Are Concepts Useful for Organizing Search Results?

P.Y. Ko, R.W.P. Luk, E.K.S. Ho, F.L. Chung
Department of Computing
The Hong Kong Polytechnic University
(852) 2766 5433
{csrluk, csksho, cskchung}@comp.polyu.edu.hk

D.L. Lee
Department of Computer Science
The Hong Kong University of Science and Technology
(852) 2358 3534
dlee@cs.ust.hk

ABSTRACT
This paper reports an explorative study on organizing search results by concepts that are selected titles of Wiki pages. Because of limited display areas and for ease of navigation, two novel algorithms identify at most three general concepts, and at most five of their specific concepts for display. Our evaluation shows that the retrieval effectiveness improvement is significant at 90% confidence level using the paired student's *t*-test, albeit our users have no access to document titles nor to the content.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *graphical user interface, evaluation.*

General Terms
Design, Experimentation and Human Factors.

Keywords
Information retrieval, Search interface and Wikipedia.

1. INTRODUCTION
This paper reports an explorative study on the feasibility of using concepts for organizing search results. It is not a full scale user study (e.g., [1]). Many search result interfaces organize the retrieved documents into a ranked list (e.g., Google). Its use is supported by the probability ranking principle [2,3]. It is not easy to find an alternative that enhances search experience.

Requests may be ambiguous [4], especially when they are short. Web queries are such examples, and their average length is 2+ words [5]. Ambiguous requests may satisfy some users, but not for all. This problem may be resolved by using individual's profile to influence ranking. However, sometimes the profile may not be related to the ambiguous request. Another approach suggests alternative queries that contain the original query terms (e.g., Yahoo!/Google term suggestions). However, this assumes that the unambiguous, intended request contains at least one original query term. Alternatively, Clusty/Vivisimo uses a similar term suggestion approach that organizes terms into a concept hierarchy (e.g., [6]). In our experience, some novice users may be lost when navigating through a complex hierarchy. Although much research (e.g., [6]) uses concepts, whether they improve retrieval effectiveness is an open issue.

2. OUR APPROACH
Our assumption is that concepts related to the query are useful for organizing search results, disambiguating requests and finding related terms, albeit not for all queries. By default, the search results are organized into ranked lists (or by clicking "ALL" in Fig. 1). If this list is good, the displayed concepts (e.g., "crude oil" in Fig. 1) will likely be ignored by users. This study does not focus on the interface layout or time-space efficiency, but on whether concepts help retrieval effectiveness.

2.1 Concepts
Our initial design uses mined word sequences as concepts. Some users find that some mined concepts are not meaningful, and their relationship with the query is not clear to them. The choice of meaningful terms seems important. As a remedy, we use titles in Wiki pages (called Wiki titles here) as our "display" concepts. They are matched in the retrieved documents by the longest (forward) match. These Wiki titles also need to be related to the query. First, they may be some contiguous sequence of query terms. Second, their Wiki pages may be linked to or from Wiki pages that have titles which may be a contiguous sequence of query terms.

Some Wiki pages have redirect links. For example, "U.S." is redirected to the "United States" in Wiki. We use the titles of these redirecting Wiki pages as related terms of the titles of the destination Wiki page. This helps the coverage of the Wiki titles, and such titles are meaningful as the following pilot study shows. We ask 10 human subjects to judge whether the titles of the redirecting and the destination Wiki pages (e.g., "U.S." and "United States") are related for 10 random samples, each of which has 30 pairs of redirection-related Wiki titles. On average, 85% of these title pairs are thought to be meaningful.

2.2 Interface Layout
Details about user acceptance of the interface layout are not presented, since this is not our focus. However, the layout constrains the number of display concepts and how these concepts are related to each other.

Fig. 1 shows our interface layout which has three rows of concepts. For each row, the leftmost concept is the most general. Its related specific concepts are on its right. To avoid taking too much space and to avoid forming complex hierarchies, at most three rows are displayed (i.e., only three general concepts), and the number of specific concepts is limited to at most five for each row. This constraint also helps to filter misidentified concepts that are not related to the query, assuming that the highly ranked concepts are highly related to the query. The document frequency of the concept within the top *N* retrieved documents is shown on its immediate right. This implicitly helps users to distinguish the general concepts from specific ones. Not all concepts need to be related to the topic, because the request may be ambiguous, or because they are used for discriminating desired documents from undesired ones

2.3 Automatic Concept Selection for Display

For selecting display concepts, a directed acyclic graph (DAG) is formed by considering each matched Wiki title as a node. Two nodes are linked if their point-wise mutual information (PMI) [7] will be positive. The link points at the Wiki title that has the higher document frequency.

Two new selection algorithms are experimented. The in-degree algorithm selects three highest in-degree leaf nodes. These are the general concepts. Their children nodes are the specific concepts, at most five of which are selected by the strength of PMI. The other algorithm is a greedy one. In this case, the DAG is simplified by deleting all edges apart from the largest weighted edge of each node. Starting from each leaf node, nodes will be combined together if the additional nodes add more unique retrieved documents to the existing set. When the algorithm terminates, the three leaf nodes that have the largest number of retrieved documents (i.e., coverage) are selected. Their specific concepts are selected by their PMI values.

3. EVALUATION

Five title queries of the TREC-7 ad hoc retrieval test collection are used because their performances have noticeable variation for experimentation. Only five queries are used because this is an explorative study. The initial retrieval ranks documents by BM25 term weights [8]. Five users including male and female, who are receiving tertiary education in computing, are asked to select concepts. They do not have access to document titles nor document content, so that the content has no influence on users' behavior. Initially, they read the title queries of the TREC topics before manual concept selection. Next, the task is repeated with the same topics but their narrative fields (i.e., long queries) are read. For measuring retrieval effectiveness, documents that contain any selected concept are ranked before those that do not. Documents in the same subset are ranked by BM25. The retrieval effectiveness measure is the commonly used R-precision. It is the precision of the top R retrieved documents, where R is the number of relevant documents.

Table 1. Average R-precisions of five TREC-7 queries. Title queries are used for the initial retrieval. Key: *G* for greedy algorithm and *I* for In-degree selection algorithm.

	Users Read:	Title Queries		Long Queries	
Query	Original	*G*	*I*	*G*	*I*
351	.396	.292	**.458**	**.583**	**.479**
352	.244	.199	**.411**	**.317**	**.415**
365	.457	**.771**	**.743**	**.771**	**.743**
387	.059	**.224**	**.271**	**.235**	**.271**
400	.568	.560	.328	.560	.472
Average	.345	**.409**	**.442**	**.493**	**.476**
p-value	N.A.	.449	.349	.053	.117

Table 1 shows the R-precision of each query and the average R-precisions across queries. The "original" column shows the performance of the initial retrieval using BM25. The other columns are R-precisions averaged across users. There are at least 5 percentage points increase compared with the initial retrieval. Although the Wilcoxon signed ranked, paired nonparametric test is significant at 85% confidence level, the highest confidence level will be at 90% using the paired student's *t*-test (Table 1) if the R-precisions averaged across users are asymptotically normally distributed across topics. If the most effective query 400 is removed, the confidence level of the greedy algorithm for long queries will reach 95% (i.e., *p*-value = 0.032; not shown in Table 1). This situation may occur in practice, because users will ignore the displayed concepts if the default initial ranking is already good. Notice that after the users read the long queries, they are able to pick more effective concepts. This will be expected in practice if knowledgeable users pose short queries and select concepts themselves.

4. CONCLUSION

Carefully selected concepts derived from Wiki seem promising for organizing search results because the poor performing queries may be improved substantially. The confidence level may reach 90% using the paired student's *t*-test.

ACKNOWLEDGEMENT

This work is supported by PolyU Project # G-U289.

REFERENCES

[1] Käki, M., and Aula, A. Controlling the complexity in comparing search user interfaces via user studies. *Information Processing & Management*, *44*, 1 (2008), 82-91.

[2] Robertson, S.E. The probability ranking principle in IR. *Journal of Documentation*, *33*, 4 (1977), 294-304.

[3] Fuhr, N. A probability ranking principle for interactive information retrieval. *Information Retrieval*, *11*, 3 (2008), 251-265.

[4] Spärck-Jones, K., Robertson, S.E., and Sanderson, M. Ambiguous requests: implications for retrieval tests, systems and theories. *ACM SIGIR Forum 41*, 2 (2007), 8-17.

[5] Spink A., Jansen, B.J., Wolfram, D., and Saracevic, T. From e-sex to e-commerce: Web search changes. *IEEE Computer*, *35*, 3 (2002), 107–109.

[6] Lawrie, D., and Croft, W.B. Generating hierarchical summaries for web searches. In *Proc. ACM SIGIR '02*, 2002, 457-458.

[7] Church, K.W., and Hanks, P. Word association norms, mutual information and lexicography. In *Proc. ACL Computational Linguistics Meeting*, 1989, 76-83.

[8] Robertson, S.E., and Walker, S. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. ACM SIGIR '94*, 1994, 345-35.

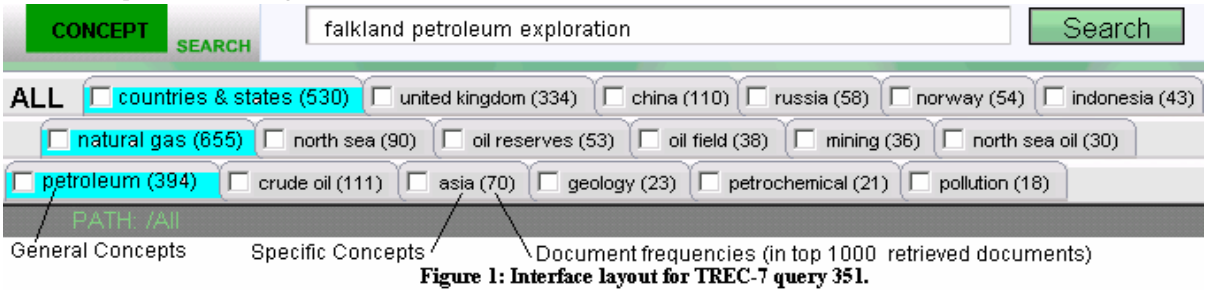

Figure 1: Interface layout for TREC-7 query 351.