# Exploring the Query Halo Effect in Site Search

Leading People to Longer Queries

Djoerd Hiemstra
University of Twente
Enschede, The Netherlands
hiemstra@cs.utwente.nl

Claudia Hauff
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

Leif Azzopardi
University of Strathclyde
Glasgow, Scotland
leif.azzopardi@strath.ac.uk

## 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, varying the rate at which the color changes in the halo were induced. 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

**ACM Reference format:**
Djoerd Hiemstra, Claudia Hauff, and Leif Azzopardi. 2016. Exploring the Query Halo Effect in Site Search. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

## 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 **R**esearch **Q**uestions:

**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.

## 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

*Conference'17, Washington, DC, USA*
© 2016 ACM. 978-x-xxxx-xxxx-x/YY/MM...$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

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. So it is not clear whether the query-length assumption holds in such a setting or for other tasks.

Karlgren and Franzén [11] were one of the first to try and lead people to longer queries in a laboratory setup by changing the query input design. They modified the single-line query input box used by Web search engines and designed a query text box with multiple lines, such that the query terms would wrap as you typed – in the belief that this would illicit longer queries than the single-line query box. In the context of Web search, Karlgren and Franzén found indeed that participants entered significantly longer queries using the multi-line input box. Belkin et al. [7] later hypothesized that this increase in query length was due to: *(i)* the larger perceived space and *(ii)* the visibility of the entire query (as in the single-line condition the query could be partially hidden if it is too long).

Belkin et al. [7] attempted to replicate Karlgren and Franzén's finding in a slightly different context and also explored whether instructing users to enter questions as their query as opposed to key words, would lead to longer queries. In the context of a lab-based setting (as part of the TREC interactive track), their results showed that when participants were in the "instruction" condition, they submitted significantly longer queries. This is not too surprising, because turning a query for "Garfield" into a question will require at least one more term, e.g. "Who is Garfield?". Interestingly, they were unable to reproduce Karlgren and Franzén's finding with respect to the different query box input modes. They found no difference between the multi-line and single line query box modes.

Belkin et al. [8] then followed up, by designing an experiment where the baseline interface was a multi-line query box and the experimental condition which used the same interface but included instructions to the participants: "Information problem description (the more you say, the better the results are likely to be)". This instruction led the lab study participants in the experimental condition to enter on average two more query terms than the participants in the control condition.

Agapie et al. [2] hypothesized that a halo around the query box that reflects the length of the query being created could mitigate people's tendency to issue short queries. Specifically, they further hypothesized that this halo effect would "nudge" people to type in more query terms. In their experiments, they set up a 2x2 factorial design, where on one dimension they included a halo vs. no-halo, and on the other dimension they provided instructions vs. no instructions. In the instruction condition, the 61 lab study participants were told that, "[the] system performs better with longer queries." Agapie et al. found the halo to lead the participants to construct significantly longer queries. However, drilling down, this was most pronounced when the halo was present, but no instructions (6.6 query terms on av.). When the halo *and* instructions were provided, they did not observe any significant increase in query length (4.5 query terms on av.). When only instructions were provided it resulted in longer queries (5.3) when compared to no halo and no instruction (4.2). In a subsequent experiment, they compared four conditions: no-halo, halo (pink to blue), inverted halo (blue to pink) and a static halo (blue). Again, they observed that the halo effect (pink to blue) resulted in significantly longer queries than no halo or a static halo. However the inverted halo did not lead people to significantly longer queries.

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.

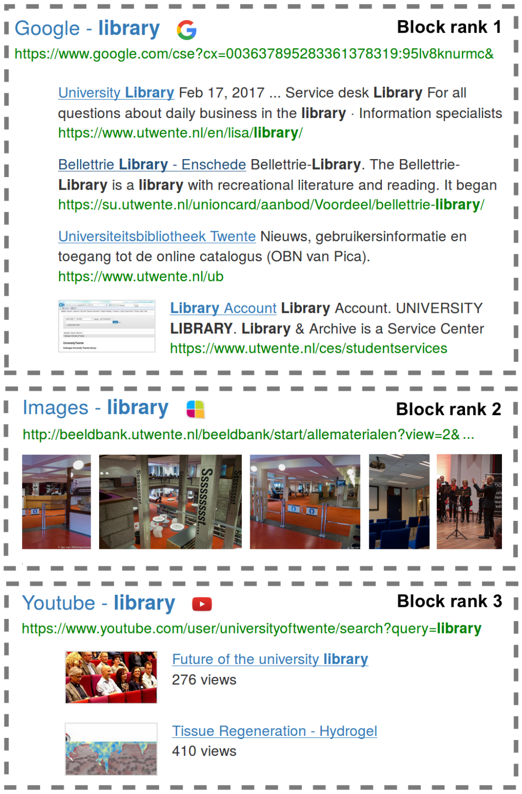
In line with those initiatives, our experiments aim to reproduce the work by Agapie et al. [2] in a live search engine. Experiments were run on the site search engine of the University of Twente[1], a federated search engine [12] searching 35 resources including Google's site search, local courses, local 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 query[2], 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 Searsia[3]. 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).

[1] https://utwente.nl/search

[2] In the 50-day period of our study, the five most popular queries were *minor*, *matlab*, *library*, *ces* and *eduroam*.

[3] http://searsia.org



Google - **library** **Block rank 1**
https://www.google.com/cse?cx=003637895283361378319:95lv8knurmc&

University **Library** Feb 17, 2017 ... Service desk **Library** For all questions about daily business in the **library** · Information specialists
https://www.utwente.nl/en/lisa/**library**/

Bellettrie **Library** - Enschede Bellettrie-**Library**. The Bellettrie-**Library** is a **library** with recreational literature and reading. It began
https://su.utwente.nl/unioncard/aanbod/Voordeel/bellettrie-**library**/

Universiteitsbibliotheek Twente Nieuws, gebruikersinformatie en toegang tot de online catalogus (OBN van Pica).
https://www.utwente.nl/ub

**Library** Account **Library** Account. UNIVERSITY **LIBRARY**. **Library** & Archive is a Service Center
https://www.utwente.nl/ces/studentservices

Images - **library** **Block rank 2**
http://beeldbank.utwente.nl/beeldbank/start/allematerialen?view=2& ...

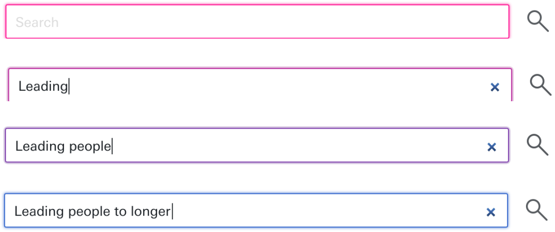
Youtube - **library** **Block rank 3**
https://www.youtube.com/user/universityoftwente/search?query=**library**

Future of the university **library**
276 views

Tissue Regeneration - Hydrogel
410 views

**Figure 1: Example block-based ranking.**

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 the university site's style, the halo surrounding the query box was slightly less wide[4] 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.



Search

Leading

Leading people

Leading people to longer

**Figure 2: Example of the character transition halo effect.**

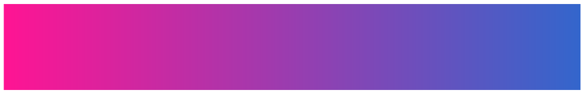


**Figure 3: Overview of the halo's full color spectrum. The extrema encode 0 terms (or 0 characters) on the left and 7 terms (or 22 characters) on the right.**

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. However to avoid it appearing mechanistic, a small (up to 100 millisecond delay) was randomly introduced.

In our work, we implemented two versions of the query halo: *(i)* a term-based transition, as done in [2] where after seven terms the halo was bluest, and *(ii)* a character-based transition query halo box, where the transition was based on characters and moved more smoothly from pink to blue, where after 22 characters the halo was the most bluish. This version shows a more immediate change of colour for every typed character.

## 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 submitted 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 the query halo effect to not be reproducible in our 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 clicks on better ranked result

[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.**

| **Experiment 1 (terms)** | | |
|---|---|---|
| | no-halo | halo |
| All queries | 2.18 (1.36) | 2.16 (1.32) |
| Single query users | 2.06 (1.11) | 2.15 (1.40) |
| 2+ query users | 2.25 (1.48) | 2.17 (1.27) |
| **Experiment 2 (characters)** | | |
| | no-halo | halo |
| All queries | 16.52 (11.81) | 16.39 (10.18) |
| Single query users | 16.44 (12.03) | 15.39 (10.14) |
| 2+ query users | 16.58 (11.65) | 17.00 (10.16) |

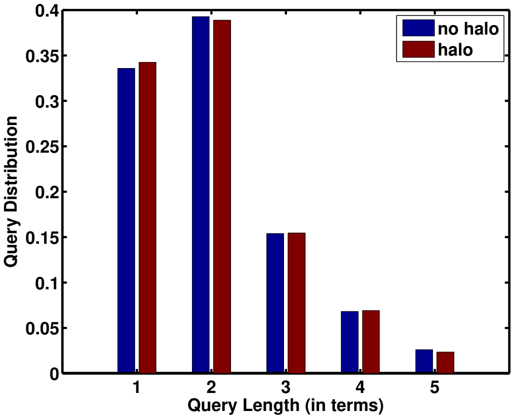


**Figure 4: Number of queries issued for each condition in Experiment 1.**

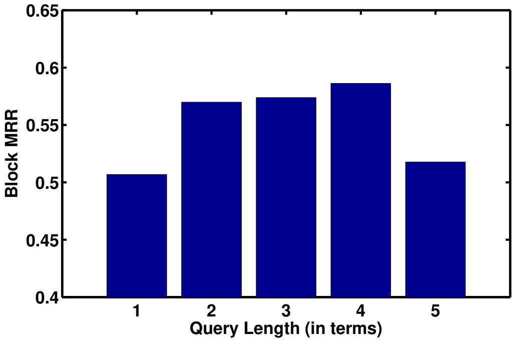


**Figure 5: Mean reciprocal rank (block-based) computed over all queries in our log.**

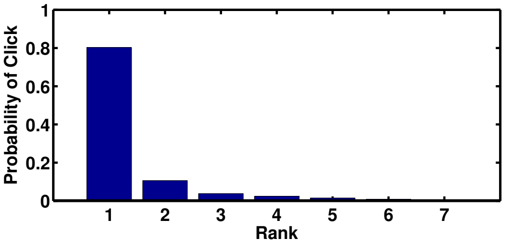


**Figure 6: Frequency of search result clicks on ranked blocks computed over all queries in our log.**

blocks, so do not lead to a higher MRR. Queries with 2 to 4 terms perform similarly. In order to verify the quality of the system, in Figure 6 we plot the distribution of search result across the ranked blocks: the vast majority (80%) of clicks appear within the first block as one would expect for a functioning production system.

## 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 there 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

[1] Nicola Ferro and Norbert Fuhr and Kalervo Järvelin and Noriko Kando and Matthias Lippold and Justin Zobel 2016. Increasing Reproducibility in Information Retrieval: Findings from the Dagstuhl Seminar on Reproducibility of Data-Oriented Experiments in e-Science. *SIGIR Forum* 50, 1, 68–82.

[2] Elena Agapie, Gene Golovchinsky, and Pernilla Qvarfordt. 2013. Leading people to longer queries. In *Proceedings of SIGCHI'13.* 3019–3022.

[3] Avi Arampatzis and Jaap Kamps. 2008. A Study of Query Length. In *Proceedings of SIGIR'08.* 811–812.

[4] Leif Azzopardi. 2009. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In *Proceedings of SIGIR'09.* 556–563.

[5] Leif Azzopardi and Guido Zuccon. 2016. An Analysis of the Cost and Benefit of Search Interactions. In *Proceedings of ICTIR'16.* 59–68.

[6] Peter Bailey, Ryen W. White, Han Liu, and Giridhar Kumaran. 2010. Mining Historic Query Trails to Label Long and Rare Search Engine Queries. *ACM Trans. Web* 4, 4, Article 15.

[7] Nicholas Belkin, Colleen Cool, Judy Jeng, A. Keller, Diane Kelly, Jay Kim, Hyuk Lee, M.-C. Tang, and X.-J. Yuan. 2001. Rutgers TREC 2001 Interactive Track Experience. In *The 10th text retrieval conference (TREC).*

[8] Nicholas Belkin, Diane Kelly, G. Kim, Jay Kim, Hyuk Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and Colleen Cool. 2003. Query Length in Interactive Information Retrieval. In *Proceedings of SIGIR'03.* 205–212.

[9] Michael Chau, Xiao Fang, and Olivia R. Liu Sheng. 2005. Analysis of the Query Logs of a Web Site Search Engine. *J. Am. Soc. Inf. Sci. Technol.* 56, 13, 1363–1376.

[10] Association for Computing Machinery. 2016. Result and Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging.

[11] Kristofer Franzén and Jussi Kalgren. 1997. Verbosity and interface design. In *SICS Technical Report: T2000:04, Retrieved online at:* http://soda.swedish-ict.se/2623/2/irinterface.pdf on May 11, 2013.

[12] Milad Shokouhi and Luo Si. 2011. Federated search. *Foundations and Trends in Information Retrieval* 5, 1 (2011), 1–102.

[13] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large Web search engine query log. *SIGIR Forum* 33, 1, 6–12.

[14] Peiling Wang, Michael W. Berry, and Yiheng Yang. 2003. Mining Longitudinal Web Queries: Trends and Patterns. *J. Am. Soc. Inf. Sci. Technol.* 54, 8, 743–758.