantity of information called World Wide Web. Due to this increase in the usage of internet, the size of the web that contains the data and information is increasing predominantly every day. Thus, managing and retrieving those data has become a difficult task for the web users. To overcome this tedious task, a tool or a platform to search and retrieve the required information has been identified and it is called as search engine. Now, the use of search engine has extremely increased among humans due to the vast improvement in technology where the required information can be accessed by anybody at any time and from anywhere.

For accessing the required information from the web, the user gives a query to the search engine. The search engine then process the given query and access the information from the web server and finally presents the documents that matches the terms in user's query. But, the retrieved document containing the information for the given query will not be always effective; possibly it contains irrelevant and redundant documents. So, presenting only relevant information without any redundancy from the web has become complex and challenging task for those search engines due to raise in the increasing amount of information stored in the web. To answer all these issues web mining has become an important research area. All search engines utilize different methods to discover useful knowledge from the web, but most of it uses conventional and traditional Information Retrieval algorithms along with data mining techniques. Applying data mining techniques with little amendments that better suits web data to mine the web is termed as Web Mining.

According to research, Web Mining is broadly classified in to three categories. They are Web Usage Mining, Web Structure Mining and Web Content Mining [1], [2], [3]. According to them, Web usage mining is the process of mining or discovering the knowledge about user access patterns from Web usage logs which is helpful to make future decisions. And, Web structure mining is the process of determining useful knowledge from the structure of the web page using hyperlinks which helps to explore the structure of web sites. Also, Web content mining is the process of mining, extracting and integrating useful data, information and knowledge from Web page contents by applying some conventional data mining techniques [4]. Many researchers carried out their research activities in all these three areas. Recently, there is a rapid growth of research activities in the Web Content Mining [5].

Web content outlier mining is a concept of finding outliers such as noise, irrelevant and redundant pages from the web documents that are presented to the users as a result of the successful search through search engines. By removing these outliers, unique documents can be retrieved by eliminating uninteresting documents obtained by mining the Web Content [6]. Web content outlier mining is not only helpful to detect outliers when a web portal is hacked but also may lead to the discovery of emerging business patterns and trends [7]. The traditional search engine is more complex system in extracting and integrating data [8] due to which it generates thousands of responses to the user, and many of them are not relevant to the user query and does not provide best result [9].

This paper focuses on improving the performance of search engines by proposing the new architecture to find relevant information and to extract potentially useful knowledge from the web [10]. This proposed Architecture improves the performance of the search engine in such a way by removing irrelevant and redundant information which minimizes the efficiency and effectiveness of a search engine. The new correlation coefficient formula has been proposed to compute the similarity between the documents and the query terms.

2. RELATED WORK

Web mining is an emerging research area that focuses on resolving problems while accessing and updating information on the web. Web Content Mining aims to extract useful information from the web pages based on their contents [11], [12]. A new approach to Web Content Mining by using page Content Rank has been introduced [13]. The n-gram based algorithm by assuming the existence of domain dictionary for mining web content outliers using full word matching, which explores the advantages of n-gram techniques as well as HTML structure of web documents has been proposed [14].

With this an improvement has been made that describes the power of n-gram and word based system [15]. Anew algorithm called WCOND-Mine algorithm has been proposed which uses vector space model for dissimilarity computation for mining web content outliers. This approach uses n-grams without assuming the existence of a domain dictionary [16], [17]. The new technique has been proposed by integrating generalized pattern mining algorithm and clustering concepts [18]. The mathematical approach based on set theoretical and signed approach for mining web content outliers is presented in [6] [19]. A better understanding of Arabic text classification techniques is achieved in [20]. The importance of using suitable measures and methods to estimate the performance of Web document classification is explained in [21]. Ioan Dzitac et al. proposes the structure into two sections where, first method briefly discusses the various web mining tasks and the second method focuses on advanced Artificial Intelligence (AI) methods for information retrieval and web search, link analysis, opinion mining and web usage mining [22].

An algorithm based on clone detection and similarity metrics to detect duplicate pages in web sites for structured web documents has been introduced in [23]. A web page de-duplication method which extracts the information from websites and web titles to eliminate duplicated web pages based on feature codes using URL hashing has been introduced [24]. But in this method the extraction of feature codes takes much time. Copy Detection Algorithm (COPS) scheme 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. This method is cost expensive for building the inverted index of the semantic keywords [25]. 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 [26] 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 explained in [27], [28]. The Linear Correlation based method to detect and remove duplicate web document is suggested in [29]. The experiment analysis has been made and it shows that 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 [11], [30]. An enhanced page rank algorithm has been adopted in [31].

To improve the quality of URLs to be listed thereby avoiding irrelevant or low-quality ones, a vertical search engine has been proposed which focuses on page relevance to a particular domain and page contents for the search keywords [32]. All the above works on web content mining which is adopted for search engines, lack simplicity of notion and computation. In this proposed work, the statistical correlation has been applied to remove the duplicates and irrelevant pages from the retrieved web pages which is more efficient than existing algorithm in time and space.

3. SEARCH ENGINE – A SUMMARY

Search engines are special sites on the web that are intended to help people to find information that is stored on other sites. Each search engine works in different way, but ultimately they all do three basic tasks:

- They search the web based on the given significant words.
- They maintain an index of the significant words they find, and where they find them.
- They allow users to look for information containing significant words found in that index.

Now-a-days a search engine is efficient than earlier version since the count of the index has been increased from thousands to million pages. Today, a top search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day. Search engines match queries against an index that they create. The index consists of the words in each document, plus pointers to their locations within the documents. This file containing an index also called as an inverted file.

Even Elizabeth Liddy explained in detail about how the search engine works. According to her study, all the documents in the web will be preprocessed to make a common format. The First preprocessing step is to delete stop words. A stop word list typically consists of those word classes known to convey little meaning, such as articles, conjunctions, interjections, prepositions, pronouns and forms of the "to be" verb. The second step in preprocessing is Stemming. 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. For example, the word analyze that stems to produce the word analy-, since the documents which include various forms of analy- like analysis, analyzing, analyzer, analyzes, and analyzed will have equal preference of being retrieved. The third step is Tokenization which 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 query for further processing. This query will be expanded since the information they need may be expressed using synonyms, rather than the exact query terms for broader search. Then query terms will be assigned a weight and then the weighted query is searched against the inverted file for required documents which may end up with duplicates. Then the weight is assigned for each term in the document [33].

4. ARCHITECTURE OF THE PROPOSED SYSTEM

The Architecture of the proposed system is 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. Then 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 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 and normalized. Then the given query Q is converted to weighted query. The proposed correlation coefficient (CC) have to be computed for each document with weighted query for relevancy based on the below equation Eq. (1).

$$CC = 1 - \frac{\sum_i d_i}{\sum_i Max(x_i, y_i)} \quad (1)$$

Where d is given by $|x_i - y_i|$ where xi and y_i are normalized term frequency 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 d_i and Q, then d_i is the relevant document. Else the retrieved document is not relevant and so it can be removed. And according to the similarity value the documents are ranked in ascending order. 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. If the word Wk is present in document Di and not in D_j then the rank of the word W_k for the document D_j will be zero. Apply the correlation coefficient Eq(1) for each document pair. If the CC value is 1 for document d_i and d_j then d_j is the duplicate of di which can be removed.

5. PROPOSED ALGORITHM

Input: Web document.
Method: Statistical Method
Output: Extraction of relevant unique web document.

Step 1. Input the query Q to Search Engine.

Step 2. Pre-process the query by removing stop words, and stemming.

Step 3. 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 4. Pre-process the entire extracted document by removing stop words, stemming, and tokenization.

//Relevancy Computation

Step 5. 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 6. Normalize the term frequency value $NTF(W_{ik})$ to each words W_k in the document D_i by dividing the $TF(W_{ik})$ by sum of $TF(W_i)$ of all the retrieved document, where $1 \leq k \leq m$. m is the number of words in document D_j.

Step 7. Assign the weight to the terms in the query, by dividing the $Max(TF(W_{ik}))$ by sum of $TF(W_i)$ of all the retrieved document.

Step 8. Perform the Proposed Correlation Coefficient given in Eq.(1) for each document with the weighted query

Step 9. If the *CC* value is greater than user threshold for document d_i and Q then d_i is the relevant document, else it can be removed which is not relevant.

//Relevancy Computation

Step 10. Initialize i=1 and j=i+1

Step 11. Find the terms T in the documents D_i and D_j by making union for the words in the documents.

Step 12. Perform the correlation coefficient using the formula in Eq(1)

Step 13. If the *CC* value is 1 then D_j is duplicate document which can be removed.

Step 14. Increment j, and repeat from step 6 to step 9 until $j \leq r$.

Step 15. Increment i, and repeat from step 6 to step 10 until $i < r$.

5.1 Explanation

Consider the table of term frequencies for four documents denoted D_1, D_2, D_3 and D_4. Compute the Term Frequency (TF) weights for the terms correlation, content, frequency and sample from each document D_1, D_2, D_3, D_4. The sample TF value is given in Table 1.

Table 1: The sample Term Frequency values

Terms/Doc	D_1	D_2	D_3	D_4	Total Term Frequency
correlation	27	4	0	27	58
content	0	33	24	0	57
frequency	0	33	17	0	50
sample	14	5	0	14	28

Normalize the frequencies by dividing each value of the term with Total Term Frequency of the term. The Normalized Term Frequency is given in Table 2.

Table 2: The Normalized Term Frequency value

Terms/Doc	D_1	D_2	D_3	D_4
correlation	0.466	0.069	0	0.466
content	0	0.579	0.421	0
frequency	0	0.66	0.34	0
sample	0.5	0	0	0.5

Assign the weight to the terms in the query, by dividing the $Max(TF(W_{ik}))$ by sum of $TF(W_i)$ of all the retrieved document. If the query term is 'correlation' then the weight for the term can be taken as 0.466. With this compute the correlation coefficient given in Eq(1) for the query term with all the documents D_1, D_2, D_3, D_4 which is given in Table 3. The CC value is given below which represents that document D_3 is not relevant and it can be removed. Based on these CC values of D_1,

D_2, and D_4 the documents can be ranked as D_1, D_4, D_2.

Table 3. The correlation coefficient for query with all documents

Terms/Doc	D_1	D_2	D_3	D_4
Q	1	0.83	0	1

Next the duplicates can be identified by computing *CC* Value for all the document pairs. The *CC* value for Document D_1 and D_4 is 1 which implies that the document D_4 is redundant and hence it can be removed. The values are given in Table 4.

Table 4. The correlation coefficient for all document pairs

Document	D_1	D_2	D_4
D_1	-	0.032	1
D_2	-	-	0.032
D_4	-	-	-

6. EXPERIMENTAL RESULT

An experimental analysis has been made for the proposed algorithm and for the existing methods with different web page size. With this It is observed that the proposed method generates high F-measure and accuracy compared against existing methods. In this experimental analysis the redundancy computation is done based on proposed method only for the retrieved documents. The input query as "Web Mining" has been given to the search engine and the documents are retrieved and processed.

The correlation between the retrieved document and the given query is calculated to find the relevant documents. The documents are relevant if the correlation value satisfies the user threshold. Using the calculated correlation coefficient values the documents are ranked. Next, the correlation coefficient has been calculated for all document pairs to find the redundancy. One document from the document pair having coefficient value as 1 has been removed since it is redundant document. The comparison has been made with the performance of n-gram method, TF.IDF, Linear Correlation and proposed method which is shown in the Figure 3. The proposed method has been evaluated by various metrics like precision, recall, F-measure and Accuracy which is defined below.

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

$$\Pr ecision = \frac{\left|\{\text{Re}\,lavant\ Documents\} \cap \{\text{Re}\,trived\ Documents\}\right|}{\left|\{\text{Re}\,trieved\ Documents\}\right|}$$

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

$$\text{Re}\,call = \frac{\left|\{\text{Re}\,lavant\ Documents\} \cap \{\text{Re}\,trived\ Documents\}\right|}{\left|\{\text{Re}\,levant\ Documents\}\right|}$$

F-measure: 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.

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

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

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

$$ErrorRate = \frac{Irrelavanat\ \ Documents}{Total\ \ Documents}$$

Figure 2, Figure 3, Figure 4 and Figure 5 show the precision, recall and F-measure and accuracy produced by different methods.

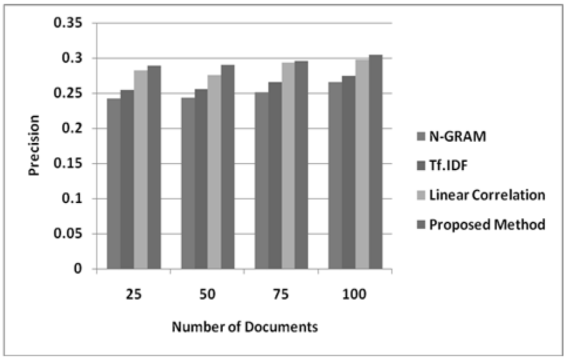

Figure 2: Comparison on precision

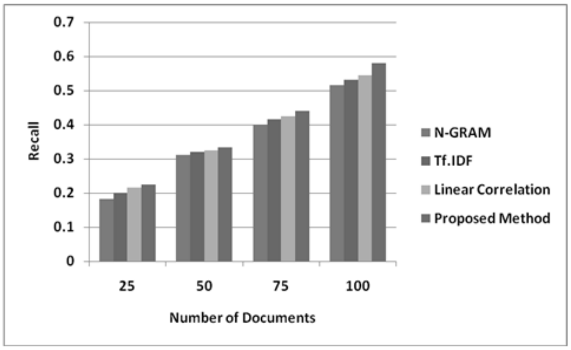

Figure 3: Comparison on recall

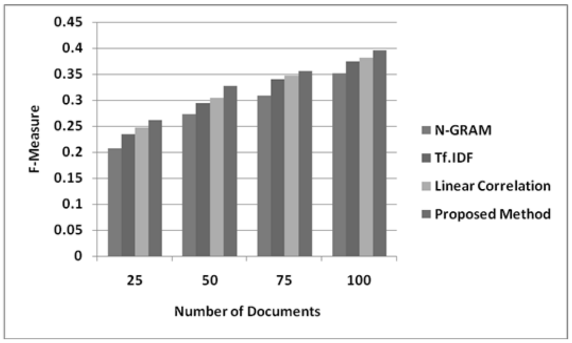

Figure 4: Comparison on F-Measure

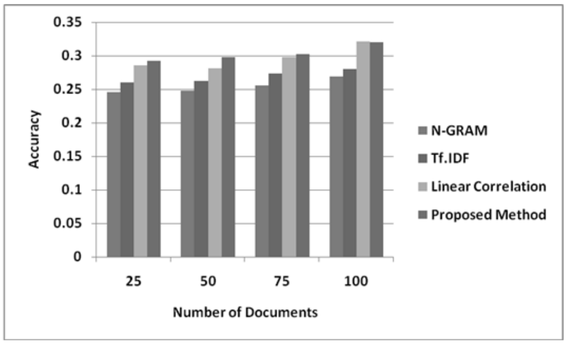

Figure 5: Comparison on accuracy

7. CONCLUSION

The enormous 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 redundant document. The strength of this algorithm and key feature is proved by the accurate results obtained from the experimental results. Future work aims at experimental evaluation of web content mining in terms of reliability and to explore other mathematical concepts for web mining to make more efficient mining in terms of space and time.

REFRENCES:

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

[2] S.Madria, S.S. Bhowmick, W.K. Ng, and E.P. Lim, "Research issues in web data mining", *Proceedings of the Conference on Data Warehousing and Knowledge Discovery*, 1999, pp. 303–319.

[3] S. Pal, V. Talwar, P. Mitra, "Web mining in soft computing framework: relevance, state of the art and future directions", *IEEE Transactions on Neural Networks*, Vol. 13, No. 5, 2002 pp. 1163–1177.

[4] S. Poomagal, T. Hamsapriya, "Cosine similarity-based PageRank calculation", *International journal of Web Science*, Vol. 1, Nos. 1/2, 2011.

[5] R. Campos, G. Dias, and C. Nunes, "WISE : Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques", *International conference on Web Intelligence, IEEE/WIC/AC*,.2006.

[6] G. Poonkuzhali, K. Thiagarajan, K. Sarukesi, G.V. Uma, "Signed approach for mining web content outliers", *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 56, 2009, pp. 820- 824.

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

[8] J. Gou, "Web Content Mining and Structured Data Extraction and Integration: An Implement of Vertical Search Engine System", *Research Report*, 2012.

[9] M. Chau, H. Chen, "Comparison of Three Vertical Search Spiders", *Computer (Journal)*, Vol 36, Issue 5, 2003, pp. 56-62.

[10] K. Pol, N. Patil, P. Shreya, C. Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data", *First International Conference on Emerging Trends in Engineering and Technology*, 2008, pp. 543–546.

[11] C. Wang, Y. Liu, L. Jian, P. Zhang, "A Utility based Web Content Sensitivity Mining Approach", *International Conference on Web Intelligent and Intelligent Agent Technology (WIIAT)*. IEEE/WIC/ACM, 2008.

[12] H. Li, Z. Wu, X. Ji, "Research on the techniques for Effectively Searching and Retrieving Information from Internet", *International Symposium on Electronic Commerce and Security*, IEEE, 2008.

[13] J. Pokorny, J. Smizansky, "Page Content Rank: An approach to the Web Content Mining", *Proceedings of the IADIS International Conference on Applied Computing*, Vol 2, 2005, pp. 22-25.

[14] M. Agyemang, K. Barker, A.S Alhajj, "Framework for Mining Web Content Outliers", *ACM Symposium on Applied Computing*, 2004, pp. 590-594.

[15] M. Agyemang, K. Barker, A.S. Alhajj, “Hybrid Approach to Web Content Outlier Mining without Query Vector”, *Springer –Berlin*, Vol. 3589, 2005.

[16] M. Agyemang, K. Barker, A.S Alhajj, “WCOND – Mine: Algorithm for detecting Web Content Outliers from Web Documents”, *IEEE Symposium on Computers and Communication*, 2005.

[17] M. Agyemang, K. Barker, A.S. Alhajj, “A comprehensive survey of numeric and symbolic outlier mining techniques”, *Intelligent Data Analysis*, Vol. 10, No (6), 2006, pp. 521-538.

[18] M. Manikandan, “Improving efficiency of textual static web content mining using clustering techniques”, *Journal of Theoretical and Applied Information Technology*, Vol. 33, No.2., 2011.

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

[20] Z.S. 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.

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

[22] I. Dzitac, I. 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.

[23] G.A.D. Lucca, Massimiliano and A.R. Fasolina, “An Approach to identify Duplicated web pages”, *proceedings of the 28th Annual International Computer Software and Applications Conference, IEEE computer Society press*, 2002.

[24] M. Wang, D. Liu, “The Research of web page De-duplication based on web pages Re-shipment Statement”, *First International Workshop on Database Technology and Applications*, 2009, pp. 271-274.

[25] Y. Weng, L. Li, Y. Zhong, “Semantic keywords-based duplicated web pages removing”, *IEEE*, 2008.

[26] Z. Han, Q. Mo, Liu, Jianzhi, “Effectively and Efficiently Detect Web Page Duplication”, *IEEE*, 2009.

[27] 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.

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

[29] G. Poonkuzhali, R. Kishore kumar, R. kripa keshav, P. Sudhakar, K. Sarukesi, “Correlation Based Method to Detect and Remove Redundant Web Document”, *Advanced Materials Research*, Vols. 171-172, 2011, pp. 543-546.

[30] S. Brian, L. Page, “The anatomy of a large-scale hyper textual Web search engine”, *Computer Networks*, 30 (1-7), 1988, pp. 107-117.

[31] S. Sathya Bama, M.S. Irfan Ahmed, A. Saravanan, “Improved PageRank Algorithm for Web Structure Mining”, *International Journal of Computer and Technology*, Volume 10, No.9, 2013, pp.1969-1976.

[32] S. Rajashree, B. Rahu, “A Vertical Search Engine – Based On Domain Classifirer”, *International Journal of Computer Science and Security*, Volume (2), Issue (4), 2008, pp.18-27.

[33] E. Liddy, “How a Search Engine Works. Searcher”. *The Magazine for Database Professionals*, Volume 9, Number 5, 2001, pp.38.

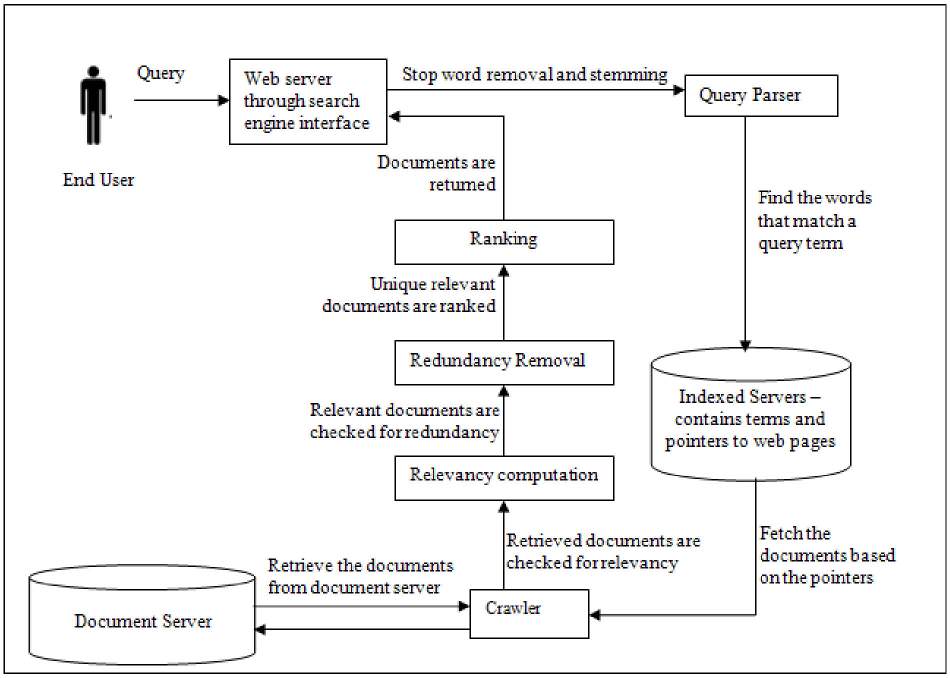
Query
Web server through search engine interface
Stop word removal and stemming
Query Parser
End User
Documents are returned
Find the words that match a query term
Ranking
Unique relevant documents are ranked
Redundancy Removal
Indexed Servers – contains terms and pointers to web pages
Relevant documents are checked for redundancy
Relevancy computation
Retrieved documents are checked for relevancy
Fetch the documents based on the pointers
Retrieve the documents from document server
Document Server
Crawler

Figure 1: Proposed architecture to improve the performance of a search engine