OCS - A System for Optimizing, Clustering and Summarizing Web Search Results Using Intelligent Agents

Y. Nawaz Ahamed Khan[a], Bhaskaran Raman[a] and Shanmugasundaram Hariharan[b]
[a]Student, [b]Assistant Professor
Department of Information Technology
School of Computer and Information Sciences
B.S.Abdur Rahman University, Chennai-48, Tamilnadu, India
nawazahamedkhan@gmail.com, ramancre_89@yahoo.co.in,
shari1981@rediffmail.com

***Abstract*-Internet has brought a major revolution in social community especially among researchers. Number of commercial search engines has emerged with lots of information available online to end user. A web surfer who wishes to surf the contents available online faces quite large number of problems. This paper in turn addresses three aspects, which we call term as OCS system. The proposed system or study analyzes the duplicates occurring in the content and link levels at first step, there by producing an optimized result. At second level, the optimized contents are clustered using top frequent clustering approach by identifying an optimal threshold. Finally the clustered contents are summarized using extraction process at query level and anchor text level. We take Google search engine and the results given by them for our case study to implement the system effectively.**

Keywords*: *Link mining, content mining, clustering, summarization, similarity and optimization, search result, google search engine.

I. INTRODUCTION

The World Wide Web is considered to be the largest form of online source accessible across the globe. The Web has created new challenges in Information Retrieval and other tasks like question answering, summarization etc. Various information sources are accessible through the web as electronic content. These online contents are usually presented in a list commonly called as hits. End users or web surfers who make simple queries or complex queries on various topics launch internet searches. Such queries generates thousands of hits and sometimes even beyond this. The challenge remains in ranking of a document's relevance based on user's query.

Search engines are programs that retrieve documents or files or data from wide database across networks. Web search engine is a tool designed to search for information on the World Wide Web. Information may consist of web pages, images, information and other types of files. Some search engines also mine data available in news books, databases, or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically as a mixture of algorithmic and human input. A search engine is a coordinated as a set of programs that creates a huge index from the pages that have been read or that receives your search request, compares it to the entries in the index and returns the results back to the users. These search engines ranges from single standard search engines [14] to Meta search engines [13].

Most Web search engines respond to a user's query by returning a list of links that are deemed relevant. The majority of these queries consist of only a few keywords, which are often ambiguous or too general in expressing the user's information need [9]. As a consequence, typically only small fractions of the search engine results are relevant. This paper solely focuses its attention on three aspects namely reranking of search results that would improve accuracy, clustering the optimized results and finally producing an effective summary to end user.

Traditional information retrieval theory offers some models like Vector Space Model, probabilistic models and fuzzy logic models for measuring the similarity between user-defined keywords and document contents. Almost all models depend on the frequency of query terms in a given document (term weights can be normalized based on document length). However, our focus will be on Vector space model only.

The web has created new challenges in Information retrieval. Huge amount of information on the web due to the rapid growth, increased number of users day to day are some factors that alerts the web researchers. People are likely to surf the web using popular search engines like Google, Yahoo, Ask jeeves etc. Google now has many versions running in many different countries, including China, Japan, the U.K., Hong-Kong and many others. Commercial search engines are those, which retrieve pages based on the user request. From survey it is found that Google search engine has the highest number of users in the world. The reasons why a user might choose appropriate search engine over another is dependent on the complexity, speed, ease of use and readability etc. The most

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

important criterion seems to be that of the relevance of the results to the search performed – at least in the way they are perceived as relevant by the user [10]. Also Google accounts for more than 85 percent of all Internet searches on a daily basis [12].So we focus our system attention taking Google search engine and the results retrieved by them for our study throughout the paper.

There are even few similar attempts made to rerank the web content to reduce the impact of Search Engine Optimization (SEO) [7], clustering of search results adopting fuzzy ants [9]. Algorithms for clustering web search results try to overcome such problems by converting the output of an existing search engine to a list of labeled clusters. Well-known clustering algorithms such as *k*-means or fuzzy *c*-means [18], ant based clustering [19] are also quiet popular. Our methodology would be to cluster the Web search results based on top frequent terms adopting vector space model. Then we finally propose a method to summarize the clustered contents.

The paper is organized as follows. Section 2 explains the related works carried out; section 3 and related sections briefs the proposed system architecture, corpus used, results including necessary discussions. Finally section 4 gives the conclusions and future improvements.

II. RELATED WORK

Yan Chen et al. [1] designed a concept-based search agent using conceptual fuzzy set (CFS) for matching contexts-dependent keywords and concepts (i.e. a word exact meaning may be determined by other words in contexts). In all possible combinations. To solve the issue of numerous combinations of words appearing in queries and documents, defining the relations between concepts, the authors have proposed a semantic tree (ST) model. Also, user's preferences for personalizing search results were applied. Finally parameters adopting fuzzy logic were used to determining the factors, semantic relations or users' preferences were investigated on the basis of obtained results.

Leung et al [2] in their paper introduced some effective approach capturing the user's conceptual preferences in order to provide personalized query suggestions. The authors developed online techniques that extract concepts from the web-snippets of the search result returned from a query and to use the concepts to identify related queries for that query as a first technique. Second, a new two-phase personalized agglomerative clustering algorithm that was able to generate personalized query clusters was adopted.

Kyung-Joong Kim et al [5] focused their attention on the relevance computing between user's query and the auto-generated text summarization of each webpage. The authors introduced the auto text summarization method based on multi-source integration and the full text of each web page is replaced by its auto-generated abstract to compute the relevance between the webpage and user query. The authors experiment results shows that the ranking results based on the summary generated by our text summarization system with 30% compress ratio can also get 11.29% of the precision improvement for the system.

Li Zhan and Liu Zhijing[6] analyzed the results produced by link-based search engine and text-based search engine. However, the authors found that there is some difficulty in producing the result fit to a specific user's preference and personalization is required. Also a search engine that uses the fuzzy concept network to personalize the results from a link-based search method was adopted based on a user profile, where the system provides a personalized high-quality result.

Lewandowski [8] took five major web search engines like Google, Yahoo, MSN, Askjeeves and Seekport for comparing the effectiveness. Results are judged by the experts using different queries. Out of the five search engines Google and Yahoo performs better and there is no significant differences between them. This study is based on a user model where the user takes into account a certain amount of results and compares results and descriptions systematically and proposes new retrieval measures.

III. PROPOSED SYSTEM

The OCS system designed for optimizing, clustering and summarizing the search results is shown in Figure 1. Initially the search query is given through the search engine (i.e. Google), results are retrieved and processed further. We have modularized the steps involved in our system as optimizer, clustering and summarizer with each component being discussed in detail shortly.

3.1. Corpus Description

Table 1 illustrates 15 queries used, statistics on the link considered, domain knowledge of the query focused, number of links taken from the web page and number of links considered. Also we have shown the percentage of duplicates occurring in the samples chosen for study.

3.2. Optimizer

The first step in our design is to optimize the search results using link and content mining. As inferred from Table 1 it is clearly understood that the user gets large amount of duplicates when he searches the web. To avoid such unnecessary or duplicate data we carry out the following steps.

a. Parsing the html file.
b. Extracting the links that are informative (i.e. result links given by the query)
c. Analysis of links structure (root of the URL) and link optimization.
d. URL extraction after eliminating unwanted links.
e. Downloading the contents to document database.
f. Performing Content level Optimization.

For a sample of 5 test cases considered, relevancy of each document with other varies by significant margin. Table 2 shows the relevance score calculated as the sum of the individual scores of each document with the remaining documents (obtained after step f). Figures 2 and 3 shows

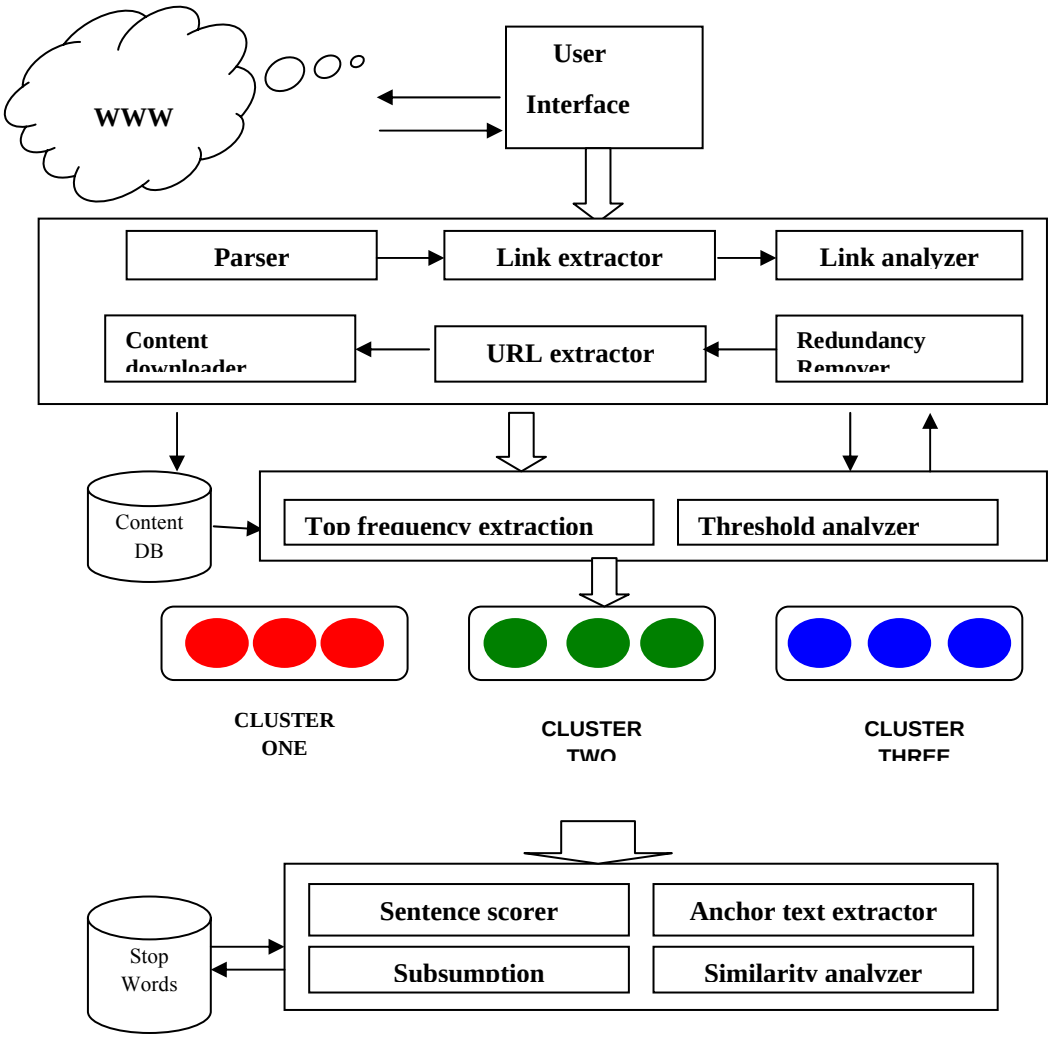

Figure 1: Proposed OCS System

the relevance of links with each other before and after reranking the content. It is inferred from Fig 2 that the content varies for each links. Hence it is essential to rerank the contents based on their informativeness (shown in Figure 3).

3.3. Clustering

Second stage of our system tries to cluster the documents effectively. We have adopted three different approaches namely content based, context based and top frequency based approaches for clustering the documents effectively. For the proposed clustering methods the results were compared with the manually generated test suite collected from different commercially available news sites. By content, context and top frequency we mean to measure the similarity of the entire documents, title and top frequent terms of the documents respectively. Table 3 shows the efficiency calculated as a measure to identify the significance of each method. From Table 3 it is analyzed that threshold of 0.20 is optimal to cluster the documents effectively. Beyond this threshold the number of outliers increases, also top frequency terms is enough to cluster the documents with maximized efficiency and less time compared to content level approach. Context based approach failed to capture the similarity of the documents for most of the cases resulting in poor efficiency. Hence we conclude saying that to frequency based clustering is superior to the other tow methods. The results were tested under the threshold guideline of 0.20 using different document types like text, xml and html contents.

Table 4 gives the details of the number of cluster formed (separated by commas) for each of the query set considered

in Table 1. For each query within the cluster, links were clustered based on the threshold (here it is 0.20) to find the commonality exiting between the documents. We have estimated the commonality between the documents adopting cosine similarity measure, which is found to be superior than several measures existing [11]. The expression for cosine similarity measure is given in expression (1).

$$Cosine(ti, tj) = \sum_{h=1}^{k} t_{ih} t_{jh} \Big/ \sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2} \qquad (1)$$

TABLE 1
STATISTICS OF THE CORPUS USED

Query ID	Cluster	No. of links taken	No. of links considered	Link Approach	Content Approach	% of Duplicates
Q1	C1	10	9	2	2	44.44
Q2		10	10	1	2	30.00
Q 3		10	9	1	2	33.33
Q 4	C2	10	8	2	1	37.50
Q 5		10	8	2	2	50.00
Q 6		10	7	2	2	57.14
Q 7	C3	10	6	1	2	50.00
Q 8		20	18	2	3	27.78
Q 9		20	17	3	5	47.06
Q 10	C4	20	19	1	2	15.79
Q 11		10	8	2	1	37.50
Q 12		30	8	3	3	75.00
Q 13	C5	10	8	3	1	50.00
Q 14		10	8	3	2	62.50
Q 15		20	15	2	4	40.00

TABLE 2
RELEVANCY SCORE BETWEEN THE LINKS IN THE SEARCH RESULTS

Link Number	Case 1	Case 2	Case 3	Case 4	Case 5
	Relevance Score	Relevance Score	Relevance Score	Relevance Score	Relevance Score
1	710	374	448	513	585
2	708	359	455	214	584
3	716	327	352	35	180
4	708	371	448	364	534
5	640	370	339	437	394
6	713	47	372	448	537
7	335	358	334	247	381
8	706	186	263	420	346
9	709	161	174	423	212
10	636	45	273	210	38

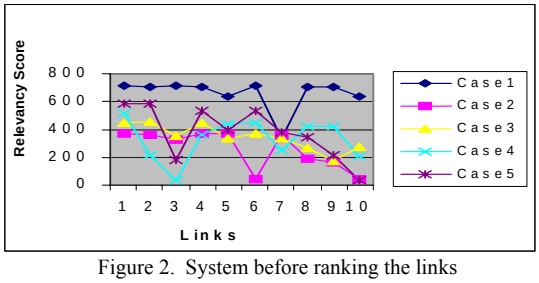

Figure 2. System before ranking the links

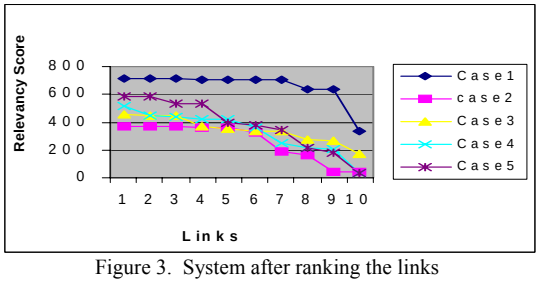

Figure 3. System after ranking the links

3.4. Summarizer

Once the contents have been optimized, finally they were summarized to the end user based on the compression rate using extraction algorithm [15]. We generate the summary based on desired compression ratio using the following steps.

a. Reranking the content based on the relevance.
b. Extraction of anchor text and query terms for assigning special weights.
c. Sentence scoring for the sentences in each document.
d. Eliminating redundancy based on threshold.
e. Generation of summary for each cluster depending on the compression ratio.

TABLE 3

EFFICIENCY OF CLUSTERED RESULTS AT SPECIFIED THRESHOLD FOR VARIOUS METHODS

Cluster ID	Feature Used	No. of clusters Required	t=0.2		t=0.3		t=0.4	
			No .of clusters Identified	Efficiency	No .of clusters Identified	Efficiency	No .of clusters Identified	Efficiency
C1	Content	6	6	82.0	7	74.0	5	43.3
	Context		6	100.0	5	98.0	4	66.6
	Top frequency		6	100.0	7	100.0	4	66.6
C2	Content	7	7	82.0	6	74.0	3	42.8
	Context		7	100.0	5	92.0	3	42.8
	Top frequency		7	100.0	6	98.0	3	42.8
C3	Content	6	6	80.0	6	73.0	3	50.0
	Context		6	100.0	5	98.0	3	50.0
	Top frequency		6	100.0	6	100.0	3	50.0
C4	Content	11	11	80.0	11	70.0	10	90.9
	Context		11	100.0	8	88.0	6	54.5
	Top frequency		11	100.0	11	100.0	10	90.9
C5	Content	8	8	80.0	8	77.0	8	100.0
	Context		8	100.0	6	90.0	6	75.0
	Top frequency		8	100.0	8	100.0	8	100.0
C6	Content	15	15	80.0	12	72.0	12	60.0
	Context		15	100.0	7	80.0	5	33.3
	Top frequency		15	100.0	12	88.0	12	80.0

TABLE 4

CLUSTERING OF DATA SETS FOR EACH QUERY

Query ID	Cluster	Domain	No. of links considered	No clusters formed
Q1	C1	Politics	9	3,3,3
Q2			10	4,3,3
Q 3			9	3,2,2,2
Q 4	C2	Sports	8	4,2,2
Q 5			8	3,3,2
Q 6			7	3,3,1
Q 7	C3	Medicine	6	3,1,2
Q 8			18	6,5,4,3
Q 9			17	4,5,4,4
Q 10	C4	Agriculture	19	4,5,6,4
Q 11			8	4,2,2
Q 12			8	2,4,2
Q 13	C5	Entertainment	8	2,1,4,1
Q 14			8	4,4
Q 15			15	5,4,3,3

Each sentence in the document is scored based on the term frequency of the document. Special weights are assigned to terms occurring in the document with that of query terms. Here a bi-level ranking is done. At first level, ranking is based on relevance of anchor text with given query (since a web user search the results based on the anchor text). For the first level the scores are sorted and the preferred link order is chosen based on the scores. At the next level document contents are ranked based on the weights they gain from sentence scoring process. Finally the sentences were retrieved by the summarizer depending on the user requirements. Subsumption is set to filter duplicates based on threshold of 0.75.

IV. CONCLUSION AND FUTURE WORK

We have presented a system called OCS, which is capable of optimizing the contents available in electronic form. Since the optimized contents are not quiet satisfactory due to diversified information in each links, we have clustered the links. Finally we have clustered the results to provide the information in condensed form.

In this paper we have not focused on refinement of user query, to optimize the results which would even provide an optimized result. Moreover our work does not focus on measuring the quality of the summarized content which we leave it for future extensions.

ACKNOWLEDGEMENT

The authors would like to express their thanks to the Dr.Kanniyappan, Vice Chancellor, Mr. Abdul Qadir A. Rahman Buhari, Pro Vice Chancellor, Dr. V.M.Periasamy, the Registrar, Dr. T.R. Rangaswamy, Dean (Student Affairs) & HOD/Department of Information Technology, B.S.Abdur Rahman University, for the environment provided

REFERENCES

[1] Chen,Y., Hou,H.L and Qing Zhang,Y, "A personalized context-dependent Web search agent using Semantic Trees", Annual Meeting of NAFIPS, pp. 1-4, 2008.

[2] Leung, K.W.T., Wilfred,Ng and Lee,D.K, "Personalized Concept-Based Clustering of Search Engine Queries", IEEE Transactions on Knowledge and Data Engineering, vol. 20, Issue: 11, IEEE Press, Los Angels, USA, pp. 1505-1518,2008.

[3] Ferragina, P and Gulli, A, "The anatomy of a hierarchical clustering engine for Web-page, news and book snippets", Proceedings of Fourth IEEE International Conference on Data Mining, pp.395-398,204.

[4] Meng,X.J., Cai Chen,Q.C., Wang,X.L and Yang,X.H, "Improving web search ranking by Incorporating Summarization Systems", Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp.3075-3080,2007.

[5] Kim,K.J and Bae Cho,S, "A personalized Web search engine using fuzzy concept network with link structure", Proceedings of IFSA World Congress and 20th NAFIPS International Conference, pp. 81-86, 2001.

[6] Zhan,L. and Zhijing,L, "Web mining based on multi-agents", Proceedings of ICCIMA, pp. 90-95, 2003.

[7] Yamamoto,T., Nakamura,S and Tanaka,K, "Rerank-by-Example: Efficient Browsing of Web Search Results", In : Wagner,R., Revell,N., Pernul,G. (eds.) DEXA 2007. LNCS 4653, pp. 801–810, 2007.

[8] Lewandowski, D, "The Retrieval Effectiveness of Web Search Engines Considering Results Description", Journal of Documentation, vol. 64, Issue 6, pp. 915-937,2008.

[9] Schockaert,S., Cock,M.D., Cornelis,C and Kerre,E.E,"Clustering Web Search Results Using Fuzzy Ants", International Journal of Intelligent Systems, Vol. 22, pp. 455–474, 2007.

[10] Veronis,J , "A Comparative study of six search engines", http://sites.univ-provence.fr/veronis/ pdf/2006-comparative-study.pdf

[11] Hariharan,S and Srinivasan,R ,"A Comparison of Similarity Measures for Documents", Journal of Information and Knowledge Management, World Scientific Publishing vol.7, No.1, pp. 1-8,2008.

[12] The Google Page Rank Algorithm, http://www.rankforsales.com/google-page-rank.html.

[13] Keyhanipoor,AH., Piroozmand,M., Moshiri,B and Lucas,C, "A Multi-Layer/Multi-Agent Architecture for Meta-Search Engines", Proceedings of AIML'05, pp.19-21,2005.

[14] Lawrence, S and Lee Giles,C, "Context and Page Analysis for Improved Web Search", IEEE Internet Computing, pp. 38-46, 1998.

[15] Hariharan,S and Srinivasan,R, " Investigations Single Document Summarization by Extraction Method", In: Proceedings of ICCCN'08. IEEE Computer Society ,2008.

[16] Porter,M.F. :An algorithm for suffix stripping", Program, 14(3), pp. 130−137, 1980.

[17] http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words.

[18] Yulan, H., Hui,S.C and Sim,Y, "A Novel Ant-Based Clustering Approach for Document Clustering", In: Ng,H.G. et al. (eds.) AIRS 2006. LNCS 4182 ,pp.537-544 ,2006.

[19] Handl, J., Knowles,J and Dorigo,M, "Ant-based Clustering and Topographic Mapping", Artificial Life,MIT Press. Cambridge,USA, vol. 12, Issue. 1,pp.35-61,2006.