ABSTRACT

People tend to type short queries, however, the belief is that longer queries are more effective. Consequently, a number of attempts have been made to encourage and motivate people to enter longer queries. While most have failed, a recent attempt—conducted in a laboratory setup—in which the query box has a halo or glow effect, that changes as the query becomes longer, has been shown to increase query length by one term, on average. In this paper, we test whether a similar increase is observed when the same component is deployed in a production system for site search and used by real end users. To this end, we conducted two separate experiments, where the rate at which the color changes in the halo were varied. In both experiments users were assigned to one of two conditions: halo and no-halo. The experiments were ran over a fifty day period with 3,506 unique users submitting over six thousand queries. In both experiments, however, we observed no significant difference in query length. We also did not find longer queries to result in greater retrieval performance. While, we did not reproduce the previous findings, our results indicate that the query halo effect appears to be sensitive to performance and task, limiting its applicability to other contexts.

KEYWORDS

Query length; Halo; A/B testing; Research replication

1 INTRODUCTION

Providing a more detailed and richer description of an information need by issuing a longer query has been considered as a fundamental way for users to indicate to the system what is relevant [7]. However, most users tend to express very short—two to three—term queries [3, 6, 13]. As a result a number of efforts have been made which attempt to lead people to longer queries [2, 7, 8, 11]. It is commonly believed that longer queries (and thus a richer description of the user’s underlying information need) will result in better retrieval performance [7].

In this paper, we explore the query halo effect, proposed in [2], where the search box glows and changes colour as the user types. In [2], Agapie et al. showed that the this halo effect led people to enter longer queries in the context of complex Web search tasks. We now attempt to reproduce this finding in a different context: site search. Crucially, while [2] was performed in a laboratory setting only, we attempt to reproduce the findings in a live setting with users issuing queries to fulfill their own information needs. To this end, we consider the following Research Questions:

RQ1 Does the query halo effect lead people to longer queries in a natural setting? Or alternatively phrased: can we reproduce the findings in [2] in a different context?

RQ2 Does the assumption of longer queries being more effective hold in this setting?

We empirically investigated these questions in a 50-day long A/B test setup implemented in a university site search engine. Our analyses are based on more than 6,000 submitted queries in that time period. We find that the query halo effect did not entice people to submit longer queries: there was no significant difference between the halo condition and no-halo condition across the two experiments performed. We also find that longer queries, in this context, do not necessarily result in better retrieval performance. We hypothesize that for the query halo effect to take hold, the retrieval performance needs to be positively correlated with query length and the search task needs to be complex as in [2].

2 BACKGROUND

Numerous studies have examined the length of queries submitted by users to search systems in a number of different settings. For Web-based queries the average number of terms is around 2.3 (2.35 Excite logs [13], 2.34 AOL logs [3], 2 (mode) MS logs [6]), while site search queries tend to be slightly shorter with a term length around 2.2 (2 (mode) UTennesse log [14] and 2.2 U.S. Government logs [9]). Since queries are typically short in the Web setting, a driving hypothesis for leading people to longer queries is based on the strong assumption that: longer queries result in better retrieval performance. However, in previous works [2, 7, 8, 11] this was not explicitly tested. In [4], Azzopardi studied the relationship between query length and retrieval performance (mean average precision) in the context of ad-hoc search on a number of TREC test collections using best match retrieval algorithms. In this simulated batch setup, performance does increase, but at a diminishing rate of return; queries of 2 to 3 terms in length result in the highest rate of return. With respect to the previous and present study on leading people to longer queries, the analysis confirmed that query length and performance are positively correlated, but only in a specific context. Production search systems, however, often use strategies beyond best match algorithms—typing more query terms often reduces the number of results returned as terms are implicitly ANDed together.
In the aforementioned studies, performance, in terms of precision-recall based measures, is not reported or considered, so it is actually unclear whether an increase in query length would translate into improved performance – and thus increased satisfaction for the user. In a recent theoretical cost-benefit analysis of query length and retrieval performance Azzopardi and Zuccon [5] show that in order for users to increase their query length either the retrieval performance needs to increase or the cost of entering terms needs to be reduced. In previous works on leading people to longer queries, neither the performance of the system is increased, nor is the cost of entering a query reduced. As such, their theory suggests that in practice larger query boxes or halo effects, which do not modify either of these, are not going to lead to longer queries.

3 REPRODUCING THE HALO EFFECT

Reproducibility, the ability of an experimental study to be duplicated independently, recently gained renewed attention in computer science and especially in information retrieval. The ACM’s policy on Result and Artifact Review and Badging [10] distinguishes repeatability (same team, same experimental setup), replicability (different team, same experimental setup), and reproducibility (different team, different experimental setup). Ferro et al. [1] investigated the problems and approaches to reproducibility in information retrieval and various other sub-fields of computer science.

In line with those initiatives, our experiments aim to reproduce the work by Agapie et al. [2] in a live search engine. Experiments were run on the site search engine of the University of Twente1, a federated search engine [12] searching 35 resources including Google’s site search, local courses, local news, the telephone directory, the university timetables, as well as results from the university’s social media feeds, such as Facebook, Twitter and Flickr. Given a query2, the search engine returns ranked resource blocks with each block containing up to four (up to seven in the case of images) ranked items; each resource can only contribute a single block to a ranking. Figure 1 shows an example result page for the query “library” with three ranked blocks. Our query log records contain for each query the URLs of the search results that were clicked as well as the block rank of the clicked result. As the vast majority of queries yielded a single click, we computed the block rank MRR (the mean reciprocal rank over all submitted queries in the respective condition) as our system’s measure of retrieval performance.

The implementation uses the open source federated search engine Searsia3. Experiments were run as A/B tests where users were assigned randomly to either the control condition (the standard search box, labelled henceforth as no-halo condition) or the experimental condition (the search box with the query halo effect, labelled as halo condition).

The query halo effect was implemented as described by Agapie et al. [2]. The empty query input box has a pink (RGB #FF1493) halo that surrounds the text box. As the user types into the box, the halo begins to change color – becoming less and less pink, and then starts becoming more and more blue (RGB #3366CC). To adhere to

---

1https://utwente.nl/search
2In the 50-day period of our study, the five most popular queries were minor, matlab, library, ces and eduroam.
3http://searsia.org
the university site’s style, the halo surrounding the query box was slightly less wide than in [2]. Figure 2 shows an example of the halo effect for queries of different lengths; Figure 3 shows the full color spectrum of the halo.

In [2], the color was interpolated between pink and blue, with queries of seven words or longer showing the bluest halo. While not explicitly noted in [2], the choice of colours seemed to be decided based on accessibility, as for example, red/pink to green may not have been distinguishable (due to red/green colour blindness). The transition from pink to blue was based on word boundaries.

4 RESULTS

4.1 Experiment 1: term halo vs. no-halo

The first experiment ran between January 6, 2017 and January 31, 2017. In this period, in the no-halo condition, 884 different users submitted 1,623 queries (1,301 unique). Assigned to the halo condition were 803 users who issued 1,367 queries (1,122 unique).

Table 1 (top) contains the results of this experiment. Users in the control condition submit on average 2.18 query terms, users in the halo condition match this almost perfectly with an average query length of 2.16. Splitting the users according to their experience with the search system – users submitting a single query across the three weeks (single query users) vs. users submitting two or more queries (2+ query users) – does not yield a different picture. Figure 4 shows the query distribution across query term length, the halo condition does not lead users to change their querying behaviour with respect to query length. More than 70% of all issued queries contain either 1 or 2 terms.

4.2 Experiment 2: character halo vs. no-halo

The second experiment ran between February 1, 2017 and February 25, 2017. In the no-halo condition, 857 different users submitted 1,395 queries (1,157 unique). In the halo condition, 919 users issued 1,641 queries of which 1,314 were unique. In Table 1 the results of this experiment are summarized. The more dynamic change of the halo has no effect: users in both the control and experimental conditions submit queries of similar length. Overall, we conclude that the query halo effect does not lead to longer queries in this context and live setting.

4.3 Longer queries lead to more clicks?

In order to evaluate the suitability of the assumption that longer queries perform better, we compute the block MRR across all the 6,026 queries in our 50-day log (combined over conditions and experiments). Figure 5 shows that in our production system the common assumption that longer queries perform better does not hold: longer queries do not lead to more clicks on better ranked result blocks, and thus do not lead to higher MRR. Queries with 2 to 4 terms perform similarly. In order to verify the quality of the

---

4 Specifically, the query box’s CSS property box-shadow contains a spread-radius of 3 pixels, the query box’s border-color was set to the same color as the halo.
Table 1: Overview of the average query length (and standard deviation) in terms (Experiment 1) and characters (Experiment 2). None of the differences are significant.

<table>
<thead>
<tr>
<th></th>
<th>no-halo</th>
<th>halo</th>
</tr>
</thead>
<tbody>
<tr>
<td>All queries</td>
<td>2.18 (1.36)</td>
<td>2.16 (1.32)</td>
</tr>
<tr>
<td>Single query users</td>
<td>2.06 (1.11)</td>
<td>2.15 (1.40)</td>
</tr>
<tr>
<td>2+ query users</td>
<td>2.25 (1.48)</td>
<td>2.17 (1.27)</td>
</tr>
</tbody>
</table>

Experiment 2 (characters)

<table>
<thead>
<tr>
<th></th>
<th>no-halo</th>
<th>halo</th>
</tr>
</thead>
<tbody>
<tr>
<td>All queries</td>
<td>16.52 (11.81)</td>
<td>16.39 (10.18)</td>
</tr>
<tr>
<td>Single query users</td>
<td>16.44 (12.03)</td>
<td>15.39 (10.14)</td>
</tr>
<tr>
<td>2+ query users</td>
<td>16.58 (11.65)</td>
<td>17.00 (10.16)</td>
</tr>
</tbody>
</table>

Figure 4: Number of queries issued for each condition in Experiment 1.

Figure 5: Mean reciprocal rank (block-based) computed over all queries in our log.

Figure 6: Frequency of search result clicks on ranked blocks computed over all queries in our log.

5 CONCLUSIONS

In this paper, we attempted to reproduce the “query halo” effect observed (in a lab setting over complex search tasks) by Agapie et al. [2] in the context of site search with users serving their own information needs. We were not able to reproduce the effect, instead observing no differences in the control vs. experimental conditions. Furthermore, in contrast to a common assumption, we found little difference in retrieval performance across query length (between 2-4 terms). These results motivate further work in unpicking the complexities and relationships between query length, performance, task and motivators.

ACKNOWLEDGMENTS

We are grateful to Anne Heining of U. Twente Media & Communication for supporting the A/B tests.

REFERENCES


