Evolution of Web Search Results within Years

Ismail Sengor Altingovde[1]
L3S Research Center
Hannover, Germany
altingovde@l3s.de

Rifat Ozcan
Computer Eng., Bilkent University
Ankara, Turkey
rozcan@cs.bilkent.edu.tr

Özgür Ulusoy
Computer Eng., Bilkent University
Ankara, Turkey
oulusoy@cs.bilkent.edu.tr

ABSTRACT
We provide a first large-scale analysis of the evolution of query results obtained from a real search engine at two distant points in time, namely, in 2007 and 2010, for a set of 630,000 real queries.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

General Terms
Experimentation.

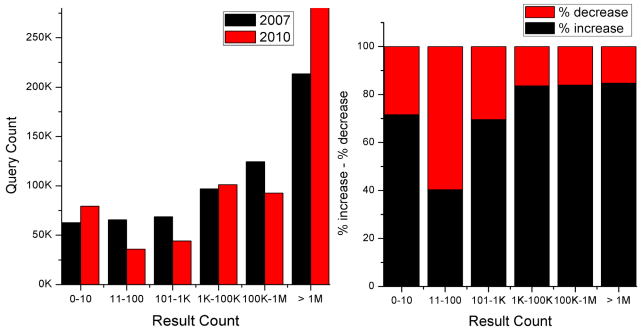
1. INTRODUCTION
The dynamicity of Web takes an increasing attention from the researchers as many studies investigating the changes in the Web content (e.g., [2]) and user queries (e.g., [7]) emerged in the last years. While these works provide quite valuable insight on the dynamics of Web search, another important dimension is usually overlooked: How do the real life search engines react to this dynamicity? That is, how the changes in the underlying collection and in the search engine's internal algorithms affect the query results presented to the end user? In this paper, we provide a first large-scale analysis of the evolution of query results obtained from a real search engine at two distant points in time, namely, in 2007 and 2010, for the same set of 630,000 real life queries.

To the best of our knowledge, only a few earlier works investigate the evolution of real search results. In one of the earliest studies, 25 queries are submitted daily to a number of search engines for a month to observe the stability of top-10 results [6]. A similar experiment with a larger set of 12,000 queries is reported in [3]. Bar-Ilan and Peritz monitor the change of a single topic, namely, "informetrics", for a period of 8 years [1]. Finally, McCown and Nelson investigate whether the application programming interfaces (APIs) provided by major search engines yield results synchronized with those retrieved from the Web interface [4]. In their experiments, 100 keyword queries are repeatedly submitted to different search APIs for five months. Nevertheless, our work differs from these in the following aspects: (i) Our queries are not synthetic, but taken from a real life query log, (ii) While the previous works involve a limited number of queries, we use a large set of 630,000 queries, (iii) We analyze real results retrieved by a commercial search engine in two points in time that are more than 3 years apart, and (iv) We focus on the properties of the results, rather than the evolution of the underlying collection.

Our analysis in this paper attempts to find answers to several high-level questions regarding the evolution of Web search results, such as: How is the growth in Web reflected to top-ranked query results? Do the query results totally change within time? Does higher number of Web documents lead to results that are located deeper in the Web site hierarchies? Do the result titles and extracted snippets exhibit any variation in time? We believe that this work, being the largest-scale longitudinal analysis of query results, would shed some light on these questions. We present our findings in the next section and summarize key results in Sec. 3.

2. EVOLUTION OF QUERY RESULTS
We used 630,000 unique queries that are randomly sampled from AOL Query Log [5]. For these queries, we obtained top-100 results from Web using Yahoo!'s public search API, twice: in June, 2007 and in December, 2010. Experiments spanned the entire month in each case, due to large number of queries. We identified a few Web sites that only listed all AOL query strings and removed them from the results as they are not real answers.

Number of query results. First, we simply compare the average number of results per query, as reported by the search engine. We are aware that these numbers are not completely reliable [4], but they still give a rough idea of how the result space has changed for our queries. We find that the average number of results per query is increased from 16.5M to 52.3M. This seems to be an expected result, as the Web has probably grown an order of magnitude from 2007 to 2010. Fig. 1(a) shows the distribution of queries based on the reported result counts in 2007 and 2010. Fig. 1(b) shows the percentage of queries whose result counts are increased or decreased in 2010 with respect to their result counts in 2007. For instance, 70% of the queries, which return at most 10 results in 2007, return more results in 2010. However, the increase percentage is lower for queries that have returned a small result set (i.e., less than 1000 results) in 2007. In contrast, queries with large number of results in 2007 can match to even larger number of results in 2010.



(a) Query count vs. result count (b) Result count change

Figure 1. Changes in the number of results in 2007 and 2010

[1] This work was done while the author was at Bilkent University.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

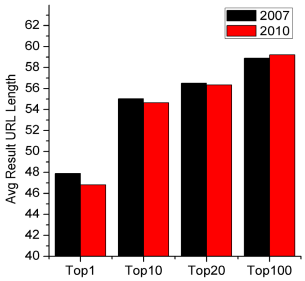


Figure 2. Result URL length

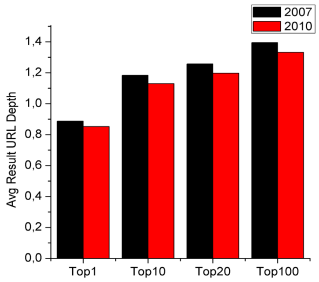


Figure 3. Result URL depth

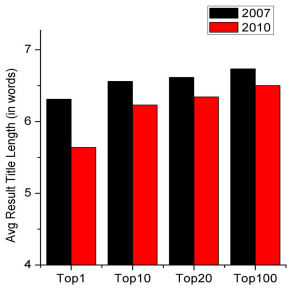


Figure 4. Result title length

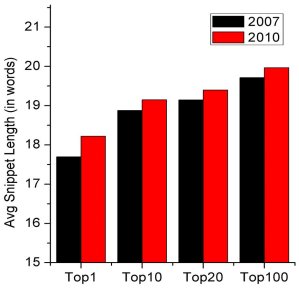


Figure 5. Snippet length

Table 1. Number of unique URLs in the query results.

	2007	2010	Overlap (% w.r.t 2010)
Top-1	475,860	437,483	87,248 (19.9%)
Top-10	4,377,299	4,456,026	476,649 (10.7%)
Top-20	8,330,692	8,737,776	836,125 (9.6%)
Top-100	34,576,357	39,437,931	3,384,122 (8.6%)

Length and depth of result URLs. In Figs. 2 and 3, we report the average length (in bytes) and depth of unique result URLs, respectively (The domain name is assumed to have a depth of 0.). We investigate whether the increase in the number of documents causes a search engine to retrieve pages that are located in a deeper position at a Web site. In contrary to this expectation, both URL length and depth decrease in 2010, which means that search engines prefer to retrieve pages at the top level of a domain most of the time. We also observe that the length and depth of URLs increase for those results that are ranked lower.

Number of unique URLs and domain names. In Table 1, we provide total number of unique URLs observed in top-*k* results ($k \in \{1, 10, 20, 100\}$) in 2007 and 2010. While this number slightly decreases for top-1 case, we observe a trend of increase with increasing values *k*. For top-100 results, the number of unique URLs increases by 14% in 2010 results. We also investigate the overlap in the retrieved URLs in 2007 and 2010. We find that 20% of the URLs returned at the highest rank in 2010 were at the same position in 2007. This implies that the "valuable" document space (i.e., documents that can answer real queries) does not grow in proportion with the entire Web.

Table 2. Number of unique domain names in query results.

	2007	2010	Overlap (% w.r.t 2010)
Top-1	230,464	242,859	90,040 (37.1%)
Top-10	1,065,881	1,362,538	373,811 (27.4%)
Top-20	1,678,452	2,249,991	599,280 (26.6%)
Top-100	4,462,468	6,599,437	1,705,899 (25.8%)

In Table 2, we make a similar analysis for unique domain names. The number of domains in top-*k* results is significantly smaller than the number of unique URLs, which implies that result documents share a smaller number of domains. The increase in unique domain names in 2010 is more emphasized in comparison to the increase in the number of unique URLs. For top-100 results, unique domains grow by a factor of 50% in 2010. On the other hand, the fraction of domains that have also appeared in the corresponding top-*k* list in 2007 is high, e.g., around 37% for top-1 results. This indicates that the domains that hosted query results in 2007 are successful for answering queries in 2010, as well.

Length of result titles and snippets. In Figs. 4 and 5, we present the average length of result title and snippet in terms of unique words. We see that, the both values increase for larger number of retrieved results. A comparison between 2007 and 2010 reveals that the search engine prefers to provide slightly shorter result titles and longer snippets in 2010. We guess that shorter titles aim to help the searcher to grasp the returned result more quickly, whereas the longer snippets possibly aim to be more informative.

3. KEY FINDINGS AND FUTURE WORK

Our key findings are: (i) we observe that although Web has probably grown significantly from mid-2007 to the end of 2010, this growth is mostly reflected to queries that already have rich results sets, but not to the other poor queries. (ii) A potentially high-quality set of URLs and domains appear in the query results of both 2007 and 2010. (iii) Result URL length and depth, as well as the title and snippet lengths, tend to increase at lower result ranks. The first three features are slightly smaller, whereas the last one, snippet length, is larger in 2010 in comparison to 2007.

In the future, we plan to make a query-wise analysis of our data.

4. ACKNOWLEDGMENTS

This work is partially supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant no. 110E135 and FP7 EU Project LivingKnowledge (contract no. 231126).

5. REFERENCES

[1] Bar-Ilan, J., Peritz, B.C. The lifespan of "informetrics" on the Web: An eight year study (1998-2006). Scientometrics, 79(1), 7–25, 2009.

[2] Fetterly, D., Manasse, M., Najork, M., Wiener, J.L. A large-scale study of the evolution of web pages. In *Proc. of WWW,* 669–678, 2003.

[3] Kim, J., Carvalho, V.R. An analysis of time-instability in Web search results. In *Proc. of ECIR,* 466–478, 2011.

[4] McCown, F., Nelson, M.L. Search engines and their public interfaces: which apis are the most synchronized? In *Proc. of WWW,* 1197–1198, 2007.

[5] Pass, G., Chowdhury, A., Torgeson, C. A picture of search. In *Proc. of the 1st INFOSCALE*, 1, 2006.

[6] Selberg, E., Etzioni, O. On the Instability of Web Search Engines. In *Proc. of RIAO*, 223–236, 2000.

[7] Spink, A., Jansen, B.J. A Study of Web Search Trends. Webology 1(2), 2004.