CONTENT EXTRACTION USING OPTIMAL CLUSTERING WITH MULTIVIEW POINT APPROACH

[1]Pradeep Kumar Sahoo, [2]Dr.S.P.Rajagopalan
[1]Research Scholar, Department of Computer Science and Engineering
Anna University, Chennai, Indias
[2]Professor, Department of Computer Science and Engineering
GKM College of Engineering and Technology, Chennai, India
Email: [1]pradeepsahoo.cs@gmail.com, [2]sasirekaraj@yahoo.co.in

Abstract:

In today's world information extraction from web play a vital role for getting 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 such as optimal clustering with multiview point approach. In Web content mining this approach 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-score is high when compared to related technique for information retrieval 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 the user requested information, search engines
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 page links. 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 novel approach to be developed to work towards ranking content of the web resource based on user query.

In the proposed work a new approach is introduced to rank the relevant pages based on clustering 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. Using web crawlers the information is retrieved. 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. According to the keyword the document is clustered. 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

We can differentiate web content mining from two different points of view. Information Retrieval view and Database view. Characteristics of dissimilarity measure on web content mining presented in [1]. 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 . 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 [3]. 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 [4]. 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 [8]. A framework is proposed to provide facilities to the user during search . In this framework 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 frame work searches all the available web pages related to the text, and the user gets the information with little efforts.

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 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.

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 [5]. 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. The indexing using both N-grams and words by using HAIRCUT (Hopkinks Automated Information Retrieving for Combing Unstructured Text) System. The efficient method for identifying replicated document collections to improve web crawlers, archivers and ranking functions used in search engines. The algorithm states that the relevance of a page increases with the number of hyperlinks to it from other relevant pages.

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.

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.

Table-I

Notation	Description	Notation	Description
N	Document count	ER	Extracted content results
P	Precision	S(CWi)	Content word strength
R	Recall	TKS	Total strength of keyword
D	Dictionary	TCS	Total strength of content
S(KWi	Keyword strength	Wt	Weight threshold

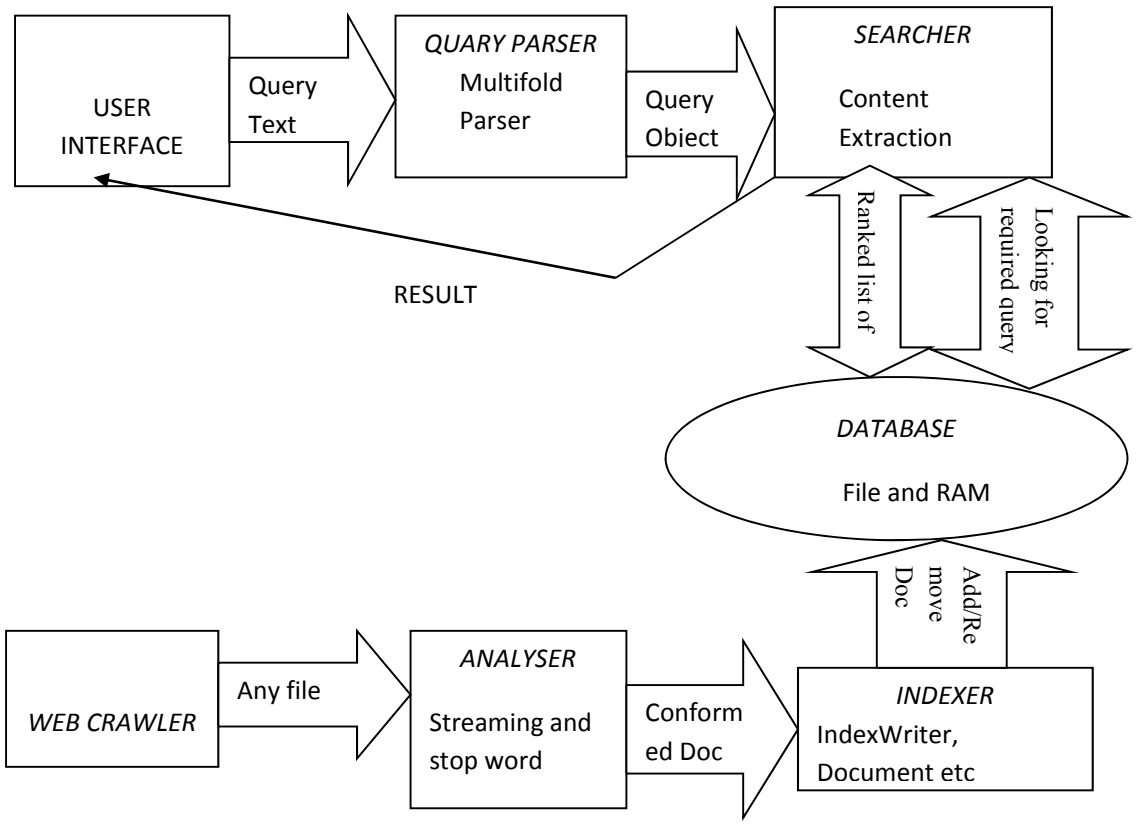

Fig: 1Architecture Diagram

Table-II(INPUT DATASET)

Sl.No	*Document Id*	*URL*
1.	D1	earthfam/…./human-survival.ht
2.	D2	en-wikipedia.org/hervert-spencer
3.	D3	en-wikpedia.orgi-survival_of_fittest
4.	D4	listverse.com/./ensure-human-survival
5.	D5	science.jrank.org/../nature-of- survival
6.	D6	www.per.research.org/papers/smith.html
7.	D7	en,Wikipedia.org/human-extinction
8.	D8	www.worldpress.com/../secreat-socity-en
9.	D9	eznerticles.com/…/personal-groth
10.	D10	http:/wikianswear.com/qfaq/23458-6

Table-III(Relevancy Ranking)

Document	*Total Relevancy*	*Rank*
D5	0.28	1
D4	0.26	2
D2	0.25	3
D3	0.25	4
D1	0.17	5
D8	0.1	6
D9	0.03	7
D6	0.03	8
D7	0.01	9
D10	0.01	10

Keyword and content based ranking approach is applied and results are listed in Table-II

IV. ALGORITHM FOR MINING WEB CONTENT

Algorithm: Relevancy and Weight based approach

Input : Extracted web content

Output : Recorded web content

Step 1. Get Extracted content results ER_i for the user query where 1<i<N

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

Step 3.Construct Dictionary D for the user query RW_j

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

Step 4.a. Compute Keyword Strength

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

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

Step 5.a: Compute Content Words

$$\text{Strength } S(CWi) = \frac{1}{\sum_{CWi}}$$

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 $S(CW_i)$

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

Step 8. Calculate Total Stregtnh 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

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

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 listed in Table II.

From the TABLE III 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 and F-Score plays a major role in text data extraction for performance measure.F-score is an equally weighted combination of precision (P) and recall(R) values used in text retrieval.

$$FScore = \sum_{i=1}^{k} \frac{n_i}{n} \max_j (F_{i,j})$$

$$\text{where } F_{i,j} = \frac{2 \times P_{i,j} \times R_{i,j}}{P_{i,j} + R_{i,j}}; P_{i,j} = \frac{n_{i,j}}{n_j}, I$$

Where n_i denotes the number of documents in class i , n_j denotes the number of document assigned to cluster j , $n_{i,j}$ the number of document shared by class i and cluster j

In this proposed work, sample Dataset in TABLE II. 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 IV.

TABLE IV 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 in term of F-score in the graph.

TABLE IV contains result for evaluating the proposed approach against various performance measures like Precision, Recall and F-Measure. The results of the performance measure are plotted in Fig.2.The red line shows the proposed system and the blue line shows the existing system in the graph of Precision vs Recall in term of f-score.From the performance measure, it is understood that the accuracy of the proposed system is superior which is high compared with existing approaches as precision is high when recall is low.

Table.III(Ranking comparision)

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

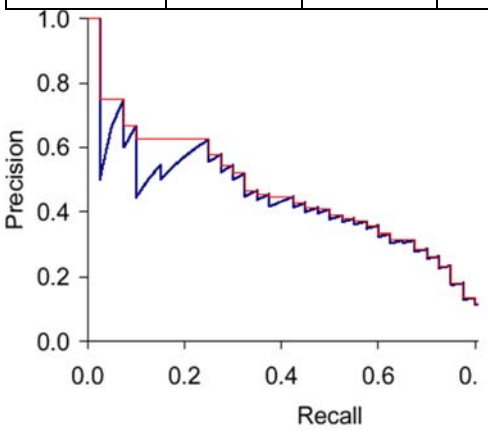

Fig 2. Performance of Proposed system

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 as well as videos and images. Forth coming research work will focus on all types of data sets.

ACKNOWLEDGMENT

The authors would like to thank Dr.S. P.Rajagopalan, Professor, GKM Engineering College &Technology, and Chennai for his intuitive guidelines and fruitful discussion with respect to this paper contribution.

REFERENCES

[1] D.Lee,J.Lee,(2010)"Dynamic Dissimilarity Measure for Support Based Clustering".IEEE Trans.Knowledge and data eng vol.22,no.6,pp.900-905.

[2] Duc Thang Nguyen, Lihui Chen,(2012)"Clustering with multiview similarity".vol 24,No.6,pp.988-1001.

[3] Bo Geng, Linjun Yang,Chao Xu,Xian-Sheng Hua,(2012)."Ranking model adaptation for specific search".vol 24,No 4.pp.745-758

[4] A.Ahmed and L.Dey,"A Method to Compute Distance Between Two CategoricalValues of Same Attribute in Unsupervised Learning for Categorical Data Set",(2007)Pattern Recognition Letters,vol.28,no.1,pp.110-118.

[5] Chakrabarti, S., Berg, M., and Dom, B(1999). "Focused Crawling: A New Approach to Topic-specific Web Resource Discovery". NY, USA.1623-1640pp.

[6] Etemadi R., Moghaddam N. (2010)"An approach in web content mining for clustering web pages",ICDIM'13,pp.279-284.

[7] Al-Ghuribi S.M., Alshomrani S.(2013)"A comprehensive survey on web content extraction algorithms and techniques", ICISA'13, pp.1-5.

[8] Guowei Chen, Pengzhou Z.(2012),"The Content Extraction Method of Webpage Information Based on Knowledge Base", CSO'12, pp.623-626.