Relevance Ranking and Evaluation of Search Results through Web Content Mining

G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi

Abstract— **Nowadays, most of the people rely on web search engines to find and retrieve information. The enormous growth, diverse, dynamic and unstructured nature of web makes internet extremely difficult in searching and retrieving relevant information and in presenting query results. The aforementioned problem has given rise to the development of web content mining. This paper proposes a correlation algorithm for web content mining. In addition to relevance ranking, this algorithm also detects redundant documents. 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. The experimental result shows that this method ranks more than 90% of the relevant documents accurately.**

***Index Terms*—Correlation, NDCG, Relevant document, Redundant document, Web Content Mining**

I. Introduction

THE growth of the World Wide Web exceeded all anticipation due to enormous amount of web pages containing several billions of structured, semi-structured and unstructured documents of various types. Therefore developing powerful tool to extract relevant content from this voluminous web has become a very difficult task. Moreover, because of its dynamic nature, the use of the web as a provider of information is unfortunately more complex than working with static databases. Another important aspect is the presentation of query results. Due to its massive size, a web query can retrieve millions of resulting web pages. Thus meaningful methods for presenting these large results are necessary to help a user to select the most interesting content. The aforementioned problems results in the development of web content mining.

this uses the ideas and principles of data mining and knowledge discovery to screen more specific data.

F. G.Poonkuzhali is with the Department of Computer Science & Engineering,, Rajalakshmi Engineering College, Affiliated to Anna University, Chennai, India, phone: +91 9444836861; (e-mail: poonkuzhali.s@ rajalakshmi.edu.in).

S. R.Kishore Kumar is with the Department of Computer Science & Engineering, Sri Sivasubramaniya Nadar College of Engineering, Old Mahabalipuram Road, SSN Nagar -603 110, Tamil Nadu, India (e-mail: rskishorekumar@yahoo.co.in).

T. P.Sudhakar is with the Department of Computer Science and Engineering , Kamaraj College of Engineering, Tamil Nadu, India (e-mail: sudhakar.asp@gmail.com).

Web content mining refers to the discovery of useful information from web contents, including text, image, audio, video, metadata and hyperlinks etc. Web content mining also distinguishes personal home pages with other web pages. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages. Two groups of web content mining specified in [4] are those that directly mine the content of documents and those that improve on the content search of other tools like search engine.

Correlation is the most popular and effective statistical technique used for analyzing the behaviour of two or more variables. The relation between variables can be verified and tested for significance with the help of the correlation analysis. In the proposed work, correlation analysis is used to find the related documents from the input document set of some particular category. The Coefficient of correlation (r) lies between -1 and +1. When r is positive, then the documents D_i and D_j are said to be positively correlated. When r is equal to 1, then the two documents have perfect positive correlation which is redundant. When r is negative, then the documents D_i and D_j are said to be negatively correlated. When r is equal to 0, then D_i and D_j are uncorrelated. The correlation between two documents is called simple correlation. The correlation in the case of more than two documents is called multiple correlation [16] .

Discounted cumulative gain (DCG) is a measure of effectiveness of a Web search engine algorithm, often used in information retrieval. Using a graded relevance scale of documents in a search engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. Search result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of *p* should be normalized across queries.

In this work, multiple correlation with the value ranging from 0 to 1 is used for detecting web content outliers. It works in a bivariate method, i e. The correlation between two documents is found iteratively. Then, correlation coefficient of each document is summed to get the total correlation coefficient. Finally, the documents are ranked in descending order based on this total correlation coefficient to detect the top n relevant documents.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. RELATED WORKS

Due to the heterogeneity and the lack of structure of web data, automated discovery of targeted or unexpected knowledge information still present in many challenging research problems. Moreover, the semi-structured and unstructured nature of web data creates the need for web content mining. In [9], the author differentiates web content mining from two different points of view. Information Retrieval view and Database view. Characteristics of web and various issues on web content mining presented in [1]. In paper [8] research areas of web mining and different categories of web mining are discussed briefly. They also summarized the research works done for unstructured data and semi-structured data from information retrieval (IR) view. In IR view, the unstructured text is represented by bag of words and semi-structured words are represented by HTML structure and hyperlink structure [8]. In Database (DB) view, the mining always tries to infer the structure of the web site to transform a web site into a database. A new method for relevance ranking of web pages with respect to a given query was determined in [5]. Various problem of identifying content such as a sequence labeling problem, a common problem structure in machine learning and natural language processing is identified in [3]. A survey of web content mining plays as an efficient tool in extracting structured and semi structured data and mining them into useful knowledge is presented in [6]. A framework is proposed to provide facilities to the user during search [7]. In this framework a user does not need to visit the homepages of companies to get the information about any product, instead the user write the name of product in the Query Interface (QI) and the framework searches all the available web pages related to the text, and the user gets the information with little efforts. In [10]-[12] Statistical approach using proportions and chi-square for retrieving relevant information from both structured and unstructured documents are presented. The authors applied correlation method to detect and remove redundant web documents.

Outline of Paper

Section 2 presents the related works. Section 3 presents architectural design of the proposed system. Section 4 presents the algorithm for correlation ranking. Section 5 presents experimental results. Section 6 presents performance evaluation. Finally section 7 presents conclusions and future work.

III. FRAMEWORK OF WEB CONTENT MINING

The proposed algorithm explores the advantages of full word matching through Correlation Approach using domain dictionary. Initially all extracted input web documents are pre-processed. The Pre-processing includes, stemming, stop words removal and tokenization. Stemming is the process of comparing the root forms of the searched terms to the documents in its database. Stop words elimination is the process of not considering certain words which will not affect the final result. Tokenization is defined as splitting of the words into small meaning full constituents

. After pre-processing, the profile for all the words are generated and stored in hash table. Following the above process, term frequency for all the words is computed. Followed by that Correlation co-efficient is computed between these two documents. If the Correlation co-efficient value is 1, then the above documents are fully redundant. Similarly correlation co-efficient is computed for all other documents. Then the total correlation co-efficient of each document is computed. This process is repeated for the remaining documents. Finally, the total correlation co-efficient is ranked in descending order. The top 'n' documents are declared as an relevant web document.

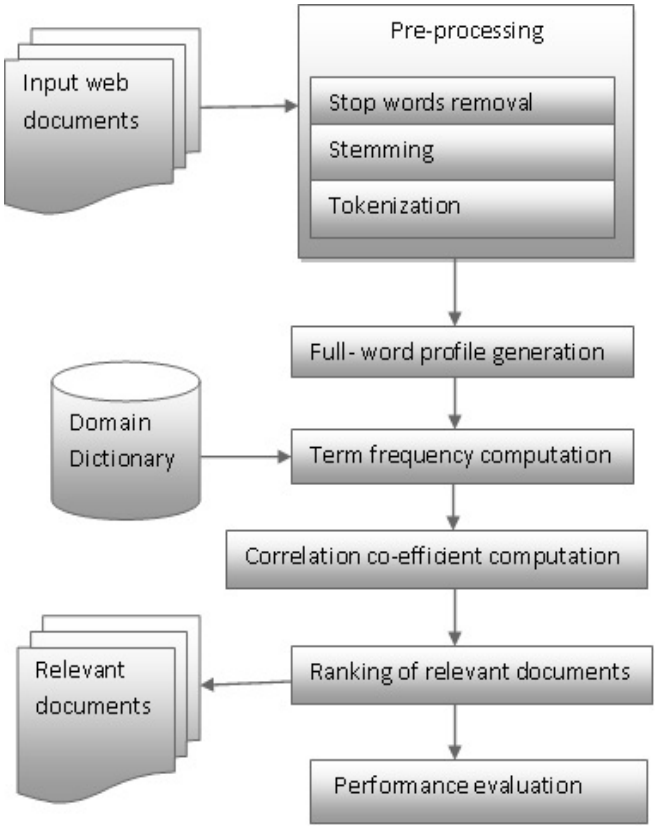

Fig. 1 Architectural design of the proposed work

IV. ALGORITHM FOR WEB CONTENT MINING

Algorithm : Correlation algorithm for relevance ranking

Input : Web document $D = \{D_1, D_2, ...D_N\}$

Method : Correlation method

Output : Relevant documents.

Step 1: Extract the input web documents D_i where $1 \leq i \leq N$.
Step 2: Pre-process the entire extracted documents.
Step 3: Initialize redundant document set RD = {};
Step 4: Initialize i = 1
Step 5: Initialize j = 1.
Step 6: Perform the correlation coefficient R_{ij} between D_i

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

and D_j

If $i == j$ then $R_{ij} = 0$ Goto step 1

else

Compute the following steps :

Extract the common words between D_i and D_j that matches with domain dictionary. Let T be the set of common words and the number of elements in the T be m, i.e. $|T| = m$.

Compute the term frequency TF $(W_k)_i$ in D_i and TF $(W_k)_j$ in Dj where $1 \le k \le m$.

Determine:

X_k = TF $(W_k)_i$ for the words in document D_i and Y_k = TF $(W_k)_j$ for the words in document D_j

Calculate: $\sum X_k, \sum X_k^2, \sum Y_k, \sum Y_k^2, \sum X_k Y_k$,

Compute: R_1, R_2 and R_3

$$R_1 = \sum X_k^2 - \frac{\sum X_k^2}{|T|}$$

$$R_2 = \sum Y_k^2 - \frac{\sum Y_k^2}{|T|}$$

$$R_3 = \sum X_k Y_k - \frac{\sum X_k \sum Y_k}{|T|}$$ Perform:

$$R_{ij} = \frac{R_3}{\sqrt{R_1} x \sqrt{R_2}}$$

Step 7: If ($R_{ij} == 1$) then D_i and D_j are redundant;

Assign $RD_i = RD_i \cup D_j$ where $1 \le i \le N$.

else D_i and D_j are not redundant;

Step 8: Increment j, and repeat from step 6 to step 7 until $j \le N$.

Step 9: Compute the total correlation coefficient:

$$\sum R_{ij}$$ where j=1 to N.

Step 10: Increment i, and repeat from step 5 to step 9 until $i \le N$.

Step 11: Sort total correlation coefficient in descending order.

Step 12: Remove redundant data set (RD).

Step 13: Display the top 'n' relevant documents.

V. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed algorithm for web content mining For testing purpose, 10 input documents listed in Table 1 are taken. These input dataset are pre-processed and then the term frequency for the common words between document D_i and D_j (j=i+1) is computed. Then the correlation coefficient is computed for these documents. If the correlation coefficient value is equal to 1, then D_j is stored in redundant document (RD) set. The same process is repeated for the remaining documents in the input dataset. Then the total correlation coefficient is computed for each document. Finally, the documents are stored in descending order based on the total correlation coefficient to detect top 'n' relevant documents. The lowest value of total correlation coefficient indicates the least relativity (dissimilarity) of that document with other documents in the input dataset, whereas the highest value of total correlation coefficient indicates the most relatively(similarity) of that document with other documents. The input documents listed in Table 1 are ranked using the correlation algorithm based on total correlation coefficient measure is presented in Table 2.

TABLE I
INPUT DOCUMENTS

Document No.	URL
D1	http://www.waset.org/journals/waset/v56/v56-150.pdf
D2	140.115.80.66/data%20**mining**%20paper%20databases/.../...
D3	http://web.cs.dal.ca/~jiz/stream_1.pdf
D4	http://dl.acm.org/citation.cfm?id=968022
D5	http://www.wseas.us/e-library/conferences/2011/Venice/ACACOS/ACACOS-12.pdf
D6	http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5708790
D7	http://www.arnetminer.org/viewpub.do?pid=893335
D8	http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1045051
D9	http://www.skybrary.aero/bookshelf/books/1125.pdf
D10	http://www.salinesystems.org/content/pdf/1746-1448-1-12.pdf

TABLE II
RELEVANCE RANKING USING CORRELATION

Top 'n' position	Total Correlation Coefficient	Relevance Ranking
1	2.4141	D5
2	2.2224	D1
3	1.9467	D4
4	1.9465	D6
5	1.8003	D2
6	1.6583	D7
7	1.5242	D3
8	1.444	D8
9	0.798	D9
10	0.4625	D10

VI. PERFORMANCE EVALUATION

The Discounted cumulative gain (DCG) is defined as the sum of the products of a relevance score (RS) and its position weight which is logarithmically proportional to the position of the returned documents. DCG can be used to evaluate the search results by the ranking algorithms. DCG method ranks the quality of returned documents by finding the Cumulative relevance gain of all the documents returned by raking algorithms. The positional parameter c represents the ranking score assign by the correlation algorithm. The positional parameter I represents the ranking score assign by the Ideal (Experts).

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The DCG accumulated at a particular rank position c through correlation ranking is defined as:

$$\mathrm{DCG_c} = \sum_{c=1}^{n} \left(\frac{RS_c}{Log_2(c+1)} \right)$$

The DCG accumulated at a particular rank position I through Ideal ranking is defined as:

$$\mathrm{DCG_I} = \sum_{I=1}^{n} \left(\frac{RS_I}{Log_2(I+1)} \right)$$

The Normalized Discounted Cumulative Gain (NDCG) is defined as searched result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of p should be normalized across queries through ideal ranking. It is also defined as ratio of Discounted Cumulative gain of the ranked documents derived from the correlation method to the ideal ranking. NDCG is computed as:

$$NDCG = \frac{\mathrm{DCG_c}}{\mathrm{DCG}_I}$$

The Cumulative Gain is normalized for the ideal result list. NDCG values vary from 0 to 1.

The results obtained through correlation algorithm listed in Table II are validated by Normalized Discounted Cumulative Gain method using the above formula. Here the relevance of each document was judged by the human experts (called Ideal ranking) and a relevant score was assigned to each document by 10 points graded scale where 10 indicated the most relevant document and 0 indicated the most irrelevant. All the documents are sorted by the relevance scores and summarized in table III.

TABLE III
NDCG EVALUATION

Documents Ranked by Human	Ideal Ranking Position(I)	Correlation Ranking Position(C)	Relevance Score (RS)
D4	1	3	10
D1	2	2	10
D2	3	5	9
D7	4	6	9
D5	5	1	8
D6	6	4	8
D3	7	7	6
D8	8	8	4
D9	9	9	0
D10	10	10	0

$$\mathrm{DCG_c} = \sum_{c=1}^{n} \left(\frac{RS_c}{Log_2(c+1)} \right) = 34.22717$$

$$\mathrm{DCG_I} = \sum_{I=1}^{n} \left(\frac{RS_I}{Log_2(I+1)} \right) = 34.48182$$

$$NDCG = \frac{\mathrm{DCG_c}}{\mathrm{DCG}_I} = 0.992615$$

The NDCG value for the results obtained through correlation algorithm for relevance ranking is 0.99 which is nearer to 1. Therefore the correlation method of relevance ranking is accurate.

NOMENCLATURE

DCG	Discounted Cumulative Gain.
NDCG	Normalized Discounted Cumulative Gain.
RS	Relevant Score.
C	Position of an item returned by correlation ranking.
I	Position of an item returned by Ideal ranking (Ranked by Human).

VII. CONCLUSION

The popularity of World Wide Web has received a tremendous attention by majority of the people to find and retrieve relevant information for various purposes. Therefore, most of the researchers pay attention to web content mining for extracting relevant documents. This paper proposes correlation method for relevance ranking and normalized discounted cumulative gain for evaluating this ranking method. The quality of search results obtained through this approach is accurate.

.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ponnammal Natarajan worked as Former Director – Research , Anna University- Chennai,India and currently an Advisor, (Research and Development), Rajalakshmi Engineering College and Dr. K..Ravi, Associate Professor, Department of Mathematics, Sacred Heart College-Tirupattur, India for their intuitive ideas and fruitful discussions with respect to the paper's contribution.

.

REFERENCES

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

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

[3] Gibson, J., Wellner, B., Lubar, S, "Adaptive web-page content identification", In WIDM '07: *Proceedings of the 9th annual ACM*

international workshop on Web information and data management. New York, USA,2007.

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

[5] Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An approach to the Web Content Mining",

[6] Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya 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.

[7] Mahmoud Shaker, Hamidah Ibrahim, Aida Mustapha, Lili Nurliyana Abdullah, "A Framework for Extracting Information from Semi-Structured Web Data Sources," iccit, vol. 1, pp.27-31, *2008 Third International Conference on Convergence and Hybrid Information Technology*, 2008

[8] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD*, July 2000, Vol-2, pp 1-15.

[9] R. Cooley, B. Mobasher, and J. Srivastava. "Web mining: Information and pattern discovery on the World Wide Web", *In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI'97), 1997.

[10] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, K. Thiagarajan, K. and K. Sarukesi, paper titled "Statistical Approach for Improving the Quality of Search Engine", *10th WSEAS International Conference on Applied Computer and Applied Computational Science,* Venice – Italy, March 8-10, 2011.

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

[12] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, "Improving the quality of search results by eliminating web outliers using chi-square*", Published in Lecture notes in CCIS – Springer,* Vol. 202, pp. 557-565, 2011.

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

[14] Chakrabarti, S. "Mining the Web: Discovering Knowledge from Hypertext Data", *Morgan-Kauman Publishers*,2002.

[15] Han, J. and Kamber, M. "Data Mining: Concepts and Techniques", *Morgan Kaufmann Publishers*,2001.

[16] D.C.Sancheti and E.K.Kapoor , "Statistics(Theory, Methods & Application" Published by *Sultan Chand and Sons*.

G.Poonkuzhali received B.E degree in Computer Science and Engineering from University of Madras, Chennai, India, in 1998, and the M.E degree in Computer Science and Engineering from Sathyabama University, Chennai, India, in 2005. Currently she is pursuing Ph.D programme in the Department of Information and Communication Engineering at Anna University – Chennai, India. She has presented and published 15 research papers in international conferences & journals and authored 5 books. She is a life member of ISTE (Indian Society for Technical Education) ,IAENG (International Association of Engineers), and CSI (Computer Society of India).

R.Kishore Kumar received B.E degree in Computer Science and Engineering from Rajalakshmi Engineering College, Anna University, Chennai, India in 2011. Currently he is pursuing M.E degree in Computer Science and Engineering in SSN College of Engineering. .He has presented 8 papers in International conferences and published 5 research papers in international journals and 3 papers in national journals. One of his paper has been selected as the Best Paper. He is also the member of Computer Society of India.

Sudhakar received Bachelor of Engineering degree under Computer science and Engineering stream Anna University Chennai-India in 2006 and Master of Engineering degree under Computer Science and Engineering stream Anna University Chennai-India in 2008. After 4 years of Software development experience on Web and Windows applications, Currently he is working as an Assistant professor in Kamaraj College of Engineering and Technology, Virudhunagar. He also presented many papers in National and International conferences and published his research works in International Journals. He is a life member of ISTE (Indian Society for Technical Education) and member in IAENG (International Association of Engineers), WSEAS.

Dr. G.V.Uma received her M.E. from Bharathidasan University, India in year 1994 and Ph.D. from Anna University, Chennai, India in 2002. She has rich experience in teaching and research; Currently working as a Professor and Head in the Department of Information Science and Technology in Anna University Chennai. Her research interests include Software Engineering, Genetic Privacy, Ontology, Knowledge Engineering & Management, and Natural Language Processing. She has organized many Workshops, Seminars and Conferences in national and International level

Dr. K. Sarukesi has a very distinguished career spanning of nearly 40 years. He has a vast teaching experience in various universities in India and abroad. He was awarded a commonwealth scholarship by the association of common wealth universities, London for doing Ph.D in UK. He completed his Ph.D from the University of Warwick – U.K in the year 1982. His area of specialization is Technological Information System. He worked as expert in various foreign universities. He has executed number of consultancy projects. He has been honored and awarded commendations for his work in the field of information technology by the government of TamilNadu. He has published over 40 research papers in international conferences/journals and 40 National Conferences/journals.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)