International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 5 (2015) pp. 13625-13635

http://www.ripublication.com

Enhancing The Search Engine Results Through Web Content Ranking

S.Sathya Bama

Department of MCA, Sri Krishna college of Technology,
Coimbatore, Tamil Nadu, INDIA
ssathya21@gmail.com

M.S.Irfan Ahmed

Department of MCA,
Sri Krishna college of Engineering and Technology, Coimbatore, Tamil Nadu, INDIA
msirfan@gmail.com

A.Saravanan

Department of MCA, Sri Krishna college of Technology,
Coimbatore, Tamil Nadu, INDIA
a.saravanan21@gmail.com

Abstract

The development of the Internet has made the rebellion in the field of information storage and retrieval. In such case, search engines play a vital role in retrieving and organizing relevant data for various purposes. But in reality, as most of the data in the web is unstructured and even structured documents contains a mixed data like text, video, audio and images, results produced by search engines are still uncertain because it returns massive amount of irrelevant and redundant results. So, there is a need to mine relevant information to satisfy the specific needs of the users by removing the outliers. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents. In this paper, a novel weighted approach based on full word matching in the organized domain dictionary with term frequency is used to mine the web documents providing the user needs. The dictionary is structured based on the number of characters in a word, searching and retrieval of documents takes less time and less space. The experimental result shows that this method ranks more than 80% of the relevant documents accurately.

Key Words: Relevant Document, Redundant Document, Weighted Approach, Web Content Mining, Web Documents, Web Document Ranking.

1. INTRODUCTION

The growth of the Web surpassed almost all expectation due to enormous amount of information in the form of structured, semi-structured and unstructured web pages containing various types of data like text, image, video, audio etc. As a result, implementing and developing powerful technique to extract relevant content from the extensive web has become a very difficult task. In most cases, Web search engines are used frequently to retrieve relevant information by giving query based on user interest as an input. These search engines normally employ conventional information retrieval and data mining techniques to discover useful and previously unknown information from web content.

But due to its huge size, normally a web query can retrieve millions of resulting web pages that contains irrelevant and redundant documents by which most of the users loss temper in navigating more number of links without getting exact information. Hence significant methods are essential to for present these large results that helps the user to access the interesting and relevant information based on their interest. These problems gave rise to the developmental work in the area of web content mining.

Generally Web content mining is one of the main category in web mining which uses the ideas and principles of data mining and knowledge discovery to mine with little amendments.

Web content mining refers to the discovery of useful information from web content that includes various types of data like text, image, audio, video, metadata and hyperlinks etc. the other interesting areas in web mining includes Web structure mining and web usage mining. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web usage mining refers to the discovery of user access patterns from web usage logs. Web content mining aims to extract/mine useful information from the web pages based on their contents. Research in web content mining incorporates resource discovery from the web, document categorization and clustering, and information extraction from web pages. Two groups of web content mining specified in [1,2,3] are those that directly mine the content of documents and those that improve on the content search of other tools like search engine.

Some of the areas of doing research in web content mining are listed below [4]:

- Structured Data Extraction
- Unstructured Text Extraction
- Web Information Integration and Schema matching
- Building Concept Hierarchies
- Segmentation and Noise Detection
- Opinion extraction

2. RELATED WORK

By taking the advantage of the HTML structure of web and n-gram technique for partial matching of strings, an n-gram based algorithm for mining web content outliers has been proposed. To reduce the processing time, the algorithm used only data captured in <Meta> and <Title> tags [5].

A utility-based Web content sensitivity mining approach has been developed by cheng wang et al. This algorithm identified a number of sensitive Web pages that traditional frequency-based methods failed to find [6].

A new algorithm for mining is proposed for web content using clustering technique and mathematical set formulae such as subset, union, intersection etc for detecting outliers. Then the outlying data is removed from the original web content to get the required web content by the user [7, 8].

An architecture called WISE has been proposed by Richard Campos, a meta-search engine that automatically builds clusters of related web pages embodying one meaning of the query. The clusters are then hierarchically organized and labeled with a phrase representing the key concept of the cluster and the corresponding web documents. The web system introduces some interesting new ideas, such as the pre-selection of the retrieved web pages, the capacity to statistically detect phrases within documents and the representation of documents based on their most relevant key concepts by using web content mining techniques. Finally the system is supported by a graph-based overlapping clustering algorithm which groups the selected documents into a hierarchy of clusters [9].

Brain et al. proposed the algorithm to improve the structure of the web pages [10]. Next the problem of identifying content is treated as a sequence labeling problem, a common problem structure in machine learning and natural language processing by Gibson. Using a Conditional Random Field sequence labeling model, they correctly identify the content portion of web-pages anywhere from 80-97% of the time depending on experimental factors such as ensuring the absence of duplicate documents and application of the model against unseen sources [11].

A new WCM method of a page relevance ranking, called Page Content Rank (PCR) based on the page content exploration has been developed by Jaroslav Pokorny. In this method, the importance of a term is specified with respect to a given query q and it is based on its statistical and linguistic features. As a source set of pages for mining we use a set of pages responded by a search engine to the query q. PCR uses a neural network as its inner classification structure [12].

Unlike traditional outlier mining algorithm designed only for numeric data sets, web outliers mining algorithm should be applicable to various types of data including text, hypertext, image, video etc. Web pages that have different contents from the category in which they were taken constitute web content outliers [13]. Statistical approach and mathematical based approach has been introduced to remove the redundancy and to improve the search results [14, 15].

A new data mining algorithm to match a large number of schemas in databases at a time has been proposed. Instead of simple 1:1 matching, they do complex (m:n) matching between query interfaces. A novel correlation mining algorithm that

matches correlated attributes with smaller cost. This algorithm uses Jaccard measure to distinguish positive and negative correlated attributes [16].

A new information retrieval method using contextual information on the Web 2.0 environment has been proposed by describing the procedure for the just-in time information retrieval [17].

Hung-yu et al[18] proposed an intra-page informative structure mining system called WISDOM (Web Intra page Informative Structure Mining Based on Object Model) which applies Information Theory to DOM tree knowledge in order to build the structure. S.Jeyalatha and B.Vijayakumar [19] proposed Web Structure Mining, using the Breadth First Search strategy. Hung-yukao et al[20] developed entropy based analysis for analysing the entropy of anchor text and links to eliminate the redundancy of the hyperlink structure so that the structure of website can be distilled.

Chun-hung Li et al [21] implemented web structure mining algorithm through automative extraction of navigational structure from a web site based on the usage instead of HyperText analyses. YongheNiu et al [22] developed webkiv tool for visualizing the web mining result. Sekhar babu boddu et al [23] reviewed two popular method HITS and Page Rank based on the availability of the information in the www as becoming one of the most valuable resources for information retrieval and knowledge discovery's using web mining technologies.

A web structure mining algorithm which computes a set of PageRank vectors, biased using a set of representative topics in order to provide more accurate search results is implemented. They also gave a brief introduction to the relationship between Hyperlinks and page content [24, 25]. A correlation algorithm for web content mining has been proposed by which not only relevance ranking is calculated but also redundant documents can be detected. Removal of these redundant documents improves the quality of search results by providing unique relevant information Normalized discounted cumulative gain method is used for evaluating this ranking algorithm [26].

3. PROPOSED ARCHITECTURE

Fig 1 shows the overall architecture of the proposed method. At first the web documents are retrieved by the user from the web using search engines by giving search query. Except text other data such as hyperlinks, sound, images etc, are removed from the extracted documents. Then the extracted web documents D_i and the search query are pre-processed. The pre-processing stage comprises of three steps: Stop word removal, Stemming and Tokenization.

A stop word list normally consists of those words that convey little meaning, such as articles, conjunctions, interjections, prepositions, pronouns and forms of the "to be" verb. Next 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. Finally the output is the list of tokens from extracted web documents and list of keywords from search query being stored in the table.

The domain dictionary is compiled for a particular field that contains important terms in that field, arranged in a way that, all 1-letter word will be indexed first, followed by 2-letter words, then 3-letter words similarly up to 15-letters word which is a very reasonable upper bounds for number of characters in a term.

After pre-processing, the term frequency for all the terms in the document is calculated.

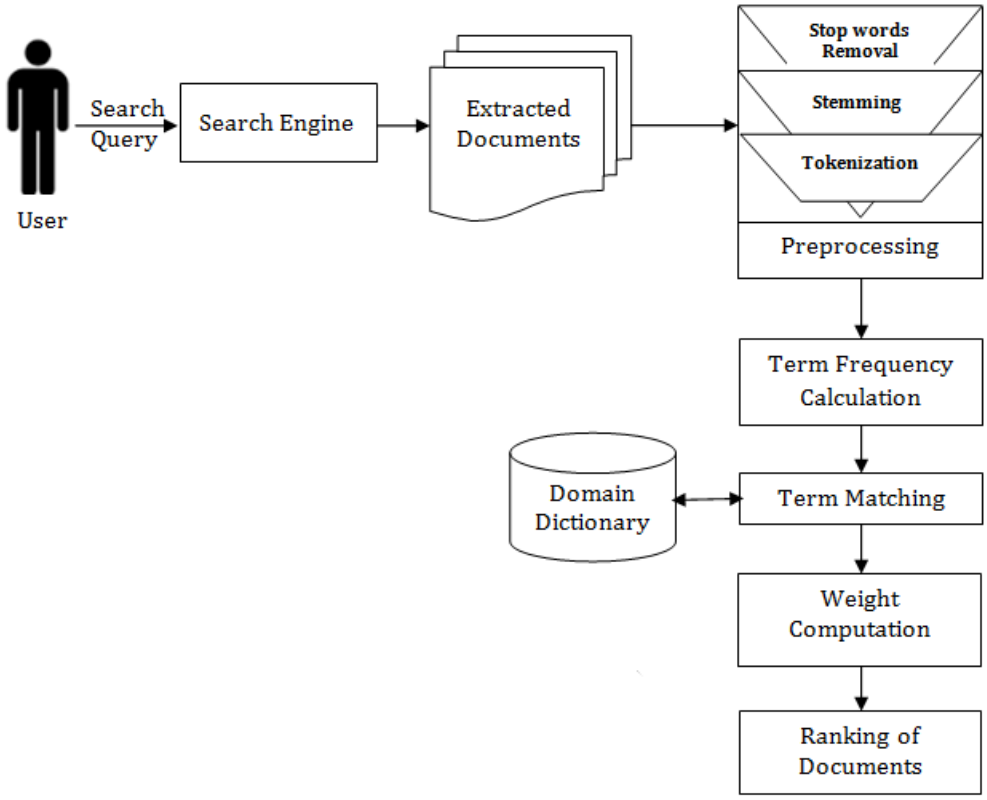

Fig 1. Architecture of the proposed System

Then each word (W_j), taken from document D_i is searched on the domain dictionary. If the word (W_j) is a keyword and is found in the dictionary, then keyword hit count is incremented by its corresponding term frequency. If the word (W_j) is not a keyword but it is found in the dictionary, then positive hit count is incremented by its corresponding term frequency. Else if the word (W_j) is not available in the dictionary then the negative hit count is incremented by its corresponding term frequency. This computation is carried out until all the words in that web documents are processed. Then the weight is applied for each category of hits and it is normalized which will be the rank of the document. The same procedure is applied for all the documents and the documents are arranged in sorted order. The least ranked document will be the

outliers which can be removed and the remaining documents are the relevant document for the given query.

3.1 Proposed Algorithm for web page relevance ranking

Input: Extracted Web Document
Method: Weighted Approach
Output: Ranked relevant web document.

Step 1: Pre-process the query by removing stop words and stemming and store it as a keyword.
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: Initialize i=1and consider Document D_i.
Step 5: Initialize Keyword_Hit(i) = 0; Positive_Hit(i) = 0; Negative_Hit(i) = 0;
Step 6: Find the term frequency $TF(W_{ik})$ for all the words W_{ik} in the document D_i where $1 \leq k \leq m$, m is the number of words in document D_i.
Step 7: If $W_{ik} \subseteq$ keyword list and also exist in domain dictionary then
Keyword_Hit(i) = Keywor d_Hit(i)+ $TF(W_{ik})$
Else if $W_{ik} \not\subset$ keyword list but exist in domain dictionary thn
Positive_Hit(i) = Positive_Hit(i)+ $TF(W_{ik})$
Else (nonexistence of W_{ik} in keyword list and domain dictionary)
Negative_Hit(i) = Negative_Hit(i) + $TF(W_{ik})$

Step 8: Increment k by repeating steps 6 and 7 until $k \leq m$;
Step 9: Compute Document Rank by giving weights to different category of hits

$$DRi = \frac{[\text{Keyword\_Hit(i)} * \alpha] + [\text{Positive\_Hit(i)} * \beta] + [\text{Negative\_Hit(i)} * \gamma]}{[\text{Keyword\_Hit(i)} + \text{Positive\_Hit(i)} + \text{Negative\_Hit(i)}]}$$

where $\alpha > \beta > \gamma$; Here the value of $\alpha=1$, $\beta = 0.75$ and $\gamma = 0.5$.

Step 10: Increment i and Repeat from step 5 until $i \leq r$;
Step 11: Sort DR_i in descending order which is the relevant document related to the user's query;

4. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed algorithm. The Experiment is conducted by compiling the domain dictionary and for the user query "Recent Research in Web Content Mining" against a Google search engine. The top 10 documents which are listed in Table 1 are retrieved and it becomes the input for the proposed method and pre-processed by removing stop words and stemming. The term frequency is calculated for each term in each document. The Keyword is extracted from the user query.

Table 1. List of Input Documents

Did	S.No	Retrieved Documents
RD1	1	http://www.cs.uic.edu/~liub/publications/editorial.pdf
RD2	2	http://www.ijarcsse.com/docs/papers/Volume_3/11_November2013/V3I11-0352.pdf
RD3	3	http://ieeexplore.ieee.org/iel5/4579839/4579840/04579960.pdf
RD4	4	http://ijirts.org/volume2issue3/IJIRTSV2I3050.pdf
RD5	5	http://ebiquity.umbc.edu/_file_directory_/papers/214.pdf
RD6	6	http://www.umiacs.umd.edu/~joseph/classes/enee752/Fall09/survey-2000.pdf
RD7	7	http://www.ijirs.com/vol2_issue-5/42.pdf
RD8	8	http://www.researchgate.net/publication/237783836_Web_Content_Mining_Research_A_Survey
RD9	9	http://www.upet.ro/annals/economics/pdf/2012/part1/Dinuca-Ciobanu.pdf
RD10	10	http://www.ijcsit.com/docs/Volume%205/vol5issue03/ijcsit20140503316.pdf

Each term in document RD1 is matched against the domain dictionary for similarity computation. If there is a match and if the term is a keyword, then the keyword hit is incremented by the corresponding term frequency, else if the term is not the keyword but is available in the dictionary, then the positive hit is incremented by the corresponding term frequency. If the word is not found in the dictionary then the negative hit is incremented by the corresponding term frequency. Finally each category of the hit is multiplied by the weight. Here the weight assigned for the keyword hit is 1, for positive hit is 0.75 and for negative hit is 0.5. Finally the values are summed up and to normalize the value, the summed value is divided by the total term frequency which will be the rank of the document. Similarly, document rank for all the documents in the list is calculated and then all the retrieved documents can be presented in sorted order based on the rank calculated. The least ranked web pages are considered as outliers and can be eliminated. Finally to calculate the accuracy, the set of document ranked using proposed method is compared against manual ranking. Table 2 shows the relevancy calculation for the above listed document.

Table 2. Relevancy Ranking

Did	Document rank value	Document Rank
RD3	0.58	1
RD4	0.53	2
RD2	0.51	3
RD1	0.49	4
RD5	0.48	5
RD7	0.45	6
RD9	0.42	7
RD8	0.39	8
RD6	0.31	9
RD10	0.28	10

5. PERFORMANCE EVALUATION

Various metrics has been used to compare the quality and relevancy of the retrieved web pages. Among them precision, recall and F-Measure plays a vital role for evaluation.

The precision is the fraction of retrieved documents that are relevant to the topic, and the recall is the fraction of relevant documents that have been retrieved.

F-Measure is calculated based on the formula

$$F= \frac{2 * \Pr ecision * \mathrm{Re} call}{\Pr ecision + recall}$$

For this proposed work, Table 1 is used as a sample data set for performance evaluation. The top 10 documents that are relevant to the user are classified manually with different user. The manual ranking is compared with the document ranking made with the proposed method. The result of this comparison is given in the Table 3.

Table 3.Comparison of the Proposed Method with Manual Ranking

Did	Search Engine Ranking	Proposed Method Ranking	Manual Ranking
RD1	1	4	4
RD2	2	3	3
RD3	3	1	1
RD4	4	2	2
RD5	5	5	5
RD6	6	9	9
RD7	7	6	7
RD8	8	8	6
RD9	9	7	8
RD10	10	10	10

The two documents RD7, RD8 and RD9 represent the mismatching of manual ranking against proposed approach. From the table, the precision of the proposed system is 0.7 whereas the precision of the search engine is 0.3 against the manual ranking.

Table 4 shows that precision calculation for varying the number of input documents.

Table 4. Precision calculation with varying data size

Dataset Size	No. of relevant documents matched with manual computation	Precision
20	17	0.85
40	33	0.83
60	50	0.83
80	71	0.89
100	85	0.85

Experimental result improves the precision of search results which also improves the efficiency of search engine.

6. CONCLUSION AND FUTURE WORK

Web mining is a growing research area in the mining community because of the information growth and information retrieval. Retrieving relevant information from the web is a very common task. Though, the results produced by most of the search engine do not satisfy the searching intension of the user. Most of the researchers pay attention to web structure mining for extracting relevant links from the web. This paper proposes a weighted approach methodology which uses term frequency and assigns weight to each term. It presents the ordered documents based on the content and also eliminates the outliers. The proposed approach produces better results compared to search engine results. Since the proposed method uses only the term matching, Future work aims at exploring pattern mining in page relevancy ranking. Also, the proposed methodology focuses only on pure text based mining to rank the web pages where relevant information may be in any format like images, audio and video. Forth coming research work will focus on all types of data sets.

REFERENCES

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

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

[3] Raymond Kosala, Hendrik Blockeel, 2000, “Web Mining Research: A Survey”, ACM SIGKDD.

[4] G. Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V.Uma, 2009, “Signed Approach for Mining Web Content Outliers”, World Academy of Science, Engineering and Technology, Vol:3.

[5] Malik Agyemang Ken Barker Rada S. Alhajj, 2005, “Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams”, ACM Symposium on Applied Computing.

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

[7] G.Poonkuzhali, K.Thiagarajan, K.Sarukesi, 2009, “Set theoretical Approach for mining web content through outliers detection”, International journal on research and industrial applications, Volume 2.

[8] Ramaswamy S, Rastogi R, Shim k, 2000, “Efficient Algorithm for mining outliers from large data sets”, proc. of ACM SIGMOD, pp 127 -138.

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

[10] Brin, S., and Page, L., 1998, “The anatomy of a large-scale hyper textual Web search engine”, Computer Networks and ISDN Systems, Vol. 30, No. 1-7, pp: 107-117.

[11] Gibson, J., Wellner, B., Lubar, S, 2007, “Adaptive web-page content identification”, WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management, New York, USA.

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

[13] Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, 2008, “A Survey on Web Content Mining and Extraction of Structured and Semistructured data”, First International Conference on Emerging trends in Engineering and Technology.

[14] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, K. Thiagarajan, K. and K. Sarukesi, 2011, “Statistical Approach for Improving the Quality of Search Engine”, 10th WSEAS International Conference on Applied Computer and Applied Computational Science, Venice –Italy.

[15] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, 2011, “Improving the quality of search results by eliminating web outliers using chisquare”, Published in Lecture notes in CCIS - Springer, Vol. 202, pp. 557-565.

[16] Shohreh Ajoudanian, and Mohammad Davarpanah Jazi, 2009, “Deep Web Content Mining", World Academy of Science, Engineering and Technology, 49.

[17] Sungrim Kim and Joonhee Kwon, 2009, "Information Retrieval using Context Information on the Web 2.0 Environment", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.10.

[18] Hung-Yu Kao, and Jan-Ming Ho, 2005, "WISDOM: Web Intra-page Informative Structure Mining Based on Document Object Model", IEEE Transactions on knowledge and data engineering, Vol. 17, No. 5.

[19] Jeyalatha.S, and Vijayakumar Design. B, 2011, "Implementation of a Web Structure Mining Algorithm using Breadth First Search Strategy for Academic Search Application", 6th International Conference on Internet Technology and Secured Transactions, Abu Dhabi, United Arab Emirates, pp.648-654.

[20] Hung-Yu Kao, and Shian-Hua Lin, 2004, "Mining Web Informative Structures and Contents Based on Entropy Analysis", IEEE Transactions on knowledge and data engineering, Vol. 16, No.1, pp. 44-55.

[21] Chun-hung Li and Chui-chun Kit, 2005, "Web Structure Mining for Usability Analysis", ACM International Conference on Web Intelligence (WI'05), IEEE.

[22] Yonghe Niu and Tong Zheng and Jiyang Chen, 2003, "WebKIV: Visualizing Structure and Navigation for Web Mining Applications", WIC International Conference on Web Intelligence (WI'03).

[23] Rajput, D.S, R.S. Thakur and G.S. Thakur, 2011, "Rule Generation from Textual Data by using Graph based Approach", International Journal of Computer Applications, Volume 31-No.9.

[24] Lili Yan and Yingbin Wei and ZhanjiGui, 2011, "Research on PageRank and Hyperlink-Induced Topic Search in Web Structure Mining", IEEE.

[25] Ramesh Prajapati, 2012, "A Survey Paper on Hyperlink Induced Topic Search (HITS) Algorithms for Web Mining", International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 2, pp. 13-20.

[26] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, P. Sudhakar, and K. Sarukesi, 2011, "Correlation Based Method to Detect and Remove Redundant Web Document", Advanced Materials Research, Vols. 171-172, pp. 543-546.