Statistical Approach for Improving the Quality of Search Results

G.POONKUZHALI[1], R.KISHORE KUMAR[2], R.KRIPA KESHAV[3], K.THIAGARAJAN [4]
K.SARUKESI[5]

[1, 2, 3] Department of Computer Science and Engineering, Rajalakshmi Engineering College,
Affiliated to Anna University- Chennai, Tamil Nadu
[4]Department of Science and Humanities, KCG College of Technology
Affiliated to Anna University-Chennai, Tamil Nadu
[5] Hindustan Institute of Technology and Science-Chennai, Tamil Nadu
INDIA
[1]poonkuzhali.s@rajalakshmi.edu.in , [2]rskishorekumar@yahoo.co.in,
[3]kripa_keshav@yahoo.co.in , [4]vidhyamannan@yahoo.com , [5] profsaru@gmail.com

Abstract: - Today, the most powerful tool in the internet world is the search engine as most of the people rely on them for retrieving interesting documents. Due to huge amount of information available on the web, most of the documents retrieved from the search engine are mostly irrelevant and causes a waste of user Therefore there is a need for Information retrieval and web mining researchers to develop an automated tool for improving the quality of the search results returned by search engines. In this research work, a statistical approach using test hypothesis with degrees of confidence at level 95% is used for retrieving the relevant web documents. This algorithm works well for both structured and unstructured web documents with high precision.

Key-Words: - critical value**,** degrees of confidence, relevant, test statistic, web document.

1 Introduction

As the information in the web world has increased, accessing information has become very difficult. Moreover, it causes a waste of user time in navigating a lot of links and finally end up with uninteresting results. This problem is mainly because of web scale due to voluminous and high dimensionalities of the documents. This has necessitated the users to make of automated tools to locate desired information resources on the web. Web mining is the application of data mining techniques to discover constellations from the Web . It is the extraction of fascinating and potentially useful patterns which are essential information related to the World Wide Web. Web mining can be categorized into three parts: web content mining, 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 [1]-[4]. Two groups of web content mining are those that directly mine the content of documents and those that improve on the content search of other tools like search engine [9]-[10].

This paper deals with web content mining. Web content mining deals with lots of issues like extraction of structured and unstructured documents, web integration and noise detection but the most important is the optimization of search engine. The search engine can be optimized by retrieving the relevant web documents. In today's search engine lots of irrelevant [7] web documents are present which causes lots of wastage in user time and retrieval time.

Existing algorithms uses n-gram technique with domain dictionary to determine the similarity of strings and expand it to include pages containing similar strings. [5]-[6]. and mathematical approach based on set theoretical and signed representation[7]-[8] using full word matching with domain dictionary for retrieving relevant documents. The proposed algorithm uses test statistic using proportions for retrieving relevant web documents.

In this work, web documents are extracted from the search engines based on the query given by the user. Then the obtained web documents are pre-processed, i.e., stop words, stem words and expect text other data such as hyperlinks, sound, images etc are removed. Each document is mined to retrieve relevant web document through test hypothesis

`

using proportions. When the value of |Z| is equal to or less than 1.645 [11] then those documents are relevant. Finally, a mined web document is obtained which contains required information catering to the user needs.

Outline of the paper:
Section 2 presents the overview of the Architectural design of the proposed system. Section 3 presents the algorithm for retrieving relevant web documents using test hypothesis. Section 4 gives the Experimental results. Section 5 presents conclusion.

2 Architectural Design

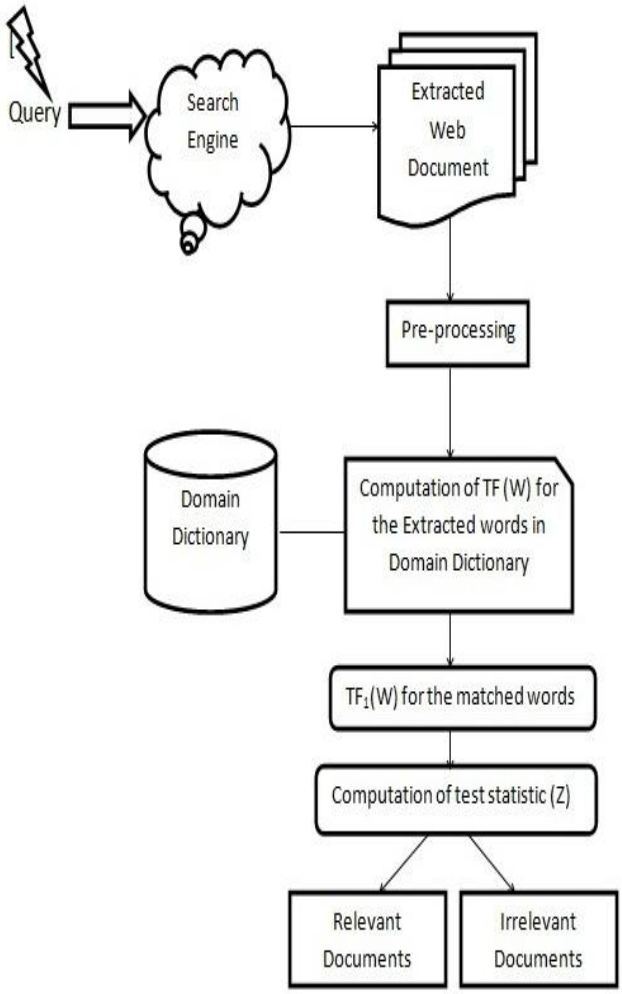


Fig. 1 Architectural Design

In this algorithm, web documents are extracted based on the user query. The extracted documents are pre-processed for making the remaining process simpler. The pre-processing step involves: removal of stop words, stemming and tokenization. Followed by this, term frequency for the words presents in the document against domain dictionary is computed for the i^{th} and j^{th} ($i+1^{th}$) documents. Then, similar words from the above documents along with their term frequencies are retrieved for performing test statistic (Z) using proportions. Finally, Z value is compared with the degrees of confidence at the level of 95% which is obtained from the table. If the calculated value is equal or less than 1.645 then both are considered as relevant documents otherwise, they are considered as irrelevant documents. The above process is repeated for all the remaining documents for computation of relevance.

3. Algorithm for retrieving relevant document through test hypothesis.

Input : Web document.

Method: Statistical Method

Output: Extraction of relevant web document.

Step1: Extract the input web document D_i where $1 \le i \le N$.
Step 2: Pre-process the entire extracted document
Step 3: Initialize i=1.
Step 4: Initialize j=i+1.
Step 5: Consider the document D_i and D_j.
Step 6: Find the term frequency for all the words TF (W_{ik}) in D_i and TF (W_{jk}) in D_j that exist in Domain Dictionary, where $1 \le k \le m$.
Step 7: Calculate TF_1 (W) the total number of words as N1 and N2 in D_i and D_j that matches with Domain Dictionary.
Step 8: Perform the Proportionate Calculation for the common words between D_i and D_j through the following steps:

Compute: $P1 = \sum X_{ik} / N1$,
$P2 = \sum Y_{jk} / N2$
where X_{ik} and Y_{jk} are the Term Frequency of D_i and D_j.

Perform Standard Error :
$S.E(P1 - P2) = SQRT [[P1 * (1 - P1)/N1] + [P2 * (1 - P2) / N2]]$.

Calculate the Test Statistic:
$|Z| = p1 - p2 / S.E. (P1 - P2)$

Step 9: Compare |Z| value with the $Z_\alpha = 1.645$ at $\alpha = 95\%$ at level of confidence, where Z_α is the Critical Value.

`

Step10: If the Z value is lesser than Critical Value then
D_i and D_j are relevant documents.
Else
D_i and D_j are Irrelevant.

Step 11: Increment j, and repeat from step 5 to step 9 until $j \leq N$.
Step 12: Increment i, and repeat from step 4 to step 10 until $i < N$.

Nomenclature:

Variables	Description
SE	Standard error
P1	Sample proportion for i^{th} Document
P2	Sample proportion for j^{th} Document

4. Experimental Results

Here 5 web documents listed in table1 are taken for test study. Initially these documents are pre-processed and then the term frequencies for the similar words taken for the first two documents are computed. Followed by that, the statistical test hypothesis using proportions is applied for those two documents to check the relevancy between them. Similarly, the relevancy for the remaining documents is computed. In this approach the degrees of confidence at 95% level which holds the value 1.645 is obtained from the statistical table. The statistical test value for the input documents is computed in table 2 and relevancy among all the documents in represented in fig.2.

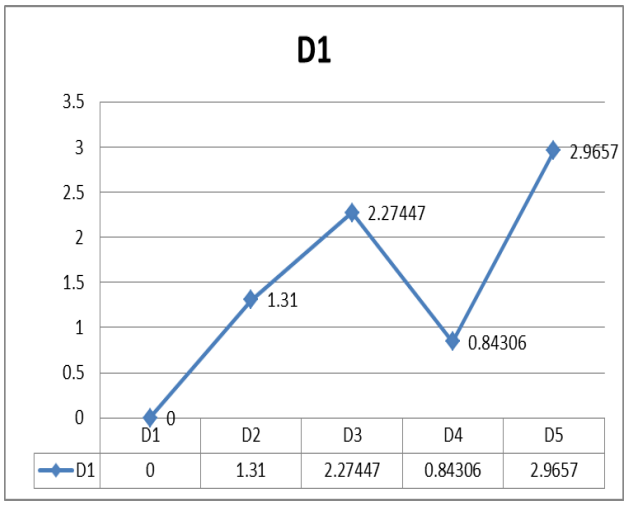
Table 1: Input documents

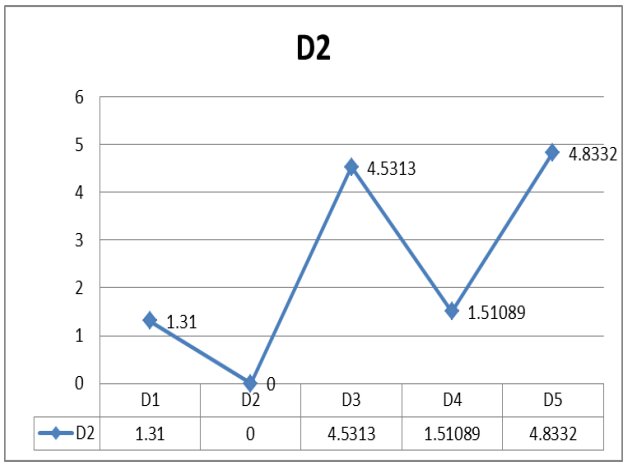
D.No	Document Name
D1	Wcm.pdf
D2	Page Content rank an approach to the web content mining.pdf
D3	Neural Analysis.pdf
D4	Deep_WCM.pdf
D5	Medical Mining.pdf

Table 2: Experimental results

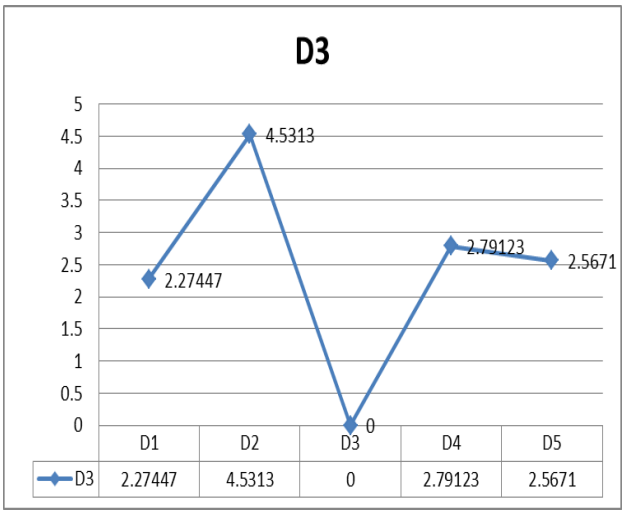
	D1	**D2**	**D3**	**D4**	**D5**
D1	*	1.310	2.27447	0.84306	2.9657
D2	*	*	4.53130	1.51089	4.8332
D3	*	*	*	2.79123	2.5671
D4	*	*	*	*	3.4015
D5	*	*	*	*	*

From the table 2, it is clear that documents 1, 2 and 4 are less than or equal to 1.645. Therefore these documents are relevant. On the other hand documents 3 and 5 have values greater than 1.645, thus concluding them to be irrelevant.





`



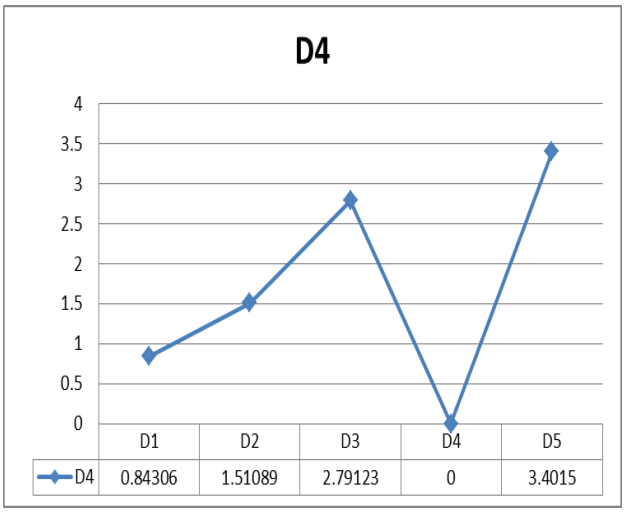


Fig 2 : Graphical results of relevancy among other input documents

Experimental results ensure that the memory space gets reduced and improves the accuracy of search results, after eliminating the irrelevant documents. As the efficiency of web content is increased, the quality of the search engines also gets increased. Precision and recall of the refined documents increases considerably.

5 Conclusion

Web mining is a growing research area in the mining community. Retrieving relevant content from the web is a very common task. However, the results obtained, by most of the search engines do not necessarily produce result that is best possible catering to the user needs. This paper proposes statistical approach using test hypothesis with 95% level of confidence for retrieving relevant web documents from structured as well as unstructured documents. The quality of search results obtained through this approach is accurate. In future, comparative study with other mathematical approaches is to be done. Also eexperimental evaluation is done in terms of precision, recall and response time. Finally, benchmark data needs to be established for evaluating the performance of this algorithm with other existing algorithms.

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] 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.

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

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

[6] Malik Agyemang, Ken Barker, Rada S. Alhajj, Framework for Mining Web Content Outliers , *ACM Symposiumon Applied Computing* 2004.

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

[8] 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.56,2009,PP 820-824.

[9] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, *ACM SIGKDD*, July 2000.

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

[11] Statistics(Theory, Methods & Application) By D.C.Sancheti and E.K.Kapoor Published by Sultan Chand and Sons, Sixth thoroughly revised Edition, 1990.

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 10 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), ISCSIT and CSI (Computer Society of India).

R.Kishore Kumar currently under-graduate student of Rajalakshmi Engineering College .He has presented 5 papers in conferences and published 4 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 International Association of Engineers and Computer Society of India.

R.Kripa Keshav currently under-graduate student of Rajalakshmi Engineering College. He is the Member of computer society of india.He has presented one paper in national conference and won the best paper award. One paper is published in international journal. He is a member of International Association of Engineers and Computer Society of India.

K.Thiagarajan working as Senior Lecturer in the Department of Mathematics in KCG College of Technology - Chennai-India. He has totally 14 years of experience in teaching. He has attended and presented research articles in 33 National and International Conferences and published one national journal and 26 international journals. His area of specialization is coloring of graphs and DNA Computing.

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 specializations 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.