On the Volatility of Commercial Search Engines and its Impact on Information Retrieval Research

Jimmy
Queensland University of Technology
Brisbane, Australia
University of Surabaya (UBAYA)
Surabaya, Indonesia
jimmy@hdr.qut.edu.au

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zuccon@qut.edu.au

Gianluca Demartini
University of Queensland
Brisbane, Australia
g.demartini@uq.edu.au

ABSTRACT

We studied the volatility of commercial search engines and reflected on its impact on research that uses them as basis of algorithmical techniques or for user studies. Search engine volatility refers to the fact that a query posed to a search engine at two different points in time returns different documents. By comparing search results retrieved every 2 days over a period of 64 days, we found that the considered commercial search engine API consistently presented volatile search results: it both retrieved new documents, and it ranked documents previously retrieved at different ranks throughout time. Moreover, not only results are volatile: we also found that the effectiveness of the search engine in answering a query is volatile. Our findings reaffirmed that results from commercial search engines are volatile and that care should be taken when using these as basis for researching new information retrieval techniques or performing user studies.

1 INTRODUCTION

On a number of occasions, information retrieval researchers have used commercial search engines and associated APIs to assist with the research of new algorithms and techniques (type A: algorithmical use), or to investigate user search behaviour (type U: user study use). Examples of this practice include, among others: Cilibrasi and Vitanyi [4] defined a word similarity function based on the number of search results retrieved by Google (A); Symonds et al. [9] used Google to perform a first round of retrieval to inform query expansion (A); Maxwell et al. [8] used Bing to retrieve documents and snippets within a user study that explored user behaviour with respect to snippet length and informativeness trade-off (U).

To help understand the extent of this practice, we systematically surveyed the literature published in the ACM SIGIR conference between 2006 and 2016 (a total of 2,138 full and short papers)[1]. We found that 158 contributions (7.4%) used commercial search engines in their experiments[2].

[1]Note that this practice is well utilised also outside of the SIGIR literature, as demonstrated by the examples cited above that were not published in SIGIR.
[2]Data available at https://goo.gl/3BLVgz

SIGIR '18, July 08-12, 2018, Ann Arbor, Michigan, USA.

DOI:

Commercial search engines are however often *volatile*: both the results retrieved and their rankings often differ given two points in time. There are multiple reasons for this volatility. On the one hand, volatility may be due to index updates and fresher information being indexed by the search engines [2, 3]. On the other hand, volatility may be due to operational reasons such as index replication, sharding and routing. Finally, differences may be due to updates to the ranking functions used by the search engines.

Commercial search engine volatility has been investigated by Altingovde et al. [1] who had experimented using a set of 630,000 queries and found that only 10.7% of top 10 results found in 2007 remained as top 10 results in 2010. Bai and Junqueira studied volatility in Yahoo! and reported that of 1.4M search results analysed over 3 weeks, ≈ 35% new URLs were added, ≈ 1% had modified content, and ≈ 0.06% were deleted [2]. Of course, these modifications would have impacted the URLs that users would have visited over that period of time, but it is unclear whether by week 3 the relevancy of the retrieved URLs was comparable to that obtained at the start of the study. While it is understandable that a query will retrieve different results when submitted many years apart and may be less of a problem for IR research such as user studies (but still a problem for replicability), such findings are more concerning when volatility is detected among searches that occur within a short period, e.g., typical length of a user study (days or weeks).

Search engine volatility may be a problem when commercial search engines are used by researchers as part of their methods or user studies. In other words: if an algorithm or technique is based on the use of a volatile search service, differences in search results and rankings may vary the effectiveness of the method, or render the replication of the experiments impossible. Similarly, if a user study relies on a commercial search engine to investigate user behaviour, volatility may be a confounding factor affecting effectiveness, especially if the user studies are carried out over a period of time, rather than all being run concurrently. Yet, this aspect is often ignored when analysing the results, e.g., volatility is not considered as factor within an ANOVA analysis of results.

To further exemplify how search engine volatility may affect information retrieval research that relies on such commercial services, consider the case of a user study relying on the Bing search APIs and investigating the capabilities of users in selecting query suggestions automatically generated by techniques that exploit the search results obtained from the initial user's query. In such case, two components of the experimental methodology rely on results from the commercial APIs: (i) the query suggestion mechanism, and (ii) the results that are retrieved (and evaluated for effectiveness)

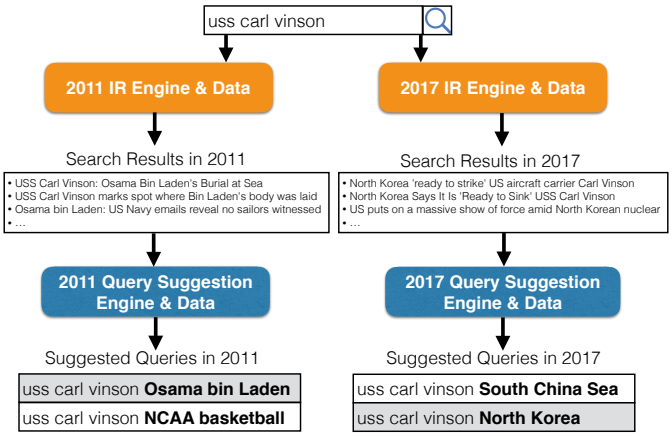

Figure 1: Example of search result volatility and its impact on an hypothetical query expansion techniques that exploits the retrieved results.

in response to the user's query and the selected query suggestion. Figure 1 shows an example of a query submitted at two different times. At each time, the query retrieved two (sensibly) different sets of results and thus returned different query suggestions.

In this study, we seek to answer the following research questions:

- RQ1: What amount of volatility do commercial search engine APIs present?
- RQ2: How does the volatility of commercial search engines affect information retrieval research?

To answer RQ1, we periodically used a search engine API to retrieve results for a large set of queries which had no specific temporal intent or seasonality effect. We then studied how results changed over time. To answer RQ2 we assessed the relevance of the top search results we collected over time and we analysed the change in search engine effectiveness over time. Details of the methods used in this study are described next.

2 METHODS

To answer RQ1, we acquired the queries used in the TREC 2013 and 2014 Web Track (100 queries in total). While these queries had no explicit temporal nor seasonal intent, a small number may have been influenced by temporal issues. For example, in our experiments, query 202: "uss carl vinson" was affected by the US decision of deploying the aircraft carrier within strike range of North Korea in early January 2018. We further acquired a set of 300 queries from the CLEF 2016 eHealth IR collection. Of the 300 queries, we removed query 129005 due to a problem with quotation mark characters in the query. These queries related to consumer health search intents, and were unlikely to be affected by temporal or seasonal intents.

We used the Bing Search API [3] to retrieve a maximum of top 50 web results in answer to the selected query sets, setting `English US` as the market and with `safe search` turned off. We performed retrieval every two days from 29 November 2017 to 31 January 2018 with exceptions of 13, 27, and 29 December 2017 where the retrieval process was not triggered due to technical problems. Hence, in total we collected 30 data samples for each query set. We used the data samples to investigate the volatility of search API by counting the number of new URLs between search results from different retrieval dates pairs. We then further investigated whether differences in ranking for an URL were also found over time.

To answer RQ2, we pooled the top ten URLs from every sampled date for the WEB2013-2014 query set, and we assessed their relevance[4]. Relevance assessments were collected using crowdsourcing. We setup tasks on Amazon Mechanical Turk, assigning each query-website pair to 5 workers. Workers, selected among those with a 90% acceptance rate and at least 1000 tasks completed, were presented with the TREC topic title and description fields and a link to the webpage to be assessed. We simplified the TREC 2013 six-point judgment scale [5] into the following four-point scale: Highly relevant (this point included Nav, Key, and HRel as defined in TREC 2013), Relevant, Not Relevant, and Junk. As suggested in [7], assessments that took less than 4 seconds were discarded. Collected assessments were aggregated as the median of the collected labels for each query-document pair. We analysed the results using ERR@10, nDCG@10, and P@10, as used in the TREC 2013.

3 EMPIRICAL EVALUATION

3.1 Volatility of Search Results

Figure 2 shows the percentage of new URLS introduced on average in the top 10 results. Each square in the heatmap corresponds to a pair of dates, and thus the percentage difference between the results obtained in the two dates. The darker the red tone, the higher the percentage of new URLs being returned on the later date: the diagonal is yellow, indicating no difference (as expected, as we are comparing a date with itself), and we removed the lower part of the heatmap for clarity (the heatmap is symmetric).

Results highlighted in blue refer to the percentage of new URLs retrieved compared to the initial date (the start of our data sampling): in the figure we further visualised this trend as a line plot to give the reader a different representation of the trend and aid interpretation. On average, we found that each day had 24.43% new URLs, compared to the initial sampling date.

Another important observation from Figure 2 is highlighted in green. This shows the percentage of new urls found on a sampling date compared to the previous sampling date. We further visualised this trend in the corresponding line plot: new urls were found at a comparable percentage overtime. On average, every two days, 10.72% of the URLs retrieved differed from those retrieved on the previous sampling date.[5]. Results using the top 50 retrieved documents showed similar trends (w.r.t. percentage); so did also the results for CLEF 2016 (available at https://goo.gl/3BLVgz).

We then further investigated the ranking distance between occurrences of the same URL across different dates. Results were again represented as a heatmap, which is reported in Figure 3. The heatmap shows the percentage of rank distances between the top 10 URLs for each pair of sampling dates using the WEB2013-2014

[3] https://azure.microsoft.com/en-au/services/cognitive-services/bing-web-search-api/

[4] We assumed that the content of the linked webpages remained the same throughout our experiments. While this may have not been the case for all results, as reported in [2], only $\approx$ 1% of urls found in the first two weeks had modified content in the third week of their study.

[5] Note we missed sampling on the 13, 27, and 29 December 2017: for the sampling date after the ones we missed, rankings were compared to those in the previous available date. In the line plot, we distinguished the data for these dates using red dots.

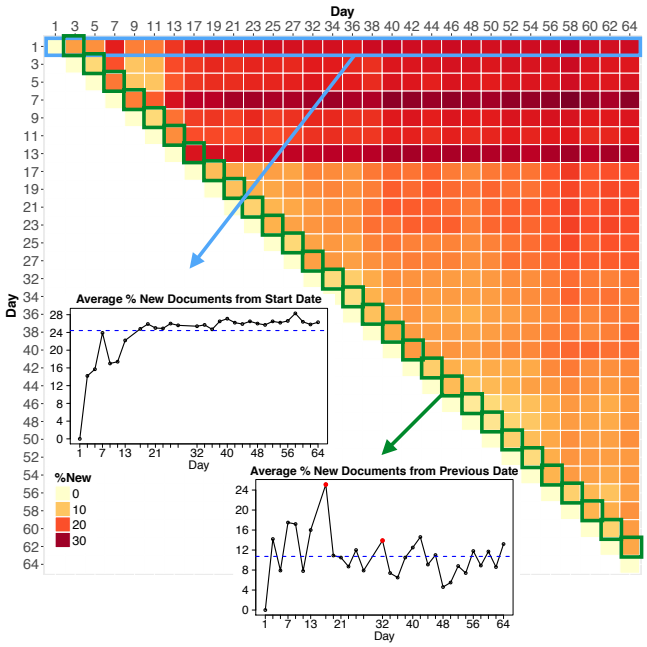

Figure 2: Percentage of average (over the query set) new URLs retrieved in the top 10 results for each pair of sampling dates for the WEB2013-2014 query set. Results for CLEF 2016 are available at https://goo.gl/3BLVgz.

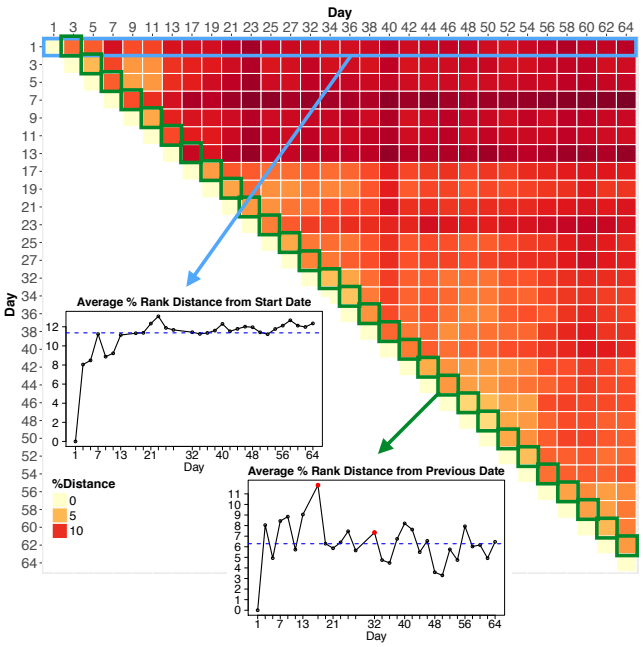

Figure 3: Percentage of average (over the query set) rank movement in the top 10 results for each pair of sampling dates for the WEB2013-2014 query set. Results for CLEF 2016 are available at https://goo.gl/3BLVgz.

query set. Blue and green colours were used to represent similar circumstances as for the previous heatmap. When considering rank movements over time with respect to the first sampling day (blue highlighting), we found that on average URLs moved 11.36% up or down the ranking, with lesser movement found in the first few days of the experiment. When considering rank movements over time with respect to the previous sampling day (green highlighting), we found that on average URLs moved by 6.29% up or down the ranking compared to the previous date, though peaks with larger rank movements did occur.

Given these results, we answer RQ1 by reporting that, on average, between two consecutive days, search engine results change by 10.72% in terms of new URLs retrieved (1.07 new URLs every 10). Furthermore, we found that the difference is even larger if a wider timespan is considered. In addition, we also report that URLs that occur in the results between two dates are likely to exhibit a rank movement of on average by 6.29%.

3.2 Impact of Result Volatility on Search Effectiveness

Figure 4 shows the search effectiveness over time for the WEB2013-2014 query set, averaged over all queries for each sampling date. The average trends show that search engine volatility had little impact on the average search effectiveness: despite new URLs were retrieved over time, and existing URLs changed rank, effectiveness on average did not vary significantly. Statistical significant differences (t-test $p < 0.05$) were found only between $\approx 6.7\%$ of the results for each pair of days (only unique pairs were considered).

The previous results analysed the impact of search engine volatility by averaging effectiveness over the query set. We next analyse

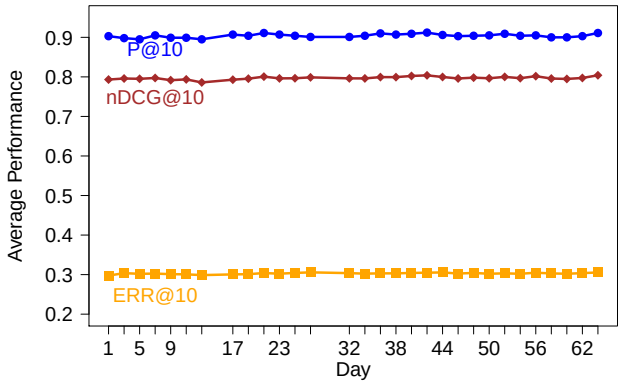

Figure 4: Average effectiveness of the top 10 URLs retrieved for the WEB2013-2014 query set.

the impact volatility had on a query-by-query basis; we did this for graded relevance (nDCG@10) – similar findings were observed for other settings. Results are reported in Figure 5. Box plots were organised such that queries were ordered in decreasing effectiveness of the results obtained on the initial date of sampling (day 1): each box summarises the effectiveness of a query over time. The box plots show effectiveness did vary over time for each query, with some queries achieving substantially different effectiveness depending on the date. Specifically, we found that 67 out of 100 queries had a change in nDCG@10 that was higher or equal to 0.1 and, on average, individual query effectiveness varied by 0.1431 over the sampling period, with the largest variation recorded being 0.4700. To further provide an intuition of the gap in effectiveness

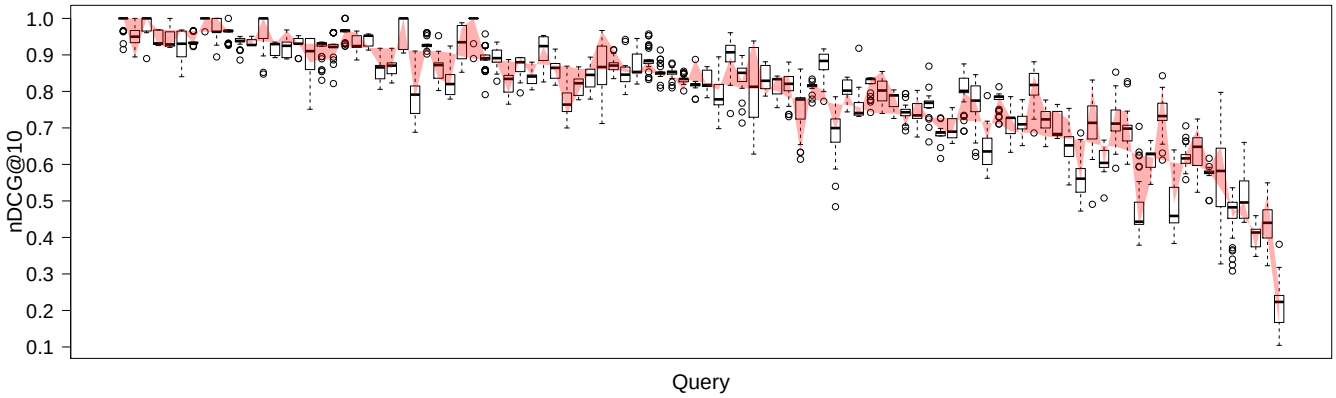

Figure 5: nDCG@10 for WEB2013-2014 queries over time. Each box plot refers to a query; queries are ordered in decreasing effectiveness as returned on day 1. The red shaded area indicates the effectiveness gap between results for day 1 and those for the day with the biggest average effectiveness gap over the query set.

that search result volatility generated, we highlighted the gap between the effectiveness of each query recorded on the first day of our experiment, and the day with the biggest average effectiveness gap over in the query set (day 64). This gap is represented by the red shaded area in Figure 5.

4 DISCUSSION

The findings from our experiments quantified the amount of volatility measured in 2-day time intervals and over longer periods. In addition, they also highlighted that not only are the search engines volatile, but their effectiveness is also volatile, given a query; although we found that for the query set used, average effectiveness remained mostly unchanged.

These findings suggest that search engine result volatility is likely to largely impact the *replicability* of results obtained by exploiting commercial search engine APIs either for algorithmical advances or within user studies. We also argue that volatility also impacts *reproducibility*, as the deterioration of results over time for some queries is large and, if results are used algorithmically, is likely to produce different outcomes.

As an illustrative example of the effect search engine volatility may have on the evaluation of information retrieval research, we consider the work of Gao et al. [6]. They proposed a number of novel query expansion techniques that exploit results retrieved from the search engine in answer to the original query. To evaluate their methods, they used a commercial search engine. Specifically, they submitted the original queries and the expanded queries to the search engine and judged the relevance of the top ten retrieved results. In their experiments, nDCG@10 improved from 0.3905 (original queries) to 0.4265 (best expansion technique) – a gain of 0.036 (19%). If we revisit those results in light of the findings of Section 3.2, it is expected that it will be impossible to replicate the same results they obtained. In addition, it becomes unclear whether the reported gains would be observed again if the results retrieved from the original queries (and their relevance) changed with the same magnitudes observed in our experiments. While we do not investigate this in the current study, we aim to empirically investigate the effect of search engine volatility on methods such those of Gao et al. [6], Symonds et al. [9], and Cilibrasi and Vitany [4] in future work.

It is unclear how researchers could mitigate the issues related to search engine volatility and yet use commercial search engines and associated APIs within their research. A possible avenue may be repeating the experiments over a sufficiently long period of time, so as to account for search engine volatility as one of the factors affecting results and study this with respect to their results.

5 CONCLUSION

In this paper, we investigated the volatility of commercial search engines and its impact on information retrieval research. By sampling the results returned by the Bing Web Search API every two days for a period of 64 days, we found that, on average, the search engine retrieved 10.72% new URLs in the top 10 ranks. Additionally, we also found that a URL that was retrieved on a previous date was subject to an average rank movement of 6.29%. When examining the possible impact such a volatility may have on information retrieval research that makes use of such search services, we found that on average, nDCG@10 varied by 0.1431 (19.88%) for each query and the biggest nDCG@10 variation for a query was 0.4700 (143.51%). These results suggest that research that uses commercial search engines as part of an algorithmical pipeline or user study should be aware of search engine volatility and its implications.

ACKNOWLEDGMENTS

Jimmy is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP). Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579).

REFERENCES

[1] I Altingovde, R Ozcan, and O Ulusoy. 2011. Evolution of web search results within years. In *SIGIR'11*.
[2] X Bai and F Junqueira. 2012. Online result cache invalidation for real-time web search. In *SIGIR'12*.
[3] R Blanco, E Bortnikov, F Junqueira, R Lempel, L Telloli, and H Zaragoza. 2010. Caching search engine results over incremental indices. In *SIGIR'10*.
[4] R Cilibrasi and P Vitanyi. 2007. The google similarity distance. *TKDE* 19, 3 (2007).

[5] K Collins-Thompson, P Bennett, F Diaz, C Clarke, and E Voorhees. 2014. TREC 2013 web track overview. In *TREC*.
[6] J Gao, G Xu, and J Xu. 2013. Query expansion using path-constrained random walks. In *SIGIR'13*.
[7] E. Maddalena, M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *HCOMP'16*.
[8] D Maxwell, L Azzopardi, and Y Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *SIGIR'17*.
[9] M Symonds, P Bruza, G Zuccon, B Koopman, L Sitbon, and I Turner. 2014. Automatic query expansion: A structural linguistic perspective. *JAIST* 65, 8 (2014).