Effective Algorithms for Improving the Performance of Search Engine Results

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

***Abstract*—**Search engine has become an important tool in today's world for searching various data like text, audio, video and images. While searching for information many users end up with irrelevant information causing a waste in user time and accessing time of the search engine. So to narrow down this problem, many researchers are involved in web mining, still the algorithms developed by them contains nearly 30% of outlaid content such as irrelevant and redundant information. In this paper, two Statistical approaches based on Proportions (Z-test hypothesis) and chi square test (T-test) is developed for mining this outlaid content. Also comparative studies between these two methods are presented. Elimination of this outlaid content during a searching process improves the quality of search engines further. The results show that the system easily provides relevancies and delivers dominant text extraction, supporting users in their query to efficiently examine and make the most of available web data sources. Experimental results revealed that statistical approach produces better results than chi square test.

Keywords-Chi Square Test, critical value**,** degrees of confidence, relevant, test statistic, web document.

I. INTRODUCTION

TODAY, the World Wide Web contains several billions of information and is still growing at a very faster rate as most of the people use the internet for retrieving interesting document. But most of the time, they lose their temper by getting lot of insignificant document even after navigating several links. Thus developing user friendly tool for retrieving the relevant content without accessing the complete data on the outset has become an important concern among the Web mining research communities. Web mining consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks which helps to investigate the node and connection structure of web sites. According the type of web structural data, web structure mining can be divided into two kinds 1) extracting the documents from hyperlinks in the web 2) analysis of the tree-like structure of page structure. Based on the topology of the hyperlinks, web structure mining will categorize the web page and generate the information, such as the similarity and relationship between different web sites Web content mining aims to extract/mine useful information or knowledge from web page contents. Web content mining is sometimes referred as web text mining, as most of the research is related on text. due to Web Content Mining applies the concepts of data mining and knowledge discovery to retrieve more specific data[1][9]-[11]. Some of the areas of doing research in web content mining is listed below:

- *Structured Data Extraction* →Data extraction is the act or process of retrieving data out of data sources for further data processing or data storage.
- *Unstructured Text* Extraction →Typically unstructured data sources include web pages, email, documents, PDFs, scanned text, mainframe report, spool files etc.
- Web Information Integration and Schema matching→ although the web contains a huge amount of data, each web site represents similar information differently. How to identify or match semantically similar data is a very important problem with much practical application.
- Building Concept Hierarchies →Concept hierarchies are important in many generalized data mining applications, such as multiple level associations rule mining.
- *Segmentation and Noise Detection* → In many web applications, one only wants the main content of the web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the page is interesting problem. A number of interesting techniques have been proposed in the past few years.
- *Opinion extraction →:* mining opinions is of great importance for marketing intelligence and product benchmarking

This paper focuses on segmentation and detection of noise issue, which implies outliers mining. Generally, Outliers are observations that deviate so much from other observations to arouse suspicions that they might have been generated using a different mechanism or data objects that are inconsistent with the rest of the data objects. Most of the web content mining algorithms works on finding frequent patterns while neglecting less frequent ones namely outliers. Web Content Outliers are web document that show significantly different characteristics than other web documents taken from the same category. Outliers identified in web data are referred to as *web outlier*. This paper focus on Web Content Outliers Mining for

extracting insignificant web content through the mathematical approach based on chi-square test.

The proposed system uses chi-square test algorithm to find the web document which are assumed to be more explicit to the user query. Next comparison between the calculated value and the degrees of freedom value at 95% level of Confidence is done. If the chi-square value is greater than chi-square distribution (table value), then that document is considered to be significant with the other document which in turn implies it is relevant to the user query. If the chi-square value is less than chi-square distribution (table value), then that document is considered to be insignificant with the other document which in turn implies it is irrelevant to the user query. Elimination of insignificant web document results in retrieval of interesting web documents to the user.

II. RELATED WORKS

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 [11]-[12]. N. P. Gopalan et al. attempts to find a solution for scalability problem by incorporating agent based systems for mining the hyperlinks on a web page to find more quality web pages. The solution is designed in such a way that it narrows down the gap between the current web content and the one found at the previous crawls [14].Malik Agyemang et al establish the presence of outliers on the web with various types of outliers present on the web and designed a framework for mining web content outliers using full word matching assuming the existence of domain dictionary. The above authors developed the work with n-gram techniques for partial matching of strings with domain dictionary[5]-[6]. Malik Agyemang et al. enhanced the same work without domain dictionary. Based on the above ideas, Malik Agyemang et al prolonged the work by presenting HyCOQ which a hybrid algorithm that draws from the power of n-gram based and word based system[7] . There is a remarkable improvement in recall with hybrid documents compared to using raw words and n-grams without a domain dictionary still it covers mining only structured web documents. G.Poonkuzhali et al presented the mathematical approach based on set theoretical and signed approach for mining web content outliers[8]-[9]. K.Thiagarajan et al. implemented weighted graph approach of trust reputation management through signed concept which can also be applied for retrieving the relevant content[4]. G. Castellano et al. explores the possibility of using fuzzy clustering to mine usage profiles from web log data. In particular, Castellano et al. [15] uses the fuzzy C-Means algorithm to categorize user sessions in order to derive groups of users which exhibit similar access patterns. The obtained clusters represent user profiles which can be exploited to implement different personalization functions, such as dynamic suggestion of links to Web pages retained interesting for the user [15]. Zakaria Suliman Zubi et al. attempts to achieve a better understanding of Arabic text classification techniques [19]. Ioan Pop et al. emphasizes the importance of using appropriate measures and methods to evaluate the performance of Web document classification [17]. Ioan Dzitac et al. proposes the structure into two sections. The first one briefly discusses the different web mining tasks and the second one is focusing on advanced artificial intelligence (AI) methods for information retrieval and web search, link analysis, opinion mining and web usage mining [18].

III. ARCHITECTURE DIAGRAM

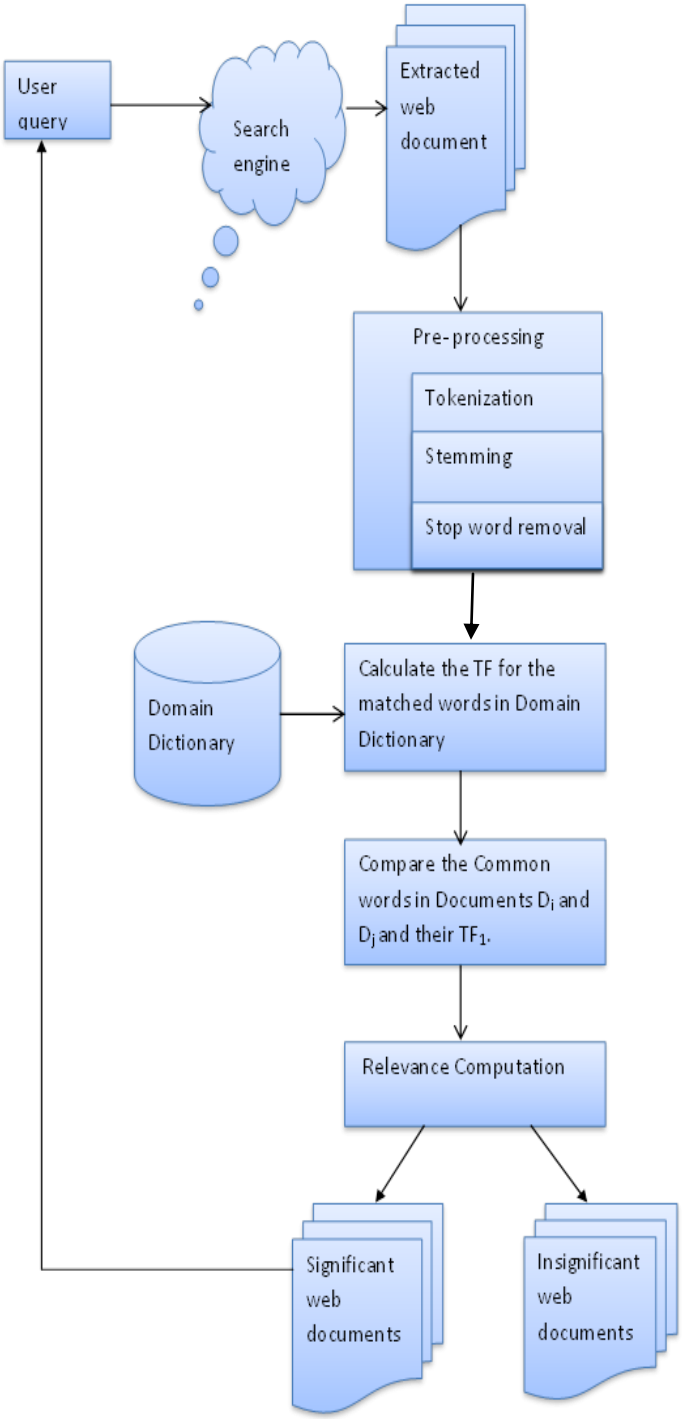

Fig 1 Architectural Design of the Proposed System.

The proposed system can be divided into 5 modules-1) user input 2) pre-processing 3) term frequency 4) comparison of term frequencies of similar word between both documents 5) Relevance computation.

The first module is where the user gives the input query. Based on that query the documents are retrieved from the search engine. Most of the documents retrieved from the search engine may or may not be relevant to the user query.

The second module is pre-processing. The various steps involved in pre-processing are stemming, stop word removal and tokenization. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. Stop words are common words that carry less important meaning than keywords. Usually search engines remove stop words from a keyword phrase to return the most relevant result. Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing.

The third module is the term frequency calculation. The words present in the document are compared with the words present in the domain dictionary. So the words that are matched with the dictionary ate taken for the term frequency calculation. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the term frequency weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

The fourth module is the term frequency calculation of the compared words that match with domain dictionary between both the documents Di and Dj.

The fifth module is the relevance calculation using the statistical methods. Here in this proposed system we use two statistical methods test hypothesis using proportions and chi square. Proportions are just means! The proportion having a particular characteristic is the number of individuals with the characteristic divided by total number of individuals. A chi-square test is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true, or any in which this is asymptotically true, meaning that the sampling distribution can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

Now the comparison is based on the precision calculation. The computed values for proportions and chi square are then compared with results obtained manually.

IV. RELEVANCE COMPUTATION THROUGH STATISTICAL METHODS.

Algorithm 1:

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 (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.
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 $j \le N$.
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
Z_α	Critical Value.

Algorithm 2:

Input : Web document.

Method: Chi-Square 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_{jl}) in D_j that exist in Domain Dictionary, where $1 \le k \le p$, $1 \le l \le q$.
Step 7: Calculate the total number of words as N1 and N2 in D_i and D_j that matches with Domain Dictionary.
Step 8: Perform the Chi Square test between Di and Dj through the following steps:
i) Calculate the term frequency between the common words in D_i and D_j.
ii) Assume the related words TF (W_{ik}) of common words in D_i as $A=\sum TF(W_{ik})$, Similarly as $C=\sum TF(W_{il})$ for the Document D_j.
iii) Assume Observed Values (O) as A, B, C, D, where B= N1-A and D = N2 - D.
iv) Compute M1, M2, N as M1=A+C, M2=B+D, N=N1+N2.
v) Similarly Calculate the Expected Values(E) as
P = ((M1)*(N1))/N,
Q = ((M2)*(N1))/N,
R = ((M1)*(N2))/N,
S = ((M2)*(N2))/N.
vi) Calculate O, E, O – E, $(O-E)^2$, and $((O-E)^2)/E$
vii)Compute Chi Square= $\sum ((O-E)^2)/E$.
Step 9: Calculate the degree of freedom (ndf) as (R-1)*(R1-1).
Where
R – Row in expected frequency table
R1 – Column in expected frequency table
Step 10: See the degrees of freedom value in the Chi Square distribution in 5% level of significance.
Step 11: If the value of Chi Square is greater than Chi Square distribution then
The Document D_i and D_j are Relevant.
Else
The Document D_i and D_j are Irrelevant.
Step 11: Increment j, and repeat from step 5 to step 9 until $j \le N$.
Step 12: Increment i, and repeat from step 4 to step 10 until $i < N$.

V. Experimental Results for Proportions Calculation

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.

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(proportions)

	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 lesser than 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.

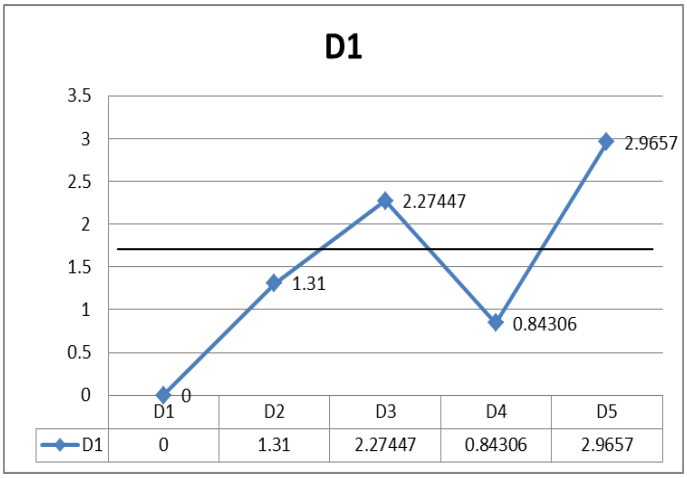

Fig 3:Graphical Representation of proportions calculation for document D1.

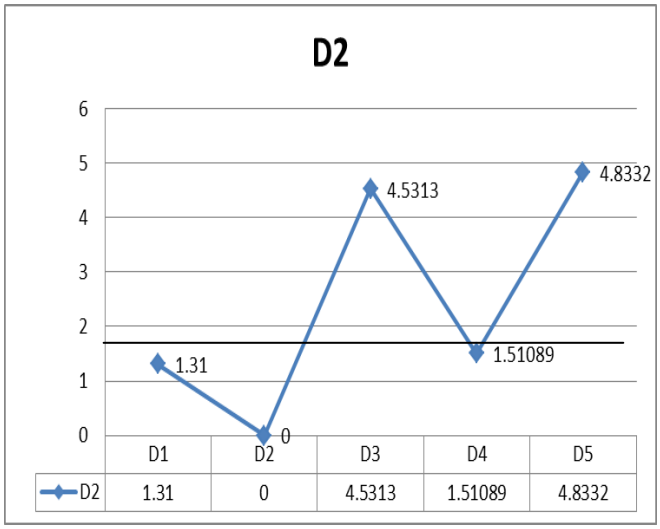

Fig 5:Graphical Representation of proportions calculation for document D2

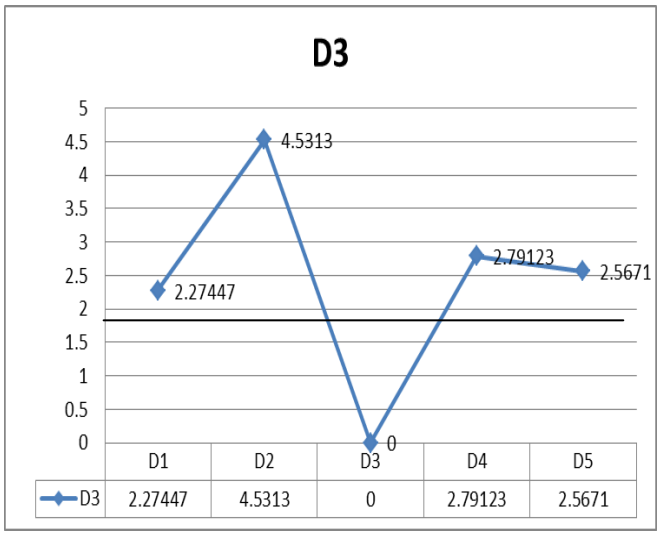

Fig 7:Graphical Representation of proportions calculation for document D3 for 3.

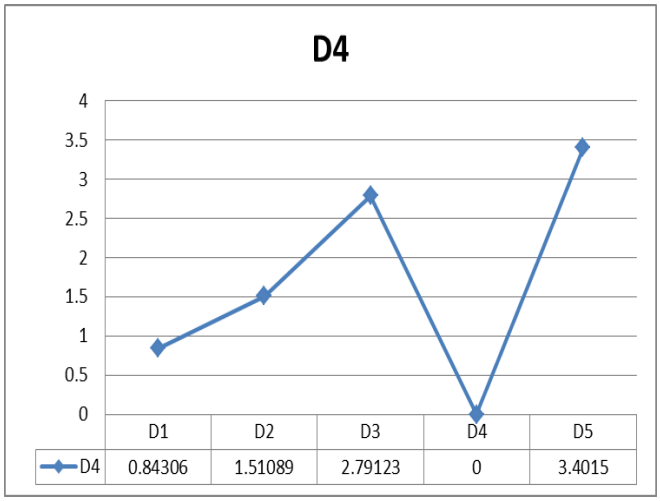

Fig 9:Graphical Representation of proportions calculation for document D4

VI. Experimental Results for Chi-Square Test

Similarly like proportions calculation, the relevancy for the documents is computed. In this approach the degrees of confidence at 95% level which holds the value 3.841 is obtained from the chi-square table. The chi test value for the input documents is computed in table 3.

Table 3: Experimental Results (Chi-Square)

	D1	**D2**	**D3**	**D4**	**D5**
D1	*	1.7161	5.4290	0.71217	9.0079
D2	*	*	22.0696	3.2483	23.2509
D3	*	*	*	8.3902	0.3214
D4	*	*	*	*	11.9084
D5	*	*	*	*	*

From the table 3, it is clear that documents 1, 2 and 4 are lesser than 3.841. 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

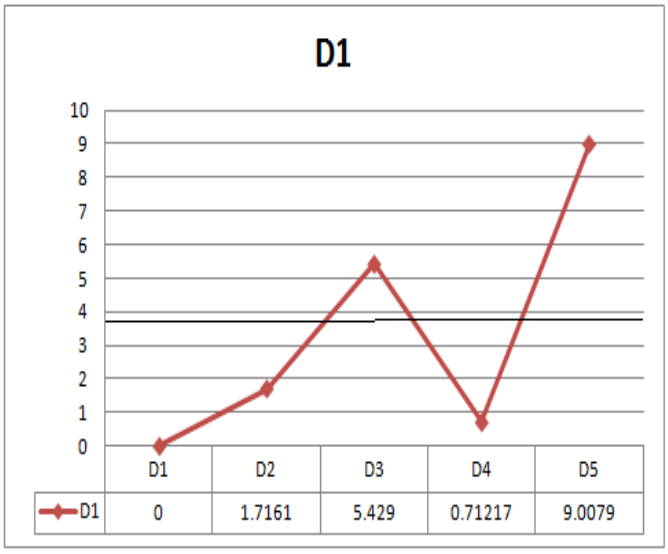

Fig 4:Graphical Representation of chi square test for document D1

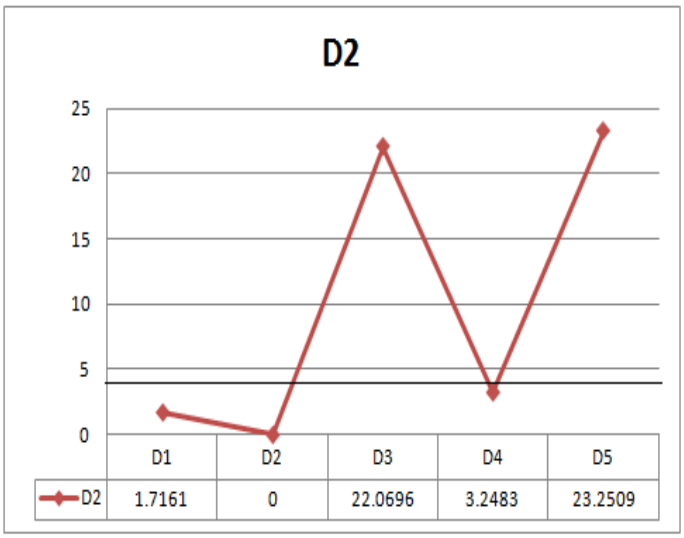

Fig 6:Graphical Representation of chi square test for document D2

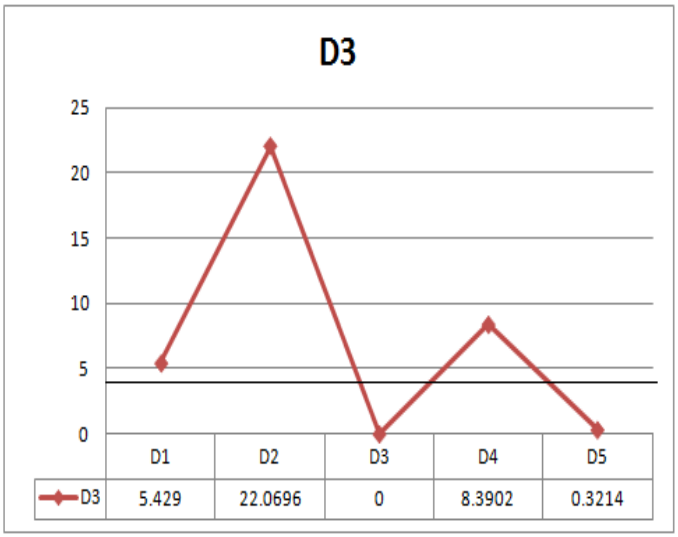

Fig 8:Graphical Representation of chi square test for document D3

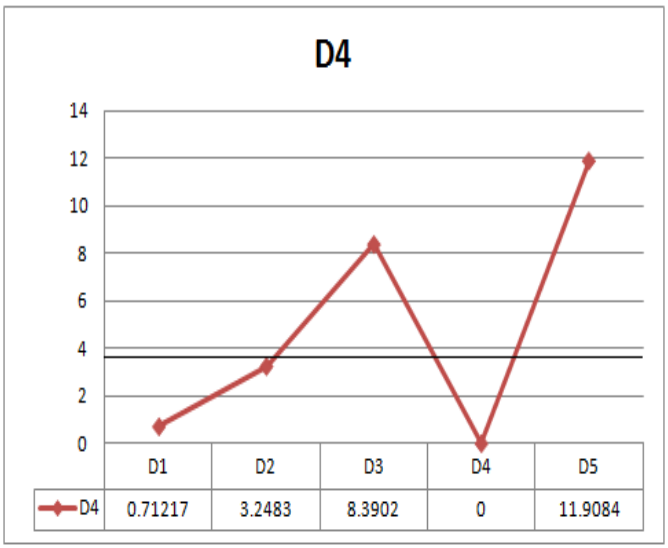

Fig 10:Graphical Representation of chi square test for document D4

VII. Comparative Study

In this paper, comparison between two statistical methods for computation of relevance is performed. The Chi-Squared Test of Association allows the comparison of two attributes in a sample of data to determine if there is any relationship between them. The idea behind this test is to compare the experimental frequencies with the frequencies that would be expected if the null hypothesis of no association / statistical independence were true. If the value of the test statistic for the chi-squared test of association is too large, it indicates a poor agreement between the observed and expected frequencies and the null hypothesis of independence is rejected. A proportion is another statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Due to the central limit theorem, many test statistics are approximately normally distributed for large samples. Therefore, many statistical tests can be performed as approximate Z-tests if the sample size is not too small. From the table it is observed that calculation of precision for both the algorithm. The statistical table shows the 100% precision for relevance. But the chi square table shows variation in precision and it is observed it is not 100%.

VIII. Precision Calculation

Precision → It is the ratio between the number of relevant documents returned originally and the total number of retrieved documents returned after eliminating irrelevant documents. Here the relevant documents indicate the required documents which satisfy the user needs.

$$\text{Precision} = \frac{\text{Relevant} \cap \text{Retrieved originally}}{\text{Retrieved after refinement}}$$

Table 4: Precision Table for the Statistical Approach. (Proposnate Calculation)

Dataset size	Relevant documents through Proposnate computation	Relevant documents computed Manually	Precision
5	3	3	**100%**
10	7	8	**87.5%**
15	10	12	**83.3%**
20	13	16	**81.25%**
25	19	23	**81.81%**

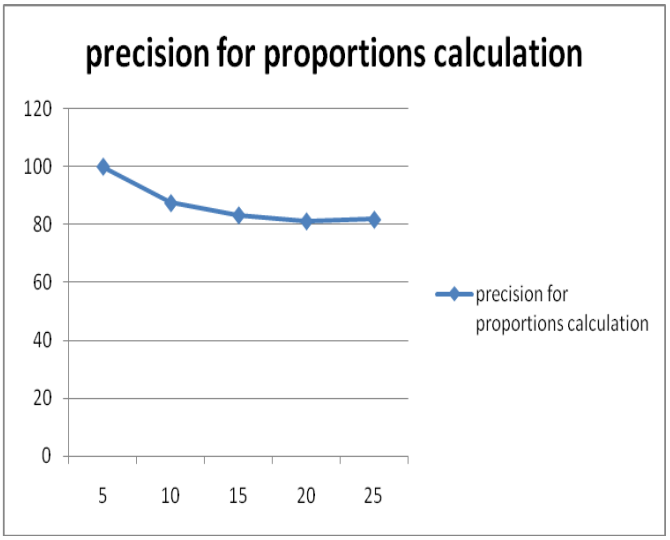

Fig 11. Precision for proportions calculation

Table 5: Precision Table for the Statistical Approach. (Chi Square test)

Dataset size	Relevant documents through Chi Square computation	Relevant documents computed Manually	Precision
5	3	3	**100%**
10	6	8	**85.71%**
15	9	12	**75.20%**
20	11	16	**75%**
5	17	23	**73.91%**

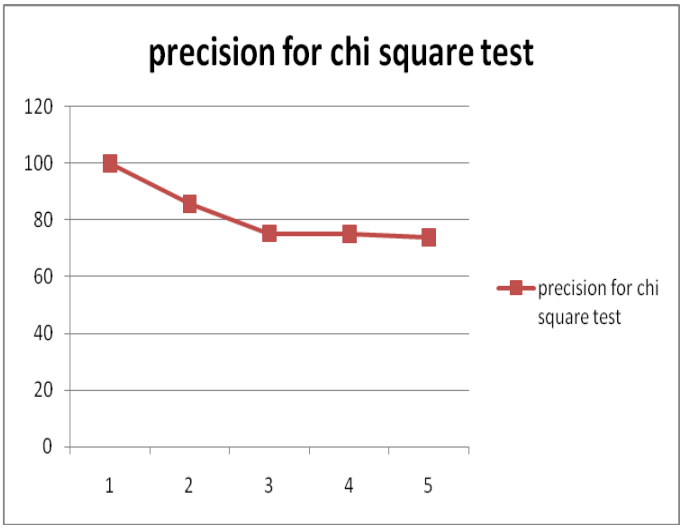

Fig 12: Precision for Chi-Square Test.

IX. 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 and chi-square test with 95% level of confidence for retrieving relevant web documents from structured as well as unstructured documents. Comparison of both the algorithm is performed. It is observed that statistical method has higher precision than chi-square test. And also the statistical method which gives results for large samples whereas the chi-square works for small samples.

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 and Rada S. Alhajj, Framework for Mining Web Content Outliersb. *In: ACM Symposium on Applied Computing*, Nicosia, Cyprus, 2004, pp 590-594.

[6] Malik Agyemang, Ken Barker and Rada S. Alhajj Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams' *ACM Symposium on Applied Computing.*, Santa Fe, New Mexico,2005, pp 482-487.

[7] Malik Agyemang Ken Barker and Rada S. Alhajj, Hybrid Approach to Web Content Outlier Mining without Query Vector. *Springer –Berlin*, 2005,Vol. 3589.

[8] 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.

[9] 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.

[10] K. Thiagarajan, A. Raghunathan, Ponnamal Natarajan, G. Poonkuzhali and Prashant Ranjan, Weighted Graph Approach for Trust Reputation Managements, *International Conference on Intelligent Systems and Technologies*, Published in Proc. Of World Academy of Science and Technology- Vol 56, 2009,pp-830-836

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

[12] 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.
[13] Statistics(Theory, Methods & Application) By D.C.Sancheti and E.K.Kapoor Published by Sultan Chand and Sons, Sixth thoroughly revised Edition, 1990.
[14] N.P. GOPALAN, J. AKILANDESWARI: Distributed, Fault-tolerant Multi-agent Web Mining System for Scalable Web Search 5th WSEAS Int. Conf. on APPLIED INFORMATICS and COMMUNICATIONS, Malta, September 15-17, 2005 (pp384-390).
[15] G. CASTELLANO, A. M. FANELLI, M. A. TORSELLO : Mining usage profiles from access data using fuzzy clustering, 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal, September 22-24, 2006.
[16] Anoop Paharia, Yachana Bhawsar, Yachana Bhawsar:Developing Web intelligence using data mining 6th WSEAS Int. Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, Tenerife, Spain, December 14-16, 2007.
[17] Web Document Classification and its Performance Evaluation: IOAN POP, 9th WSEAS International Conference on EVOLUTIONARY COMPUTING (EC'08), Sofia, Bulgaria, May 2-4, 2008
[18] Advanced AI Techniques for Web Mining: IOAN DZITAC, IOANA MOISILMATHEMATICAL METHODS, COMPUTATIONAL TECHNIQUES, NON-LINEAR SYSTEMS, INTELLIGENT SYSTEMS.
[19] ZAKARIA SULIMAN ZUBI, Using Some Web Content Mining Techniques for Arabic Text Classification, RECENT ADVANCES on DATA NETWORKS, COMMUNICATIONS, COMPUTERS.

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

R.Kishore Kumar currently under-graduate student of Rajalakshmi Engineering College. He has presented 7 papers in conferences and published 4 research papers in international journals and 3 papers in national journals. Also won the best paper award in National conference at Knowledge Utsav. 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 has presented 3 papers in International conference. Also won the best paper award in National conference at Knowledge Utsav. 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.