Content Based Ranking for Search Engines

P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, *Member IAENG*

***Abstract*— In today's e-world search engines play a vital role in retrieving and organizing relevant data for various purposes. However, in the real ground relevance of results produced by search engines are still debatable because it returns enormous amount of irrelevant and redundant results. Web content mining and Information retrieval is an ample and powerful research area in which retrieval of relevant information from the web resources in a faster and better manner. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents. In this research work, a novel approach using weighted technique is introduced to mine the web contents catering to the user needs. Experimental results prove that the performance of the proposed approach in terms of precision, recall and F-measure is high when compared to other search engine results.**

***Index Terms*— Content mining, Mathematical approach, Relevant Information, Web page ranking**

I. INTRODUCTION

World wide web plays a starring role for retrieving user requested information from the web resources. In-order to retrieve user requested information, search engine plays a major role for crawling web content on different node and organizing them into result pages so that user can easily select the required information by navigating through the result pages link. This strategy worked well in earlier because, number of resources available for user request is limited. Also, it is feasible to identify the relevant information directly by the user from the search engine results. When the Internet era increases, sharing of resource also increases and this leads to develop an automated technique to rank each web content resource. Different search engine uses different techniques to rank search results for the user query. This leads to business motivation of bringing up their web resource into top ranking position. As the competition and web resource increases, ranking of web content become tedious and dynamic with respect to user query.

This also affects user interest on looking for search engines to identify the web content relevant to their needs. So A noval approach to be developed to work towards ranking content of the web resource based on user query.

F. P.Sudhakar is with the Department of Computer Science and Engineering, Kamaraj College of Engineering and Technology, Virudhunagar, Affiliated to Anna University, Tamil Nadu, 626001 – India. (email : sudhakar.asp@gmail.com)

S. G.Poonkuzhali is with the Depatment of Computer Science and Engineering, Rajalakshmi Engineering College, Affiliated to Anna University, Chennai, India (email:poonkuzhali.s@rajalakshmi.edu.in)

T. R.Kishore Kumar is with the Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, SSN Nagar, Chennai 603110 – India (email rskishorekumar@yahoo.co.in)

In the proposed work a new approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and page ranking provided by search engines. Based on the user query, search engine results are retrieved. Every result is individually analyzed based on keywords and content. User Query is pre-processed to identify the root words. Every root words are considered for Dictionary construction and Dictionary is built with synonyms for the user query. Every result page keywords and content words are pre-processed and compared against the dictionary. If a match is found then particular weight is awarded to each word. Finally, the total relevancy of the particular link against user request is computed by summarizing all the weights of the keyword and content words. The page which contains total relevancy value nearest to 1 are ranked as first page and 0 are ranked as last page.

Outline of Paper

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

II. RELATED WORKS

Due to the heterogeneity of network resources and the lack of structure of web data, automated discovery of targeted knowledge retrieval mechanism is still facing many research challenges. Moreover, the semi-structured and unstructured nature of web data creates the need for web content mining. In Paper [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 given query was determined in paper[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

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

user does not need to visit the homepages of companies to get the information about any product, instead the user write the name of the 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.

Nowadays, most of the people rely on web search engines to find and retrieve information. When a user uses a search engine such as Yahoo or Google or bing to seek specific information, an enormous quantity of results are returned containing both the relevant document as well as outlier document which is mostly irrelevant to the user. Therefore discovering essential information from the web data sources becomes very important for web mining research community.

Chakrabarti et al (1999) describes a new hypertext resource discovery system called focused crawler which analyze its crawl boundary to find the links that are likely to be most relevant for the crawler and avoids irrelevant regions of the web. Mei Kobayashi and Koichi Takeda (2000) discussed the development of new techniques targeted to resolve some of the problems such as slow retrieval speed, noise and broken links associated with web based information retrieval and speculates on future needs. Mayfield et al (1998) explores the indexing using both N-grams and words by using HAIRCUT (Hopkinks Automated Information Retrieving for Combing Unstructured Text) System. Junghoo Cho et al (2000) present the efficient method for identifying replicated document collections to improve web crawlers, archivers and ranking functions used in search engines. Sungrim Kim and Joonhee Kwon (2009) propose an information retrieval method using the context information on the web 2.0 environment by adopting page rank and context tags algorithms. Brin et al (1998) gives an in-depth description of large scale web search engines and described the page rank algorithm. The algorithm states that the relevance of a page increases with the number of hyperlinks to it from other relevant pages. Bin et al (2003) explained web mining process and the Taxonomy of web mining. Georgioes (2007) provide an overview of web mining and the latest developments on web mining application in beneficial to society.

III. ARCHITECTURAL DESIGN

Architecture of the proposed work uses the advantage of full word matching against Dictionary. User request is processed for search engine to obtain the results. Search results are extracted and sent for pre-processing. Pre-Processing is an important step in text based mining. Real-world data tend to be dirty, incomplete and inconsistent. Data pre-processing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data pre-processing is an important step in the knowledge discovery process, since quality decisions must be based on quality data. All user query, keywords and content words are preprocessed to remove noisy words. After pre-process, Dictionary is built for user query with related words (synonyms). Every result of the keywords and content words are compared against dictionary by full word matching. If a match is found then a point is awarded to each words based on their position (keyword / content) using weighted technique. Finally all matched keywords and content words are summarized and normalized so that the cumulative total must be less than or equal to 1.

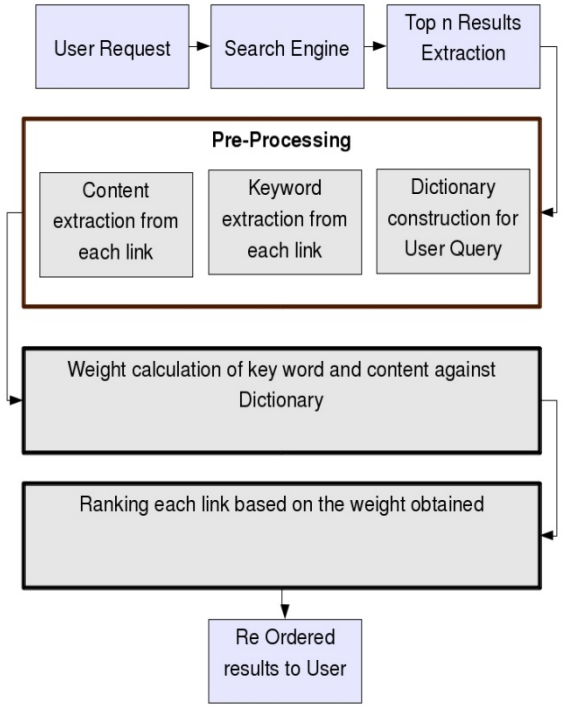

Fig 1. Architecture design

At last, the normalized value of each result is sorted in descending order to get the most relevant content for the user query. Re-ordered results are sent back to the user so that the top most page is more relevant for the user query.

IV. ALGORITHM FOR MINING WEB CONTENT

Algorithm : *Relevancy and Weight based approach*
Input : *Extracted Web Contents*
Output : *Reordered Web Content*

Step 1. Extract Search Engine results SR_i for the user query where $1<i<N$

Step 2. Pre-process user query and extract root words RW_j where $1<j<N$

Step 3.Construct Dictionary D for the user query RW_j

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

Step 4. Extract and Pre-process the keywords KWi for the search results SR_i

Step 4.a. Compute Keyword Strength

$$S(KW_i) = \frac{1}{\sum KW_i}$$

Step 5. Extract and Pre-process the Content words CW_i for the search results SR_i

Step 5.a: Compute Content Words

$$\text{Strength } S(CW_i) = \frac{1}{\sum CW_i}$$

Step 6. Compare each keyword KW_i against Dictionary D.

Step 6.a. If match is found then award strength $S(KW_i)$ to particular keyword

Step 6.b. Else award 0 as a strength for particular keyword.

Step 7. Compare each content word CW_i against Dictionary D

Step 7.a If match is found then award Strength $S(CW_i)$ to particular content word

Step 7.b. Else award 0 as a strength for particular content word.

Step 8. Calculate Total Strength for Keyword

$$TKS(SR_i) = \sum S(KW_i)$$

Step 9. Calculate Total Strength for Content Word

$$TCS(SR_i) = \sum S(CW_i)$$

Step 10. Compute Total Relevancy for the particular link

$TR_i = TKS(SR_i)*(Wt) + TCS(SR_i)*(1- Wt)$ where $0 < Wt < 1$

Step 11. Repeat step 4 to 10 for all Search Results (SR)

Step 12. Sort the result set SR based on TR_i in Descending order.

The Topmost Search Result SRi is the most relevant for the user query and bottom most search result is the least relevant for the User query.

V. EXPERIMENTAL RESULTS

Experiment is conducted with a generic user query ("human survival in society") against specific search-engine. Top 10 web pages from that search-engine are taken as an input dataset and are listedin Table I.

TABLE I.
INPUT DATA SET

S.No	Document Id	URL
1	SR1	earthfamilyalpha.blogspot.com/.../respectism-and-human-survival.ht...
2	SR2	en.wikipedia.org/wiki/Herbert_Spencer
3	SR3	en.wikipedia.org/wiki/Survival_of_the_fittest
4	SR4	listverse.com/2010/08/30/8-ways-to-ensure-human-survival/
5	SR5	science.jrank.org/.../Social-Darwinism-Human-Nature-Struggle-Survi...
6	SR6	www.personalityresearch.org/papers/smith.html
7	SR7	en.wikipedia.org/wiki/Human_extinction
8	SR8	kauilapele.wordpress.com/.../ben-fulford-7-11-11-secret-societies-an...
9	SR9	ezinearticles.com › Self Improvement › Personal Growth
10	SR10	http://wiki.answers.com/Q/FAQ/2872-61

Keyword and content based ranking approach is applied and results are listed in TABLE II.

TABLE II.
RELEVANCY RANKING

Document	Total Relevancy	Rank
SR5	0.28	1
SR4	0.25	2
SR2	0.25	3
SR3	0.25	4
SR1	0.17	5
SR8	0.1	6
SR9	0.03	7
SR6	0.03	8
SR7	0.01	9
SR10	0.01	10

From the TABLE II results, it is understood that if the total relevancy value is high, it is ranked as first and vice-verse. The same set of document is given to different users to compare the system results against user ranking. These results are discussed in section 6.

VI. PERFORMANCE EVALUATION

Performance evaluation of the proposed approach is done based on classification context scenario. Precision, Recall, Accuracy and F_1 - Score plays a major role in classification

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

based performance. Precision measure is calculated based on the formula Precision = $\frac{tp}{tp+fp}$

Recall is calculated based on the formula

Recall = $\frac{tp}{tp+fn}$

Accuracy is calculated based on the formula

Accuracy = $\frac{tp+tn}{tp+tn+fp+fn}$

where

tp – True Positive (*Correct result*)
tn – True Negative (*Correct absence of result*)
fp – False Positive (*Unexpected Result*)
fn – False Negative (*Missing result*)

F-Measure is calculated based on the formula

F= $2.\frac{Precision\,.\,Recall}{Precision+Recall}$

In this proposed work sample Dataset TABLE I. is consider for evaluation purpose and top 10 documents that are more relevant to the user based on user decision is classified manually with different users . Now the same relevant dataset is evaluated against retrieved dataset. Comparison results of the proposed approach are given in the TABLE III.

TABLE III.

RANKING COMPARISON

Document	Search engine ranking	Manual Ranking	Proposed approach ranking
SR1	1	5	5
SR2	2	9	3
SR3	3	4	4
SR4	4	2	2
SR5	5	1	1
SR6	6	3	8
SR7	7	8	9
SR8	8	6	6
SR9	9	7	7
SR10	10	10	10

TABLE III represents the matching of manual ranking against proposed approach ranking. Document SR2, 6, 7 represents the mismatching of manual ranking against proposed approach. From the table, it is understood that precision of the proposed system is 0.7 out of 1 where as search-engine precision is 0.1 out of 1.

TABLE III contains result for evaluating the proposed approach against various performance measures like Precision, Recall, Accuracy and F-Measure. The results of the performance measure are plotted in Fig.2.

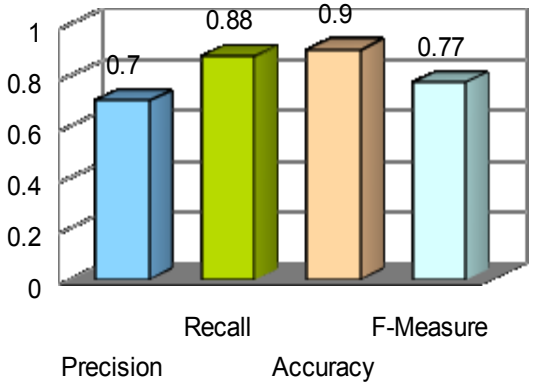

Fig 2. Performance of Proposed system

From the performance measure, it is understood that the accuracy of the proposed system is 90% which is high compared with existing approaches.

VII. CONCLUSION

The Proposed approach gives far better results compared with search-engine ranking. However, more fine tuning process to be needed to bring the best result. Proposed methodology focus only on text based mining to rank the relevancy of the web pages where nowadays relevant information may be available in any format like images, audio and video files. Forth coming research work will focus on all types of data sets.

ACKNOWLEDGMENT

The authors would like to thank Dr.K.Sarukesi, Vice-chancellor, Hindustan University, & Advisor, Kamaraj College of Engineering and Technology for his intuitive guidelines and fruitful discussion with respect to this paper contribution.

REFERENCES

[1] Bing Liu, Kevin Chen- Chuan Chang ," Editorial: Special issue on Web Content Mining" , *SIGKDD Explorations*, Volume 6, Issue 2.
[2] Bin W, LiuZhijing, Web Mining research, *5th International Conference on computational Intelligence and Multimedia Applications*, 2003
[3] 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.
[4] Chakrabarti, S. "Mining the Web: Discovering Knowledge from Hypertext Data", *Morgan-Kauman Publishers*,2002.
[5] Chakrabarti, S., Berg, M., and Dom, B. Focused Crawling: A New Approach to Topic-specific Web Resource Discovery. *Computer Networks, Amsterdam, Netherlands*, 1999.
[6] 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.
[7] 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.
[8] Georgies Lappas, An overview of web mining in societal benefit areas, *The 9th IEEE International Conference on E-Commerce Technology, IEEE* 2007.

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

[9] 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.
[10] Han, J. and Kamber, M. "Data Mining: Concepts and Techniques", *Morgan Kaufmann Publishers*,2001.
[11] 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.
[12] Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An approach to the Web Content Mining",
[13] Junghoo Cho Narayanan Shivakumar Hector Garcia-Molina, Finding replicated web collections , *MOD* 2000, Dallas, TX USA
[14] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD, July 2000, Vol-2, pp 1-15.*
[15] 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.
[16] Mayfield, J., McName, P. Indexing Using Both N-Grams and Words. In proceeding of NIST Special Publication 500 - 242: *The Seventh Text Retrieval Conference (TREC 7)*, 1998, pp 419 -224
[17] 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
[18] Mei Kobayashi and Koichi Takeda, Information Retrieval on the Web , *ACM Computing Surveys*, Vol. 32, No. 2, June 2000
[19] 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.
[20] 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
[21] 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.
[22] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD*, July 2000, Vol-2, pp 1-15.
[23] Shohreh Ajoudanian, and Mohammad Davarpanah Jazi," Deep Web Content Mining", *World Academy of Science, Engineering and Technology*,49 2009.
[24] Sungrim Kim and Joonhee Kwon, 'Information Retrieval using Context Information on the Web 2.0 Environment', IJCSNS International Journal of Computer Science and Network 62 Security, VOL.9 No.10, October 2009.

P.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.

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.

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