An Analysis of Time-instability in Web Search Results

Jinyoung Kim[1] and Vitor R. Carvalho[2]

[1] Department of Computer Science
University of Massachusetts Amherst**
jykim@cs.umass.edu
[2] Microsoft Bing
vitor@cs.cmu.edu

Abstract. Due to the dynamic nature of web and the complex architectures of modern commercial search engines, top results in major search engines can change dramatically over time. Our experimental data shows that, for all three major search engines (Google, Bing and Yahoo!), approximately 90% of queries have their top 10 results altered within a period of ten days. Although this instability is expected in some situations such as in news-related queries, it is problematic in general because it can dramatically affect retrieval performance measurements and negatively affect users' perception of search quality (for instance, when users cannot re-find a previously found document).

In this work we present the first large scale study on the degree and nature of these changes. We introduce several types of query instability, and several metrics to quantify it. We then present a quantitative analysis using 12,600 queries collected from a commercial web search engine over several weeks. Our analysis shows that the results from all major search engines have similar levels of instability, and that many of these changes are temporary. We also identified classes of queries with clearly different instability profiles — for instance, navigational queries are considerably more stable than non-navigational, while longer queries are significantly less stable than shorter ones.

1 Introduction

It is natural for web search results to change. Web documents are frequently updated by users as well as publishers. Search engines try to capture as much of these changes as possible and experiment with new retrieval techniques continually. The problem is that, as noted by Selbert and Etzioni [4], the amount of changes in web search results far exceeds the variation of the web itself. That is, for the same query, the top search documents rankings fluctuate even without any real content change in top documents.

As an example, we show daily rank changes of top 3 documents for query 'com' in a major search engine. Figure 1 shows that the relevant document

** This work was done during the author's internship at Microsoft Bing.

(about the Microsoft COM object model) changed its position 4 times within 9 days of measurement, resulting in the fluctuation of 22 percentual points in NDCG$_5$. This is a typical example of instability – a change in top results that lasted only for a few days. In fact, this issue of instability is quite common and is not confined to a single search engine. Our analysis in Section 5.1 shows overall instability trends for all three major search engines, which suggest that nearly 90% of general web queries experience some change at top 10 results within 10 days.

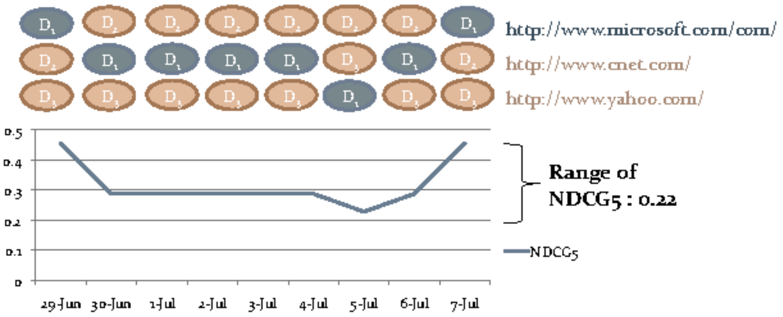

Fig. 1. Daily change of top 3 documents for query 'com' and corresponding fluctuation in NDCG$_5$.

This instability causes serious issues for both end users and search engine companies. For end users, this can be a source of confusion and frustration, especially when they want to get back to earlier results. Teevan et al. [5] suggests that 40% of web queries are re-finding queries, and that rank changes make it hard for users to get back to previously clicked results. For search engine companies, which monitor search quality all the time, this variation in search results mean unstable performance measurement, making them difficult to track down performance precisely and make important decisions (e.g., shipping of new ranking algorithm).

In this paper, we aim to analyze this phenomenon of time-instability in web search results. By instability, we mean the day-to-day change in top 10 results of the same query over a period of time. Our notion of change includes the change in the position of existing documents as well as the insertions (i.e., new documents added to the top 10 list) and deletions (i.e., a document was removed from the top 10 list) of documents. We focus on this short-term change in top rank list because of its high visibility to users.

To better understand the problem, we conceptualize different aspects of instability, based on its relation to structural change of search engine, its source and its duration. We also suggest a set of metrics for quantifying the degree of instability. Then we present an analysis based on large-scale data (12,600 queries), which shows that top results from web search engines experience a considerable amount of changes, and that many of these changes are temporary. We also found that many characteristics of a query – length, frequency and intent – affect its degree of instability.

The rest of this paper is organized as follows. In the next section, we provide an overview of related work. We then introduce several types of instability in Section 3, followed by a description of the data set and metrics we used for measuring instability in Section 4. Results from our data analysis are presented in Section 5. We then conclude this work in Section 6.

2 Related Work

Several previous studies focused on the changes in the contents of webpages, instead of chronological changes in retrieval ranks of the same documents. For instance, Fetterly et al. [2] crawled 150 million pages weekly for 3 months and found that 65% of pages remain the same during the period. Ntoulas et al. [3] performed similar studies and showed that major source of change is the creation of new pages as opposed to the update in existing pages. Adar et al. [1] crawled 55,000 webpages hourly for 5 weeks and found that 34% of pages did not have any change during the period. They attributed the differences in document sampling method to much higher frequency of change. In contrast, our work focus on the changes in the rankings of search engine results as opposed to the changes in document contents.

Teevan et al. [5] studied the re-finding behavior of web search users and showed that as much as 40% of all queries are re-finding queries. They also concluded that rank change has detrimental impact on this type of task, finding that users are less likely to re-click the results and take more time to do so when a previously clicked result changed its position. Building on their insights, our work presents the first quantification and comparison of search results time instability in major commercial search engines in large scale, as well as a detailed analysis on various aspects of this instability.

Selberg et al. [4] analyzed the instability of web search results more than ten years ago. They showed that 54% of top 10 documents are replaced over a month, and that web search results change much faster than the web itself. They also suggested the caching of results to improve the response time as a possible cause of this instability. Our study is different in that we used large-scale data which include ranking scores and relevance judgments. Also, while they focused on changes as the set overlap in results between two time periods, our analysis include many varieties of instability as well as the correlating factors.

3 Types of Instability

3.1 Structural vs. Non-structural

Modern commercial search engines are permanently experimenting with new ideas. Often search engines can introduce new features or ranking/indexing techniques that can cause sweeping changes to the rankings for a large number of queries. We refer to the instability derived from such expected and controlled events as structural instability. In contrast, non-structural instability is the one found even when there is no such events. In this work, we focus on non-structural instability.

3.2 Indexing Issues vs. Ranking Issues

Another distinction we can make about instability concerns the source it is stemming from, which can be grouped into indexing issues and ranking issues. Indexing issues are related to the availability of a document in the search engine index. That is, if a document is not found in the index, there is no chance it will be ranked among the top 10 documents, even if a few moments earlier the document was readily available. Various factors can be behind this type of instability, including, but not limited to, different (re)crawling policies, spam detection policies, click fraud policies, architecture and capacity limitations, to name a few. From a top 10 results page perspective, the common characteristic of all indexing issues is that they manifest as insertions or deletions of documents at top 10 list, leaving no impact on the relative ranking of the rest of documents.

Ranking issues include fluctuation in the value of ranking features, which can be caused by document content changes, link structure changes, anchor text changes, among many other factors. In contrast to indexing issues, ranking issues cause relative positions of existing documents to change, leading to rank swaps among existing documents. Note that, in our definition, swaps are defined in terms of two documents. The query in Figure 1 exemplifies this kind of change, showing 4 swaps during 9 days of measurement.

3.3 Short-term vs. Long-term

As previously mentioned, changes in search results are somewhat expected, and our notion of instability includes only such fluctuations which last only for a few days. In this sense, it is meaningful to classify the change (insertion, deletion or swap) with regard to its duration. For insertions, the duration measures how many days inserted documents stay at top 10 results. Similarly, the duration of a swap measures the period during which two documents keep their relative positions after a swap has happened. In this paper we arbitrarily define short-term changes as the ones that are revoked within 5 days, whereas any change that lasts longer than 5 days is considered as long-term.

4 Measuring Instability

4.1 Data Set

We collected daily snapshots of the top 10 results for the same set of 12,600 random queries over a period of four weeks between June and July of 2010. These queries were randomly selected from the Microsoft Bing [3] search engine query logs. For each day and for each top 10 query-document pairs, we also extracted the values of all its ranking scores. In addition, we also collected the human relevance judgments in a 5-grade scale (Perfect, Excellent, Good, Fair and Bad) for each query-document pair.

[3] http://www.bing.com

During the first two weeks of data collection, Bing deployed structural changes in the ranking/indexing techniques that significantly affected the profile of the collected top 10 query-documents. Yet no such changes happened during the last two weeks of data gathering. While our notion of instability focuses on the variation in results without any structural change, we compare the instability trend between two periods in Section 5.2.

News-related queries are expected to present large instability. Since we wanted to quantify the impact of these queries in our study, we identified from the collection any query that showed a *news* document (a document that was recently created) during the test period. Eventually, this process found 951 queries in which there were at least one insertion of *news* documents during the test period, which is about 7.7% of entire query set. In Section 5.4 and 5.5, we present how the characteristics of news queries are different from regular queries.

In order to compare instability trends across different commercial search engines, we also created a second data collection with the top 10 results for 1,000 queries issued at Google [4] and at Yahoo! [5] using their APIs during two weeks in August 2010. These queries were also randomly selected from the Bing query logs. Note that we used the API results for Bing here as well to ensure consistency in measurement. Table 1 summarizes our data set, where we named the collection from Bing as Bing Set, the collection from three major search engines as Big3 Set.

Table 1. The statistics of our two data collections.

Name	Date	#Queries	Source
Bing Set	6/12 – 7/8	12,600	Bing internal
Big3 Set	8/2 – 8/16	1,000	Bing / Google / Yahoo! API

4.2 Metrics of Instability

There can be several ways of measuring the instability in web search results during a time period between t and $t+n$. For instance, to measure how many of documents changed over time, we can use the ratio of set overlap between top k results of two rank lists.

$$\text{Overlap}_k(R_t, R_{t+n}) = \frac{|R_t \cap R_{t+n}|}{k} \tag{1}$$

If we are more concerned about the change in ranking, we can use the ratio of agreement in pairwise preference between two rank lists, which are defined as follows:

$$\text{PairAgree}_k(R_t, R_{t+n}) = \frac{|\text{PairPref}(R_t) \cap \text{PairPref}(R_{t+n})|}{k(k-1)/2} \tag{2}$$

Although above measures will summarize the change in the rank list itself, since we are concerned about the change in the quality of ranking in many cases,

[4] http://www.google.com

[5] http://www.yahoo.com

we can use the range of NDCG$_k$ (henceforth rNDCG$_k$) for each day during the period of our analysis. Alternatively, we can use the variance in NDCG$_k$ (vNDCG$_k$) instead of the range.

$$\begin{aligned} r\text{NDCG}_k(R_t, R_{t+n}) =& max(\text{NDCG}_k(R_t), ..., \text{NDCG}_k(R_{t+n})) \\ &- min(\text{NDCG}_k(R_t), ..., \text{NDCG}_k(R_{t+n})) \end{aligned} \tag{3}$$

Table 2. Pearson correlation coefficient between the measures of instability.

Measure	PairAgree$_5$	vNDCG$_5$	rNDCG$_5$
Overlap$_5$	0.986	-0.131	-0.347
PairAgree$_5$		-0.139	-0.362
vNDCG$_5$			0.798

Table 2 shows the correlation among these measures on the Bing Set. Since the Overlap$_5$ and PairAgree$_5$ are defined between two rank lists, we used the data from the first and the last day of measurement. In the end, we decided to use the range of NDCG$_5$ as main metric since it correlates relatively well with other metrics, and captures what is more likely to be noticed by users: the worst-case scenario of instability in ranking quality.

5 Instability Trends

5.1 Overall Trends

We first present overall level of instability for Big3 Set in Table 3. Although Yahoo! seems to have slightly more unstable results, all three search engines seem to have similar level of instability using the metrics we used. Over the two-week period of our measurement, around 20% of documents were replaced in top 10 results, and around 40% of pairwise preferences were changed. Average range of $NDCG_5$ was around 6%.

Table 3. Overall level of instability for three search engines.

Measure	Bing	Google	Yahoo!
Overlap$_5$	0.958	0.958	0.948
Overlap$_{10}$	0.819	0.815	0.798
PairAgree$_5$	0.911	0.911	0.895
PairAgree$_{10}$	0.641	0.642	0.603
rNDCG$_5$	0.057	0.058	0.063

As the daily trend of instability, Figure 2 shows the number of queries (out of the 1,000 sample queries) with change (insertion or swap) in the top 10 results for three major search engines. Here, it is evident that nearly 30% of queries had some change at top 10 every day and that nearly 90% queries experienced change within 10 days.

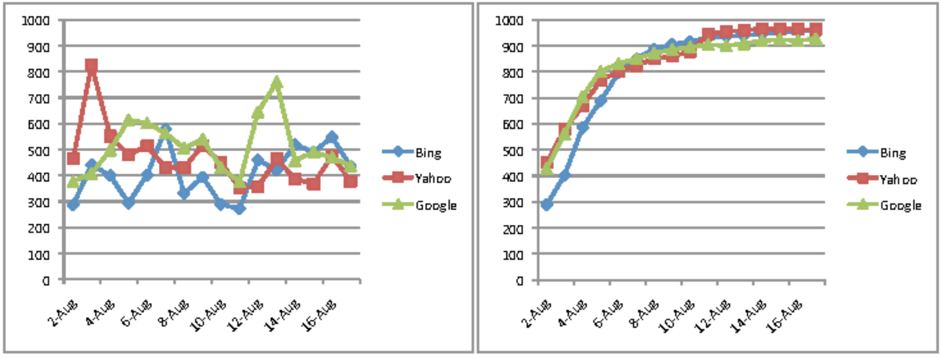

Fig. 2. Number of queries with some change in the ranking of top 10 for Big3 Set. The left figure shows a day-to-day trend, and the right figure shows cumulative trend for the same data.

Another observation from Figure 2 is that the daily amount of change among different search engines does not tend to correlate. (i.e. the day with big change in Google is not necessarily the same in Bing) This provides evidence that these instabilities are mostly due to internal factors specific to each search engine.

We further analyzed the correlation in query-level instability ($r\text{NDCG}_5$) between each pair of search engines. The result showed little correlation (less than 0.1 in Pearson correlation coefficient), whereas the correlation between retrieval performance measured in NDCG_5 was very high (around 0.7). In other words, an easy query in one search engine is likely to be easy in another search engine as well, yet a query with unstable result in one place is not necessarily unstable in another. This can be another evidence that instability in web search result is somewhat separate phenomenon from the dynamics of the web itself.

5.2 Structural vs. Non-structural

Here we compare the instability of the search engine when affected by a structural change versus non-structural changes using Bing Set. The graph in Figure 3 shows the number of queries with some change in the ranking of top 10 (*Chg@10*) and top 5 (Chg@5), improvement (*Imp@5*) and degradation (*Deg@5*) at NDCG_5 for former (left) and latter (right) 2-week period. It shows that all the queries had some change at top 10 at former period, and around 57% of all queries had changes in NDCG_5 (*Imp@5* + *Deg@5*). Comparing the change in ranking and the change in ranking quality top 5, we can see that around 30% of the change in ranking results did not cause any change in NDCG_5, as examplified by the query in Figure 1.

If you recall from Section 4.1 that Bing had a structural change in the former period, this level of change is not surprising. However, even in the latter period where we did not have any such event, we can still find that 90% of queries had some change at top 10, with 37% of queries had changes in NDCG_5. This clearly indicates that the top search results experience a great deal of instability even without any structural change. Also, here you can see that the number of

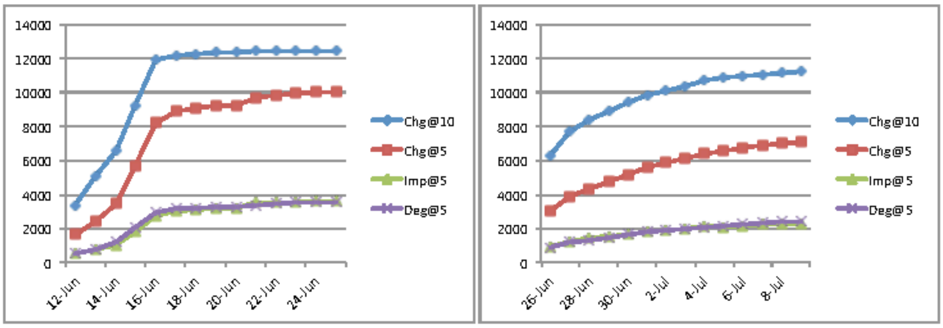

Fig. 3. The comparison between structural instability and non-structural instability. Each plot shows the cumulative number of queries with some change in the ranking of top 10 and top 5, improvement and degradation at NDCG$_5$ for Bing Set.

positive and negative changes are roughly equal, meaning that these changes does not impact the average retrieval effectiveness.

5.3 Insertions vs. Swaps

Next we present the amount of changes in the form of insertions and swaps. As mentioned in Section 3.2, an insertion means the addition of new document among the top 10 results, whereas a swap means the change in relative positions of two documents. Note that some of the insertions within top 10 results would be swaps if we consider rank positions beyond top 10.

The distribution of rank positions for insertions (left) and swaps in Figure 4 supports this point, showing a huge increase in insertions at rank position 9 and 10 overlapping with sudden drop at the number of swaps around the same rank positions. Otherwise, it shows that both the number of insertions and swaps steadily decrease as the rank position gets lower. We hypothesized that this relative stability at lower positions might be due to the fact that these documents have much higher scores than other documents, for which we provide some evidence in Section 5.5.

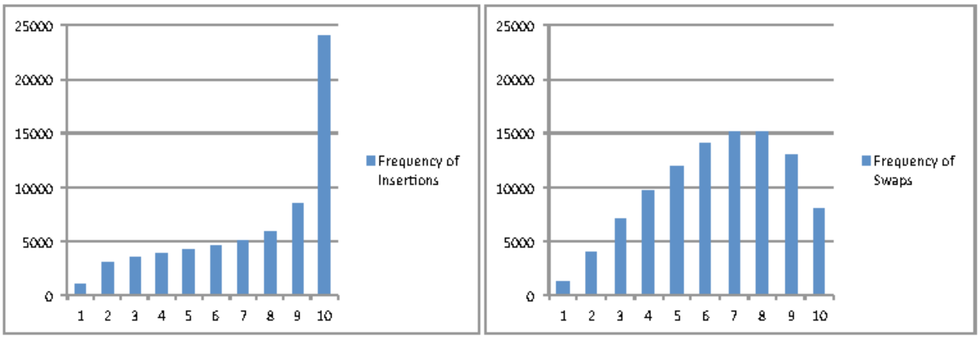

Fig. 4. The distribution of rank positions for insertions (left) and swaps (right).

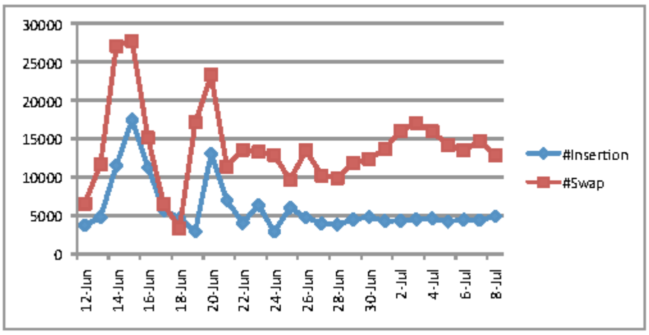

Fig. 5. Daily number of insertions and swaps at top 10.

Figure 5 shows the daily trends in insertions and swaps for Bing Set, where it indicates around 5,000 insertions and 12,000 swaps per day in latter two weeks (without structural change in search engine). This means that approximately one document was inserted for every other query, and that one pair of documents swapped positions for every query. We also found that the number of 'news' documents inserted was around 600 per day, from which you can see that the majority of top 10 insertions happens from non-news sources.

5.4 Duration of Change

We then analyzed the duration of change as defined in Section 3.3 using Bing data set. Figure 6 shows the number of insertions at June 26th, and the number of insertions that was revoked in the following days (that is, given all document insertions in the top 10 that happened on june 26th, the percentage of these documents that disappeared from the top 10 in the following days). Here you can see that queries with news intent (951) have larger portion of temporary insertions (90% if we use the threshold of 5 days). However, even for non-news queries (11,650), 50% of inserted documents are taken out of top 10 within 5 days.

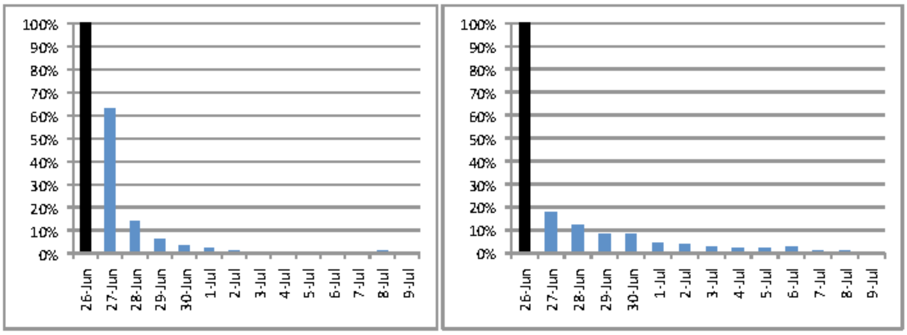

Fig. 6. The percentage of insertions on June 26th that was revoked in the following days for queries with news intent (left) and regular queries (right)

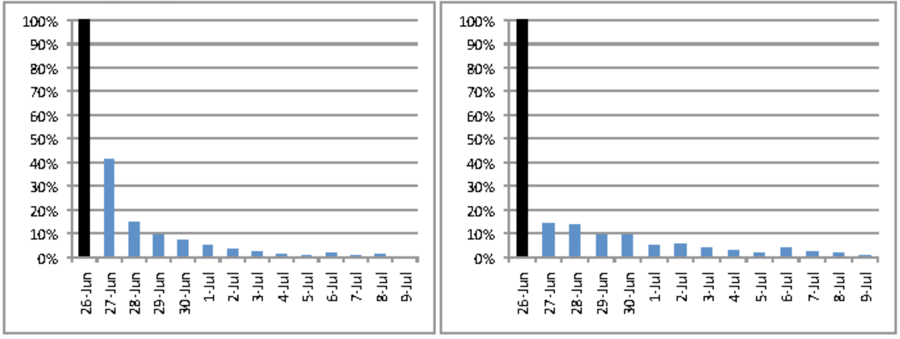

Fig. 7. The percentage of swaps on June 26th that was revoked in the following days for queries with news intent (left) and regular queries (right).

Figure 6 shows the number of swaps at June 26th, and the number of swaps that were revoked in the following days. Here you can see that queries with news intent have still larger portion of temporary swaps (70%), yet 50% of rank swaps for regular queries are revoked within 5 days. In overall, we can see that a majority of changes that happens to top results last no longer than a few days, as was the case of the example query in Figure 1.

5.5 Factors that Affect Instability

Finally, we investigate a few factors regarding queries that affect instability. We first look at how queries with different frequency and length have different levels of instability. Here, the frequency is calculated from Bing's query logs, and the length indicates word counts. Note that we use the range of NDCG$_5$ (rNDCG$_5$) as the metric of instability henceforth.

Figure 8 shows a boxplot with the instability of queries for different frequency ranges (in log) and different lengths. The width of each boxplot is proportionate to the square root of the number of data points in each range. Although instability and query frequency does not seem have straightforward correlation, it is clear that longer queries have higher instability than shorter queries. Given that most of long queries are known to be tail queries, the trends in both plots are consistent.

Next, the left side of Figure 9 shows the degree of instability for different levels of query difficulty, as measured by NDCG$_5$. Although there seems to be some indication that the hardest queries as well as the easiest queries have lower instability ranges, overall query difficulty does not show a clear correlation with time instability. The right side of Figure 9 shows the clear relationship between instability and the difference in ranking scores between the top document and 5th document (deltaScore5). This makes intuitive sense in that queries with smaller differences between document scores are more likely to have rank swaps even at the small change in input feature values. Also, combined with the finding that top positions in a search result have higher differences between documents (plot omitted for space constraint), this explains the result in Section 3.2 that top positions in web search results are usually more stable than others.

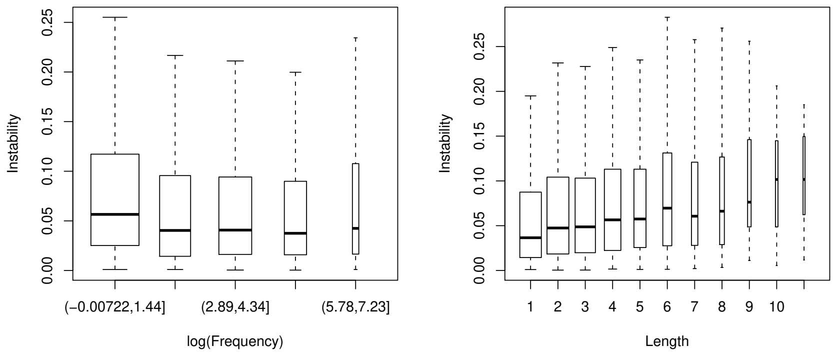

Fig. 8. The instability of queries with different log of frequency (left) and length (right).

We finally looked at the relationship between instability and some query classes. The left plot in Figure 10 shows that navigational queries have much less instability, which is consistent with the trend in Figure 8 since navigational queries are typically shorter and more frequent. The plot on the right confirms that queries with news intent has more changes in the top rank list, which is expected considering that search engines inject lots of news documents into those queries.

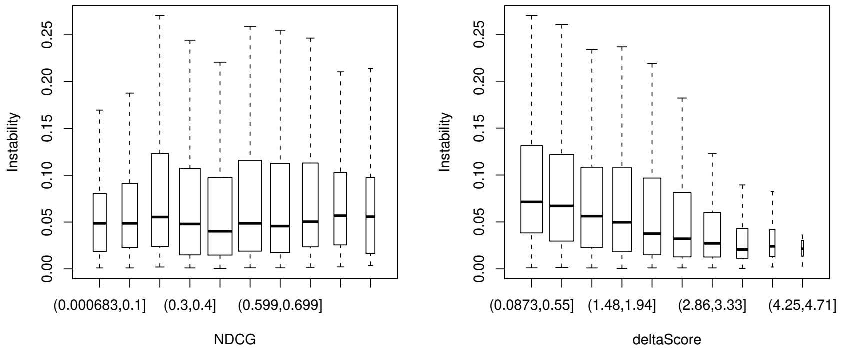

Fig. 9. The instability of queries with varying difficulty (left) and score difference between the 1st and the 5th document (right).

6 Conclusions

This paper presented the first large-scale study on the time instability of web search results. We suggested several classes of instability, and the metrics for quantifying it. Our data shows that top results in all major search engines have

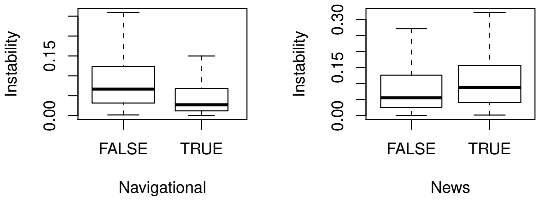

Fig. 10. The instability with respect to two query classes: Navigational and News.

similar levels of day-to-day fluctuation and that many of these changes are temporary. We investigated several factors that impact time instability of each query, and we found that longer queries are more unstable than their counterparts, and certain classes of queries such as navigational present lower instability than average, while other classes show higher than average instability, such as news queries.

Being the first large-scale analysis of this kind, our study can lead to various directions as future work. While previous studies provide some evidence that the the instability causes problem with re-finding, a more careful user study should be followed to quantify the impact of instability in user's search experience. Also, given the negative impact of instability in user perception and search quality measurement, we need to find out how to control unwanted fluctuations in search results.

7 Acknowledgements

The authors would like to thank Rich Caruana, Giridhar Kumaran, Manish Malik and Mohammad Makarechian for their helpful feedback during our project. This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of the sponsor.

References

1. E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The web changes everything: understanding the dynamics of web content. In *WSDM '09*, pages 282–291, New York, NY, USA, 2009. ACM.
2. D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. In *WWW '03*, pages 669–678. ACM Press, 2003.
3. A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from a search engine perspective. In *WWW '04*, pages 1–12. ACM Press, 2004.
4. E. Selberg and O. Etzioni. On the instability of web search engines. In *In Proceedings of RIAO '00*, pages 223–235, 2000.
5. J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *SIGIR '07*, pages 151–158, New York, NY, USA, 2007. ACM.