Negligence of Phrase Indexing in Search Engines

Pit Pichappan
Department of Computer and Information Sciences
Al Imam University
Riyadh
Saudi Arabia
ppichappan@gmail.com

P. Vijayakumar
Venkateswara College of Engineering
Sriperumbudur
Chennai
India
pvijai@gmail.com

ABSTRACT: *Currently, search engines suffer from conceptual problems despite the application of refined techniques. The fundamental problem lies in the absence of classical information processing concepts in search engine processing. The limitation of search engines in phrase processing is highlighted in this paper. The results of search engine retrievals are compared with propriety database search results. The inferences if applied in search engine processing would lead to high precision in search engine retrievals.*

Keywords: Search Engines, Term processing, Phrase Index, Propriety databases

Received: 11 November 2008. Revised 8 January 2009. Accepted 11 January 2009

1. Introduction

Today's web environment is impelling individuals and organizations to shift into the so-called online-dependent information world. This emerging growth requires suitable information systems, electronic content processing support, effective search system, etc.

The information processing techniques followed in web are complex as each search engine is unique in deploying methods and technologies. However the processing systems in domains follow lightweight algorithms to match new documents with those stored in a database, rapidly focuses on the information that is required(1) In a rather different way, general techniques address the diversified requirements and present solutions to offset the complexity, forcing the existing systems refined.

2. Query Syntax

Currently despite the complex algorithms, there are many limitations in the web content processing systems. The poor precision of retrievals is the result of both the search engines limitations and ambiguity in concept expression by users in terms of too general words. Despite the availability of search assistance such as search keys, query refinements and other advance search options, typical users employ free text words and pose queries without using qualifiers for search terms. To offset the limitations and to improve the search effectiveness, both query expansion during either before search or after preliminary search are employed and efforts are made to improve the content processing.

Despite the structured ways of processing, the popular search engines such as ‘google’ suffer from poor precision in retrievals. They are away from the application of concept-based indexing. This problem is compounded by the fact that most of the web users are novice with the lack of search knowledge. Many users tend to make too general queries where they use mostly one word as query. About 70% of searches in *Infoseek* contain just one word (2). A way to improve the search results is to employ more search terms or phrase. However, users cannot be directed to use phrases and more terms in searching. A flexible and robust search mechanism should allow for free text search where users can use any word and term and the retrieval could result in better precision. Such a system considers the ambiguity in the queries and brings forth the refined results.

All search engines follow a typical processing which dictates the search process. Most of the searching systems use the keywords given in queries and match with descriptors identified in the indexed texts and retrieve the pages and documents. The matching between text descriptors and index terms leads to retrieval. The basic problem here is that all retrieved results are not relevant and the relevant ones are not totally relevant to the query. Hence, the retrievals are ranked based on the degree of relevance. Ranking algorithms employed in retrieval are many; some are explicit and others are intriguing.

3. Conceptual units in processing

The search engines unfortunately provide weight to terms rather than concepts. In indexing, the fragmented term feeding plays a central role and the design should address this issue in a way to offset the limitations from the users. While many experiments tested hyperlink structure to determine relevance, they provide little emphasis for the term frequency in retrieval. The above propositions lead to understand the inputs essential for designing retrieval systems. In many earlier design studies(3,4,5,6), term frequency was given prominence and designers (see for example *Kilgarriff* [3]) supported for accepting the term which occurs within a document with more frequency than other terms of a document.

Term Proximity or word occurrence in documents as phrase captures the texts that are related to the query terms. Phrase constitution with associative or proximity words which occur with high frequency than the rest of the phrases have semantic relatedness with the query term. The retrieval design could consider this factor and incorporate as an element.

The ‘disband’ of terms in the retrieval is a primary cause for retrieval inefficiency. Words are the carriers of concepts but the proper sensing of the concepts by words is the determining factor in retrieval design. Identifying the word association leads to improve the concept detection in texts and this is enabled by the identification of phrases from the text against the query terms.

4. Related studies

The information retrieval research has addressed the term frequency widely and attributed a positive value to this approach. The use of phrase as query than the words has positive impact in retrieval. The limitation of short query without using phrase is highlighted in the TREC experiments. These experiments reported that retrieval effectiveness dropped when short queries were used. (7, 8) As indicated in the earlier passages, the users, particularly the novice, continue to pose short queries. It is the design that could address the issues by interfacing the query with the phrase in the text so that the context preserved in retrieval.

The relationship between the documents and the query terms in retrieval is extensively addressed in (9, 10) which offer empirical analyses to support the refining the query terms by offering proper interfaces. Phrases have long been used to supplement word-based indexing in IR systems [11, 12, 13], yet these methods have not been widely applied to document clustering. (14) An enhanced technique for text processing is the use of the co-appearance of pairs of words as the attributes of the documents’ vector representations [15]. The words and phrase frequency is not the only option for web page retrieval; equally significant is the ranking of such retrieval. Ranking of retrieval needs to be properly focused to ensure the rich retrieval and has been given emphasis in many experiments (16, 17, 18, 19, and 20).

Thus, the propositions posted here give strength to the term and phrase frequency and the word frequency dependent ranking in retrieval. In the discussions below we have presented empirical results emerged from the testing of web search retrieval and presented the keys to the search engine retrieval systems.

5. Systems

In the light of the foregoing discussions, we have initiated exercises to address the limitations by drawing illustrations from search engine retrieval and shown how the keywords and their frequency play central role in web information retrieval.

To understand how the web searches retrieve pages, the normal procedure is to feed searching terms and test the retrieved results for relevancy, volume of hits and other parameters. In many search environments, users typically employ keyword searches which consist of free words in disassociate form. The faulty retrieval continues to occurs because of the disassociate feeding of query terms. Hence, we employed typical associative searches by feeding phrases in selected (popular in terms of users' preference) search engines such as *Google, Lycos, All the web* and *Alta vista*. A search, even in a perfectly constructed query would retrieve large files with varying relevance. The indexing algorithm is believed to be rich in semantics and syntactic when the retrievals are ranked by decreasing degree of relevance. However, in search engine retrieval, the voluminous retrievals are distributed unevenly without any logical relevance structure. When users restrict their access to the top ones the semantically rich text remains untapped. Thus, the retrievals can be weighed for relevance and it is beneficial to estimate how the high relevant retrievals occupy the top ranks. The ranking techniques employed by the search engines have considerable influence on their retrieval effectiveness. The retrieval thus is estimated for the relevance as judged by experts. The top hundreds in each of the search engines are posted to the focus group experts and their evaluation is measure in a five point scale ranging from very low or no relevance to high relevance. The mean score of them is estimated for each retrieval and the total 100 retrievals are grouped into ten groups, each group constitutes ten pages. The relevance score in an ideal retrieval is likely to decrease from the first cluster to the succeeding ones. The exercise measures how the mean relevance values are distributed over the clusters and the extent of relevance dispersion.

Further, the retrieval from the propriety databases are examined for the degree of relevance in accordance with the above described method. Since the propriety databases are constructed manually based on the pre-determined key words, they could ensure more efficient retrievals. The next step moves ahead to find the phrase concentration index. The phrase concentration index is defined as the ratio between the phrase occurrence in texts and the total words in the text. This measure seems to be somewhat raw as it measures the relation between terms and phrases. However, if an exercise measures the total phrases in texts, multiple phrases cause overlapping and counting problems. The theory behind the phrase concentration index measure is to know the correlation between the search phrases and text keys.

Another measure employed is the phrase frequency index of texts. The occurring phrases in all retrieved texts are subjected to all phrases identification in the texts and the phrases are ranked for their frequency. The search phrase matching is carried out for all retrieved texts and the rank of the search phrase is fixed which the phrase frequency index is. If both the phrase frequency rank and the phrase concentration index decreases with cluster, the ranking technique employed in searches would be highly acceptable and would match with users' relevance.

6. Analysis

While many studies have presented the web retrievals and examined the relevance by evaluators' perception, our search results have presented a different type of analysis. The results from selected search engines for phrase search are posted for expert validation which is a departure from the previous studies. The simple reason behind the presentation of the results of phrase search is that even users pose structured queries, search engines return voluminous answers and still most of them are irrelevant.

In order to test the validity of the prepositions described above, we have extracted from web against queries and conducted relevance measurement. The web pages collected from the searches were parsed to obtain the word frequencies contained in the resources. The pages range from snap shots to text files, home pages and large text sets.

Four different selected phrases are included for searching in four search engines, G*oogle, Alta vista, Lycos* and *All the Web*. These phrase queries return large voluminous retrieval as given in the table 1.

The number of hits from search engines depends on the size of the pages indexed by them. The large sets of retrievals pose formidable problem to the users in one way and however the root of this disorientation problem is not in *number per se*, but in one aspect of an otherwise unknown ranking algorithm. The searches return weakly related documents and snippets that do not enjoy any relevance to the query. The search engines hide the fact that the relevance is unevenly distributed among retrieved files.

Search terms	**Google**	**Alta vista**	**Lycos**	**All the web**
1. Grid Computing	9,893,000	6,340,000	873,700	887000
2. Computer Network	12,306,400	7,040,000	2,498,005	8950000
3. Research Collaboration	1,45 0,000	945,000	149,300	947000
4. Information Growth	47,000	57,500	17,000	53000

Table 1. Results of Phrase Search in Search Engines

While it is demonstrated about the size of retrieved files from search engines retrieval, the structured propriety databases constructed in disciplines have different way of indexing where most of the coverage comes from manual effort. These databases are considered to be highly structured as the system of predefined format enable to ensure more relevance than the web searching. However, we would project the fact that still the way of processing is also prone to irrelevance and low precision.

The phrase searching of the four selections are carried out in three databases including the beta scholar search engine, 'google scholar'. As google scholar is produced form the database sources and believed to be a refined search engine than the general search engines, it is also used to analyse. The Science Direct is a major full text database which is indexed from the full text of 1900 plus scientific journals. The indexing system of it, even not explicitly given, is highly structured and the key words are based on the terms given in the full text. It permits to search for the key terms available not only in titles or journal key words but from the full text also. The third is the web of science produced by the *Institute for Scientific Information* which is based on citations and logically connects the scientific papers.

The search carried out for the search engines are also used to retrieve files or full text from the propriety databases. The results given below enable to sense the size as well as to understand the degree of structured results.

Search terms	**Scholar**	**Science Direct**	**Web of Science**
Grid Computing	11000	122	**763**
Computer Network	**191,000**	**1220**	**368**
Research Collaboration	**7,980**	**117**	**162**
Information Growth	357	76	9

Table 2. Results of Phrase Search in Propriety Databases

The number of hits is determined by the volume of coverage rather than the relevance size. The inclusion of files in the above semi/full structured sources is highly skewed. This is mainly because for the databases other than google scholar, the target sources are limited as they index only highly selected ones.

While the searches brought a large volume of hits, the present study scanned the first selections, i.e., 100 of all web and propriety databases searches. As the search processing systems do not say about the degree of relevance between query term and retrievals, and all retrievals are not equally related to queries, it becomes essential to understand the relevance. As indicated in the section 4, the relevance score is calculated for both search engine and database retrievals. The relevance measurement carried out in the current exercise is manual based on expert validation with a focus group. 7 Experts were invited to measure the relevance of retrieval to the queries posted. The queries and search results are transferred to them thro servers and analysed their judgments for relevance.

Hence, it could not be possible to measure for all hits, and the top 100 of them are selected and presented to experts group who validated the relevance by using a five point scale that range from 1 to 5 (no or very little relevance to perfectly relevant) which is converted into the relevance score like correlation measurement, 0 to 1. To apply this parameter, the top 100 full text files were given to them. These 100 are grouped into ten clusters based on their ranks where the first ten forms the cluster one. The ten clusters of retrievals with their mean relevance score is given in the table 3.

Cluster	**Mean Relevance Score**
1	0.42
2	0.32
3	0.16
4	0.54
5	0.18
6	0.34
7	0.06
8	0.28
9	0.12
10	0.14

Table 3. Mean Relevance of web searchers validated by experts

Correlations			
		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	-.560
	Sig. (2-tailed)	.	.092
	N	10	10
VAR00002	Pearson Correlation	-.560	1
	Sig. (2-tailed)	.092	.
	N	10	10

Correlation between clusters and relevance score

The correlation values between the clusters and relevance score is calculated. A logical ranking should offer high level relevance for the top and lower values for the further ranks.

The correlation is found to be insignificant at the 0.05 level (2-tailed). In an ideal search process, the retrieval relevance decreases with rank. However, the web searches result in skew ness where the relevance is not only low but values are deviated and the following figure (figure 1) also supports the interpretation.

It is documented that the users rely more on structured databases than the web processing for information access. We have proceeded to estimate the relevance between user queries and retrievals in databases search. We took the first 100 for the google scholar and 20 each for the Scopus and Web of Science databases with an exception for the retrieved files from web of science for the search phrase ‘information growth’ as the hit is very less for which we took only the top five and the relevance scores are given below in the table 4.

The probability and reality if matches with each other, the validity of ranking could be documented. The observation of the above table reveals the inconsistent relevance in the clusters with respect to all the propriety databases, however the WOS and science direct have less scattering and it is evident from the figure below. The mean relevance among ten clusters for the propriety databases are 0.44, 0.59, and 0.62 for scholar, Science direct and WOS respectively.

In the next stage, we have identified how the search terms and phrases are available in the text of the retrieved files and how significant they are to the text. This is measured as detailed in section ….., the phrase index by two principal mechanisms, viz., the rank of the phrase frequency in the text and phrase concentration index.

Both the web searches and database searches are subjected to these two measures and the results are available in the following tables 3 and 4.

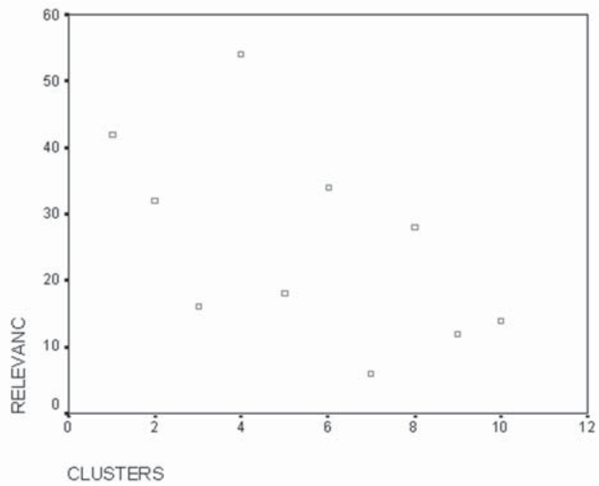

Figure 1. Dispersion of relevance score of web search with clusters

Relevance Score			
Clusters	Google Scholar	WOS	Science Direct
1	0.35	0.87	0.93
2	0.50	0.67	0.83
3	0.41	0.53	0.39
4	0.26	0.64	0.64
5	0.51	0.73	0.60
6	0.30	0.51	0.58
7	0.66	0.62	0.49
8	0.38	0.42	0.69
9	0.52	0.39	0.55
10	0.54	0.48	0.49

Table 4. Mean Relevance of Propriety Databases searchers validated by experts

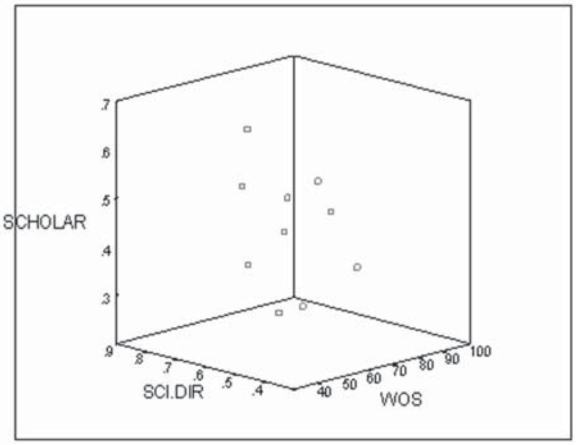

Figure 2. Dispersion extent of relevance consistency for three searches

In web searches using phrases, when the phrase used for search occurs with more frequently in the web files, the relevance is high. It is visible from the two measures, Phrase Frequency Rank and the Phrase Concentration Index. In the case, the cluster 4, the concentration index is high and the occurrence of the search phrases is also very high in the retrieved files. The concentration values decrease with the decrease of the frequency rank thus support the argument that the term frequency and particularly the phrase frequency ensure a high semantic retrieval.

Clusters	Phrase Frequency Rank	Phrase Concentration Index
1	13	0.035
2	15	0.033
3	6	0.027
4	2	0.058
5	18	0.031
6	21	0.029
7	38	0.019
8	24	0.016
9	53	0.004
10	28	0.012

Table 5. Phrase Frequency Rank and Phrase Concentration Index of Web search results

Correlations			
		VAR00001	**VAR00002**
VAR00001	Pearson Correlation	1	.683
	Sig. (2-tailed)	.	.029
	N	10	10
VAR00002	Pearson Correlation	.683	1
	Sig. (2-tailed)	.029	.
	N	10	10

Correlation is significant at the 0.05 level (2-tailed).

The correlation between Phrase Frequency Index and Phrase Concentration Index are proved to be significant and recalls the earlier mismatch between the clusters and experts focus group decision on relevance.

The high occurrence of search phrases in texts and their top ranks in frequency are very important and they should form the core key words in the key words database.

7. Summary

Improvements in indexing of web pages could be possible if search engines can review their strategies of indexing particularly the term and phrase frequencies. We have documented that from the testbed phrase relation between the query and text is significant in retrieval and the experiments conducted reinforce indexing process.

The major limitation of the search engines is that users feed multiple key words that are conjunctive from user's perception, where they are treated by search engines as disjunctive. The retrieval depends on other factors such as relevant documents in the collection, the ability of parsing web pages by search engines, relevance assessment, etc.

The processing is not limited to word or phrase frequency as they consider syntactic aspect in indexing. The semantic richness of retrieval is preserved only when the indexing concentrates on concepts.

References

[1] Samuel W.K. Chan. Beyond keyword and cue-phrase matching: A sentence-based abstraction technique for information extraction, Decision Support Systems, 2005

[2] Steve Lawrence, Lee Giles. Searching the web, General and scientific information access, IEEE Communication Magazine 1999. 116-122

[3] Kilgarriff A. Which words are particularly characteristic of a text? a survey of statistical approaches. In: Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition, Brighton, UK, Apr. 1996. 33–40.

[4] Kleinberg J. M . Authoritative sources in a hyperlinked environment. Journal of the ACM 1999. 46; (5): 604–632.

[5] Robertson S. E, Spark Jones K. Simple proven approaches to text retrieval. Technical report 356, Cambridge University Computer Laboratory, May 1997.

[6] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 1988. 24; (5):513–523.

[7] Voorhees E. M, Harman D. Overview of the Fifth Text Retreival Conference (TREC-5). In: E. M. Voorhees & D. K. Harman (Eds.), Proceedings of the Fifth Text Retrieval Conference (TREC-5) Gaithersburg, MD: National Institute of Standards and Technology.1996. 1–28

[8] Harman D. Overview of the fourth Text Retrieval Conference (TREC-4). In Harman D. K.(Ed.), Proceedings of the fourth Text REtrieval Conference (TREC-4Gaithersburg, MD: National Institute of Standards and Technology. 1995.1–23.

[9] Hearst M. TileBars: visualization of term distribution information in full text information access, in: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'95), 1995. 59-66.

[10] Veerasamy A, Belkin N. J, Evaluation of a tool for visualization of information retrieval results, In: Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), 1996. 85-92.

[11] Fagan J. L. Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and nonsyntactic methods, Ph.D. Thesis, Cornell University, 1987.

[12] Hull D. A, Grefenstette G, Schulze B. M, Gaussier E, Schütze H, Pedersen L. O. Xerox TREC-5 site report: routing, filtering, NLP, and Spanish tracks, in: Harman D. K (Ed.), The Fifth Text Retrieval Conference (TREC-5), NIST Special Publication, 1997.

[13] Salton G, Yang C. S, Yu C. T. A theory of term importance in automatic text analysis, Journal of the American Society for Information Science, 1975. 26;(1): 33-44.

[14] Oren Zamir, Oren Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results

[15] Maarek Y. S, Wecker A. J. The Librarian's Assistant: automatically organizing on-line books into dynamic bookshelves, in: Proceedings of the International Conference on Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94), 1994.

[16] Can F, Altingovde I. S, Demir E. Efficiency and effectiveness of query processing in cluster-based retrieval. Information Systems, 2004. 29;(8): 697–717.

[17] Clarke C. L. A., Cormack G. V, Tudhope E. A. Relevance ranking for one to three term queries. Information Processing and Management 2000. 36;(2): 291–311.

[18] Wilkinson R, Zobel J, Sacks-Davis R. Similarity measures for short queries. In Fourth text retrieval conference TREC-4 Gaithersburg, Maryland. 1995. 277-285.

[19] Long X, Suel T. Optimized query execution in large search engines. In Proceedings of the 29th international conference on very large databases. Berlin, Germany. 2003.

[20] Barla Cambazoglu B, Cevdet Aykanat. Performance of query processing implementations in ranking-based text retrieval systems using inverted indices, Information Processing and Management 2005.