Do Longer Queries Retrieve More Diverse Results?

Michael Bendersky
bemike@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

ABSTRACT

In this paper, we demonstrate that verbose and grammatically complex queries retrieve, on average, more diverse results across different search engines than the short keyword queries. Our evaluation using both commercial and open source search engines shows that the overlap between search engine results decreases by up to 50% as a function of query length.

1. INTRODUCTION

Web search engine users today are facing the paradox of choice [4]. Web search engines come in many flavors, including keyword search engines, meta-search engines, vertical search engines and natural language search engines. Users are typically loyal to a single search engine, and ignore the rest [6]. However, research shows that, in fact, different search engines retrieve very different results [5]. Hence, it is important for both IR researchers and search engine users to understand which searches can be improved by using multiple search engines [6].

In this paper, we explore the relation between the overlap among the results retrieved by the different search engines and the search query verbosity. Recently, researchers found that verbose natural language queries pose a challenge to modern search engines. They both exhibit lower clickthrough rates than the keyword queries [3], and retrieve less relevant results [1].

High search engine overlap indicates that querying a single search engine would be sufficient for the query. Low overlap among search engines characterizes challenging information needs that can be potentially better answered by fusing the results from different sources. For instance, Beitzel et al. [2] showed that such result fusion can be especially successful when search engines retrieve diverse *relevant* documents at high positions. In this paper, we demonstrate that verbose and grammatically complex queries retrieve, on average, more diverse results across different search engines than the short keyword queries.

l(q)	Example
1	*yahoo*
2	*gary valenciano*
3	*treasury saving bonds*
	...
5	*galloway township community charter school*
6	*what does your IQ score mean*
	...
10	*a former Soviet state that lies along the 30th meridian*

Table 1: Queries of different length.

2. DATA AND METRICS

For the purpose of this study we used 5,000 queries, which were randomly sampled from a search log of a major commercial search engine[1]. In our work, we focus on studying the relation between result overlap in web search and query length. Accordingly, we apply the following stratified query sampling approach.

First, we define query length, $l(q)$, as the number of word tokens (sequences of characters separated by space) present in query q. We then randomly sample 500 unique queries for each query length in the range $l(q) = (1, \ldots, 10)$.

Some examples of the sampled queries are presented in Table 1. Note that as the query length increases, the complexity of the grammatical structure increases as well. In the range $l(q) \leq 3$, queries address a single entity. Slightly longer queries usually describe a simple relation between two entities. For $l(q) > 5$, search queries include wh-questions and semi-complete sentences that describe complex information needs.

For each of the 5,000 queries, we retrieve the *top-k* results using a search API of two commercial search engines: Bing and Google. In addition, we use the Indri toolkit[2] to create two simple search engine variations (described in Section 3.2) and retrieve results from the ClueWeb09 corpus[3]. We set $k = 10$, to limit the retrieved list to the first result page shown to the user.

We use $\mathbf{r}_{x,q}$ to denote the list of *top-k* retrieved results by a search engine x for query q. As search engines may retrieve several pages from a single domain, we normalize the results on a domain level. For any given query q, we can express the overlap between the results of search engines x and y as

$$\mathcal{O}_{x,y}(q) = \frac{|\mathbf{r}_{x,q} \cap \mathbf{r}_{y,q}|}{|\mathbf{r}_{x,q} \cup \mathbf{r}_{y,q}|}.$$

[1]Available as a part of Microsoft 2006 RFP dataset.
[2]`http://www.lemurproject.org/indri/`
[3]`http://boston.lti.cs.cmu.edu/Data/clueweb09/`

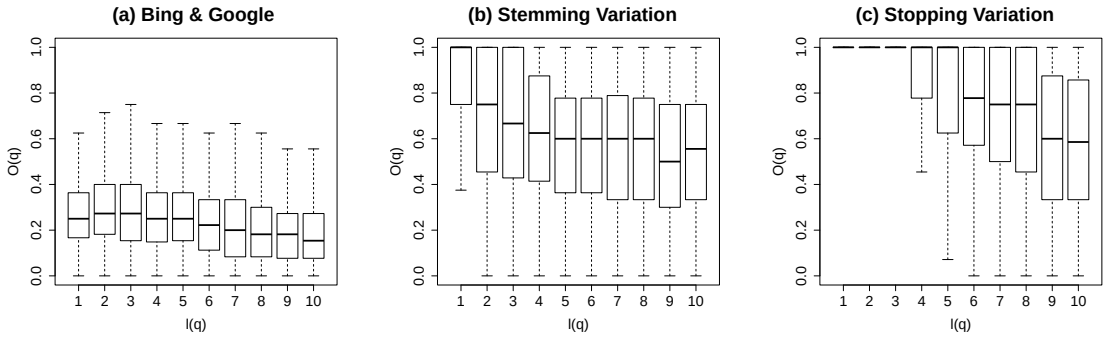

Figure 1: Overlap in search results as a function of query length.

Note that $0 \leq \mathcal{O}_{x,y}(q) \leq 1$, with $\mathcal{O}_{x,y}(q) = 0$ when x and y share no common results, and $\mathcal{O}_{x,y}(q) = 1$ when x and y retrieve the same results (not necessarily at the same ranks).

3. EVALUATION

3.1 Overlap in Web Search

In general, the overlap (when considering all 5,000 queries) between the search engines is quite low. The mean overlap is 0.25, i.e., Bing and Google search engines, share, on average, 2.5 results per query on the first page. This low overlap is in line with previous work [5], and indicates that a web meta-search engine that could successfully fuse the results from multiple search engines would be beneficial to the users.

Figure 1(a) demonstrates the distribution of $\mathcal{O}(q)$ by query length for Bing and Google. It is interesting to note that as the query length increases, the overlap between the results of the search engines generally decreases, indicating less consensus among search engines for more verbose queries. Mean overlap drops from 0.27 for single-term queries to 0.18 for queries with ten terms, a 50% decrease.

The decrease in overlap is gradual. For $l(q) \leq 3$, an average overlap is 0.29; for $4 \leq l(q) \leq 6$, it is 0.25. Finally, for $l(q) > 6$, the average overlap drops to 0.20. Thus, we conclude from Figure 1(a) that the complexity of the grammatical structure of the queries increases the diversity of possible query interpretations by the search engine.

3.2 Search Engine Variations

A comprehensive investigation of all the possible factors that may impact the divergence in search engine results is out of the scope of this paper. However, using the Indri toolkit, we conduct a simple experiment that demonstrates the effect of two such factors – stemming and stopword removal – on queries of different length.

3.2.1 Stemming Variation

In the stemming variation, we create two alternate search engines which differ solely in their stemming strategy:

A1-No stemming *A2-Porter stemmer.*

Figure 1(b) demonstrates the overlap as a function of query length for these alternatives. *A2* has a very limited effect on the single-term queries, many of which are not influenced by stemming (e.g., queries like *yahoo* or *hotmail*). However, as the query length increases, there is a clear divergence between *A1* and *A2*, especially for queries with more than four terms. Overall, Figure 1(b) shows a decrease of 50% in overlap between *A1* and *A2* as a function of query length.

3.2.2 Stopping Variation

In the stopping variation, we create two alternatives which differ solely in their stopword removal strategy.

A1-No stopword removal *A2-INQUERY stopword list.*

Figure 1(c) demonstrates the overlap as a function of query length for these two alternatives. *A2* strongly influences the queries with six or more terms, which is consistent with the fact that many of these queries are wh-questions or semi-complete sentences that contain many function words. Overall, Figure 1(c) shows a decrease of 42% in overlap between *A1* and *A2* as a function of query length.

4. CONCLUSIONS

In this paper, we demonstrate, using a pair of commercial web search engines as well as two variations of an open source search engine, that verbose natural language queries retrieve significantly more diverse results across search engines than the short keyword ones. This result indicates that verbose queries are both a challenge and an opportunity for information retrieval: there is less consensus on what is the best way to answer these queries, and hence they carry more potential for improving retrieval through fusion of different retrieval methods.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by ARRA NSF IIS-9014442. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring reductions for long web queries. In *Proc. of SIGIR*, pages 571–578, 2010.

[2] S. M. Beitzel, O. Frieder, E. C. Jensen, D. Grossman, A. Chowdhury, and N. Goharian. Disproving the fusion hypothesis: an analysis of data fusion via effective information retrieval strategies. In *Proc. of SAC*, pages 823–827, 2003.

[3] M. Bendersky and W. B. Croft. Analysis of long queries in a large scale search log. In *Proc. of WSCD*, pages 8–14, 2009.

[4] B. Schwartz. *The paradox of choice: why more is less.* HarperCollins, 2004.

[5] A. Spink, B. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major web search engines. *Information Processing and Management*, 42(5):1379–1391, 2006.

[6] R. W. White, M. Richardson, M. Bilenko, and A. P. Heath. Enhancing web search by promoting multiple search engine use. In *Proc. of SIGIR*, pages 43–50, 2008.