PHRASE CO-OCCURRENCES AS INTERFACES FOR SEARCH ENGINES RETRIEVAL

P.PICHAPPAN
Department of Information Science
Annamalai University
Annamalainagar 608002. TN
India
ppich@vsnl.net pichappan@dirf.org

LATHA PILLAI
National Assessment and
Accreditation Council
Dr Rajkumar Road
Bangalore 560012.
India
lp407@yahoo.com

P.VIJAYAKUMAR
Sri Venkateswara
College of Engineering
Sriperumbudur 602 105.
India
vijai@svce.ac.in

ABSTRACT

The degree of complexity of web content is increasing and hence the prototype experiments to parse the large web collections are initiated recently. The multi-tier architecture is found to be quite effective to ease the complex information space. Through this paper an interface for search engine is proposed to minimize the distorting results that occur while hits are ranked. The degree of low relevance between adjacent hits is not addressed in web retrieval experiments. This problem is highlighted and co- phrase occurrence frequency is employed to re-rank the retrieved hits. The results are encouraging and lead to further directions of research in refining search engine results.

PRELUDE:

The complexity of information space in web is addressed extensively in literature and the concern to solve the formidable heterogeneity is paid more attention. Extracting semantically and cognitively related knowledge depends on the understanding the semantic connections underlying in the web content. To support concept access, web content processing with due emphasis on semantic relevance is required which would ultimately enable the semantic link across disparate files and data structures. As a consequence, semantic integration issues have now become a key bottleneck in the deployment of a wide variety of information management applications.

BACKGROUND:

The crucial issue in web retrieval is the estimation of relevance between retrieved hits and query and also within the retrieval. The second component is neglected in many optimization studies. The search engines' mechanism of ranking the pages for query is carried out mainly based on certain assumptions. First, the term occurrence is believed to be the determinant factor in page retrieval and also in ranking them. Many papers and studies have identified and critically examined the ways in which the pages are ranked. '*Google*' achieved notoriety by ordering in accordance with the number of pointers it found to the target webpage. This is not unlike the idea of listing articles in order of the number of other papers citing the one in question. For the most part, search criteria tended to be simple or conditional text matching Simple heuristics might rank pages by the number of times they contain the query term, or they may favor instances in which that text appears earlier. But such approaches can sometimes fail spectacularly[1].

Search engines such as *AltaVista*, *Infoseek*, *HotBot*, *Lycos* and *Excite* use heuristics to determine the way in which to order the pages and fix priority. The rules of the engines are collectively known as a ranking function, which are prevalently applied.

Building semantic linkages, is thus has impact on effective retrieval. The intrinsic less efficient semantic properties of web can be enhanced if algorithms are developed by incorporating conceptual analysis skills. Deploying strategies for weaving semantic web in a highly distributed and decentralized environment is a difficult task. Understanding the semantic relationships and the use of special-purpose heuristics are highly challenging as the base for doing the objects is highly complex, raw and unstructured. It is true that the content correlation in the web pages occurs beyond key terms.

Web has distributed services to transfer the raw objects across repositories by incorporating processing activities from different locations. The need is to develop the distributed services to translate concepts from index terms across domains and to create perfect linkages in the content. The parsing mechanisms could extract generic units from the objects, and the indexing statistically correlates these uniformly across sources. For doing this processing refinement, co-occurrence frequency offers more promise.

RELATED WORKS:

Many information retrieval models including the Vector Space Model, probabilistic models, and fuzzy logic models depend on the frequency of query terms in a given document and they also introduce ways to normalize the weighting of terms to account for document length. [2,3,4]

If there is a link from page v to w, then the link is prescribed by author of v, or by crawlers or by algorithms. It is not the quantity of links, but the quality of links pointing the sites. Since all links are not created equal, the engines attempt to rank the importance of each link, and to understand the context of the link. Search engine ranking is widely discussed in[5]. Since the search engine links to pages are not perfect, efforts are undertaken by researchers to improve the efficiency of search process. One such algorithm for refinement is proposed by *Jeffrey Dean* and *Monika Henzinger* which was labeled as ‘The Companion algorithm’ derived from the HITS algorithm proposed by Kleinberg for ranking search engine queries.[6] The problem of resource discovery is addressed in *Soumen Chakrabarti* who found the giant all-purpose crawler not adequate for resource discovery.[7]

In many retrieval experiments, the selection of seed web pages is given priority as the top ranking pages in the retrieved list from search engines are found to be quite important even they do not ensure relevance. Hence, the deployed algorithms advocate for the selection of these few web pages as seed pages from which links could be established. The linkages in the pages and the page content are inseparable components in web page processing; the extended linkage concept has given significance in many recent studies and architecture[8].

Resource discovery systems, prior to the web environment exist around the world. Many of them are virtually subjective as the relevance is measured not in cognitive and semantic terms. The resource discovery of online knowledge takes a different course as machine processing systems lead which provides a different kind of atmosphere. The processing systems thus could be shaped and oriented to ensure objective measurement.

PROBLEM STATEMENT:

In the proliferation of online sources, scholars, scientists, government agencies, and individuals need to familiar with diverse especially online information resources available for the effective utilization of knowledge. Information seekers need to identify those information resources that are related to the query, and should able to navigate those resources. The highly distributed online knowledge across the emerging networked landscape is a crucial issue. The challenge is to decide how to mix, match, and combine one or more search queries with one or more knowledge repositories.

In the search environment, users log on to the retrieved pages in accordance with the ranking algorithm and find a vast content dissimilarity between adjacent hits. In the distributed environment, it happens that a large number of retrieval results for a query term, each query correlates to the retrieval with varying degree of relevance. Index parsers try to establish the correlation between the query and retrieval and fail to recognize the fact that the retrieval results are largely disconnected them selves. Listing the retrieval in the order of relevance remains the major challenge to search mechanisms. Cluster search engines offer some sort of solutions by grouping the retrievals for the queries. Even, it is an improvement over the normal search mechanisms, the retrieval among the different groups also suffers from the same conceptual problems. Search mechanisms connect the query and the retrievals but fail to connect the retrievals.

The designing of algorithms by exploiting the cognitively existing fundamental interconnections among online knowledge would aid the retrieval process.

MULTI-TIER ARCHITECTURE:

Multi-tier architecture offers more promise as the complexity of web is growing in different directions and no single and simple level architecture would capture all features. This problem is well observed by *Risvki* et al.[9] In the multi-tier architecture, automatic search engine result extraction is found to be very effective and productive.[10] In the multi-tier architecture, one can proceed to offer two extraction processes - one responds to the link analysis and the other is associated with the analysis of the content.[11] Several methods recommend less human effort in searching inlcuding a recent algorithm proposed in the keyword spicing[12]. Thus a multi-tier architecture with more automated processing and less human effort may offer better equation. In the passages below, an interface to the search engines is given.

ARCHITECTURE TO ESTABLISH CO-RELEVANCE

The work for the current study is structured in the following way. The retrieved results in search engines for a given query are subjected to further processing for clustering based on content similarity. The pages for any query are analysed for phrase indexing. The phrase indexing specified in an earlier paper[13] by us is found useful and applied for phrase identification. Each retrieved file for the query is indexed for phrase extraction by frequency count and stop words elimination that results in meaningful phrases. The key words that occur more than the threshold and the words that co-occur with the key words with a 'n2' proximity operator both pre and post qualifiers produce better precision in content description. The application of this algorithm produces set of phrases for retrieval which then grouped for relevance in terms of co-occurrences. One could expect a high relevance between adjacent hits in terms of content. However, this is not the case in many instances. The phrases that occur with a high frequency for each hit are now matched with each other for co-frequency between the retrieved files. High co-frequent terms for files are now placed for content similarity. This exercise will certainly produce a total re-ranking of hits and even the isolation of some files retrieved. Some of the isolate files would be totally irrelevant of query.

The figure 1 presents the architecture for refined retrieval. For the given query, the crawler retrieves large number of pages. The pages are now subjected by the indexer or processor given in the figure. The files that range from 'p1' to 'pn' are extracted for phrases particularly for 'above threshold'. The clustering of the phrases for similarity is based on the premise that high co-occurred phrases form clusters. The result is the formation of clusters for a query where the clusters are relevant to the query that ensures high semantic retrieval.

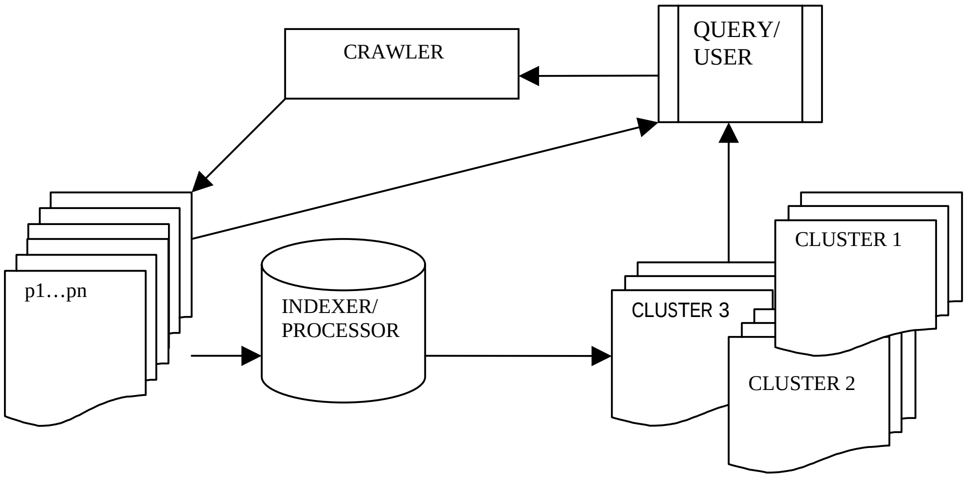


FIGURE 1 Architecture that shows the clustered retrieval based on content similarity

EXPERIMENT:

The sample queries are placed for search engines that given rise to a large number of hits; top retrieved ones are captured and placed in the experiment files. There were 12455 pages, for a given query extracted from meta search engines and combined results that ultimately eliminated the overlapping results. Of these, 127 pages were given among the top 500 ranks in the combined results. The total phrases extracted from the analyzed files are 13452 and the phrases are unified for all files and top occurred ones are compared for co-frequency. The total number of unique phrases generated was 799.

The total 127 full text files with 799 phrases are tested for co-frequency. Interestingly, the co-retrieved files particularly the adjacent ones have little co-frequency leaving to get placed at different clusters. The clustering activity is based on co-occurrences of phrases. The next stage after the co phrase extraction was to create concept map to serve as the experimental test bed for users. The initial results are encouraging.

Given below is the sample page with the estimated relevance between the analyzed files. The table 1 presents the statistical relevance score between the seed page and the pages that occur in the search mechanisms retrieved adjacent top pages.

TABLE 1

Mapping of Web pages based on Content Similarity

Sample Study “Allergic Rhinitis”

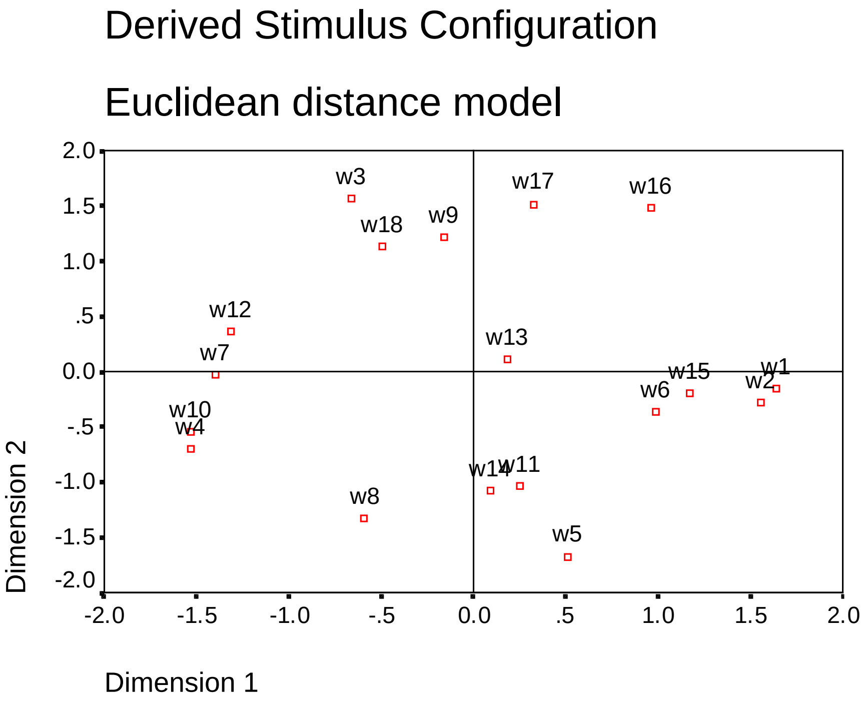
Ranked List of pages as per correlation

Source page: w6

Related pages to w6 according to the correlation values:

W15	0.95
w16	0.85
w11	0.85
w10	0.85
w8	0.85
w4	0.84
w5	0.84
w17	0.80
w18	0.79
w7	0.78
w9	0.77
w13	0.74
w1	0.69
w3	0.68
w14	0.64
w12	0.42
w2	0.37

The resulting pages are now mapped to explicitly show the scattering of retrieved pages in terms of semantic content. The top pages retrieved are widely distributed when they are parsed for concept correlation. The figure 2 shows the varying degree of placement of pages related by search engines but not related in semantic correlation.



w1…………wn   denotes web pages

FIGURE 2   Mapping of Web pages based on Content Similarity Sample Study "Allergic Rhinitis"

SUMMARY:

Semantic contents processing attempts to aid the information retrieval to ensure high relevance. The semantics proposed advocate the deep parsing for small collections, while scalable semantics towards shallow parsing for large collections. For text documents, the analyzed units are phrases while the statistical indexes record the co-occurrence frequency, of how often each phrase occurs with each other phrase within a document within the collection. The system ensures high rate of success in establishing a semantic linkages among the retrieved collections for queries.

The initial querying and its output can further strengthen by measuring the content similarity between co-retrieved pages particularly in the adjacent hits, which enable to refine the retrieval in many ways. First the degree of relevance between the retrieved pages is measured; the irrelevant ones could be separated from the relevant ones. This process ensures automatic parsing of terms based on the phrase analysis. The re-ranking can be scaled so that a better interface for search engine build up. Further experiments are being carried out to validate results and in the subsequent research, re-ranking algorithms will be presented.

REFERENCES:

[1] Hypersearching the Web**,** *Members of the Clever Project* , 1999.

[2] D. Harman, Ranking Algorithms, *In: Information Retrieval Data Structures and Algorithms,* W.B Frakes and R. Baeza-Yates, Eds., Upper Saddle River, N.J., Prentice Hall, 1992.

[3] G. Singhal, A. Salton, and C. Buckley, Length Normalization in Degraded Text, *Fifth Symp. Document Analysis and Information Retrieval*, 1996.Available online at http://www.research.att.com/~singhal/ocr-norm.ps

[4] Yanhong Li, Toward a qualitative search engine, *IEEE Internet Computing,* July - August 1998 , pp. 24-30.

[5] http://searchenginewatch.com/searchday/

[6] http://elib.cs.berkeley.edu

[7] S. Chakrabarti, M. van den Berg, and B. Dom, *8th World Wide Web Conference*, Toronto, 1999.

[8] Jaroslav Pokorn, Web Searching and Information Retrieval, *Computing in Science & Engineering*, July/Augsut 2004, pp. 43-48.

[9] Risvki et al. Multi-tier architecture for search engines, *Proceedings of the 15th Symposium on Computer Architecture and High Performance Computing, SBAC-PAD03*, 2004 .

[10] Zonghuan Wu, Vijay Raghavan Hua Qian, Vuyyuru Rama K Weiyi Meng, and Hai He Clement Yu, Towards Automatic Incorporation of Search engines into a Large-Scale Metasearch Engine, *Proceedings of the IEEE/WIC International Conference on Web Intelligence (Wl'03) 2003*, Computer Society.

[11] Ah Chung Tsoi, Structure of the Internet, *Proceedings of the 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing* ,Hong Kong, 2001.

[12] Satoshi Oyama, Takashi Kokubo, and Toru Ishida, *IEEE Transactions on Knolwedge and Data Engineering*, 16(1) , 2004, pp. 17- 27.

[13] P. Vijayakumar, et al. The implications of deploying proximity operators in web content processing, *Journal of Digital Information Management*, 2(4) ,2004, pp. 183-185.

APPENDIX 1

List of URLs studied (a few top studied URLs are alone given)

http://www.nationaljewish.org/MFhtml/ALR_MF.html
http://www.nationaljewish.org/medfacts/allergic_rhinitis.html
http://www.mckinley.uiuc.edu/health-info/dis-cond/allergy/allergrh.html
http://www.aaaai.org/public/fastfacts/rhinitis.stm
http://www.uoregon.edu/~uoshc/allergicrhinitis.html
http://www.geocities.com/nutriflip/Diseases/HayFever.html
http://www.med-help.com/Allergicrhinitis.html
http://www.allergyusa.com/allergicrhinitis.htm
http://www.jcaai.org/Param/Rhinitis/Complete/non_allergic_rhinitis
http://www.montana.edu/wwwebm/AllergicRhinitis.htm
http://www.allergy.mcg.edu/advice/rhin.html
http://www.icarus.med.utoronto.ca/carr/manual/allergic.html
http://www.outlinemed.com/demo/allergy/6706.htm
http://www.hon.ch/Library/Theme/Allergy/Glossary/rhinitis.html
http://www.aaaai.org/public/publicedmat/tips/rhinitis.stm -
http://www.atlallergy.com/rhinitis.html
http://www.respiratorybenchmarks.com
http://www.umm.drkoop.com/conditions/ency/article/000813.htm