Leading People to Longer Queries

Elena Agapie
Harvard University
33 Oxford St, Cambridge, MA 02138
eagapie@seas.harvard.edu

Gene Golovchinsky
FX Palo Alto Laboratory, Inc.
3174 Porter Drive, Palo Alto, CA 94304
gene@fxpal.com

Pernilla Qvarfordt
FX Palo Alto Laboratory, Inc.
3174 Porter Drive, Palo Alto, CA 94304
pernilla@fxpal.com

ABSTRACT
Although longer queries can produce better results for information seeking tasks, people tend to type short queries. We created an interface designed to encourage people to type longer queries, and evaluated it in two Mechanical Turk experiments. Results suggest that our interface manipulation may be effective for eliciting longer queries.

Author Keywords
Interactive information seeking; query construction; persuasive computing

ACM Classification Keywords
H.5.m

INTRODUCTION
Keyword queries are a familiar way of representing information needs. Research literature shows that longer keyword queries are more effective at retrieving useful documents in exploratory search [3, 4]. Although users formulate more diverse queries when having difficulty finding results [8], it is well documented that people tend to run short queries [2].

It is also interesting to note work on shortening long queries to improve precision of search results (e.g., [9]). While shortening to improve query clarity [6] and coherence [8] may be useful, in many cases longer queries may be desirable either to improve recall or to refine results in topics with many documents. However, shorter queries become less reliable in situations with insufficient information on the relative utility of relevant documents. In these more complex search situations, longer queries are more likely to efficiently retrieve the desired information.

We hypothesized that this propensity to create short queries could be mitigated through a novel interaction design that used a halo to reflect the length of the query being constructed. We hypothesized that a pleasant, affective design [7] that modifies the visual characteristics of the text input area would nudge people to type more query terms.

We tested the effectiveness of this visualization by running two Mechanical Turk studies that asked people to find information on topics we constructed. Experimental results suggest that an interactive halo around the search box is effective in encouraging people to construct longer queries, although the phenomena are complex.

CHI 2013, April 27–May 2, 2013, Paris, France.

USER INTERFACE
Our interface design creates a halo around the query text box that varies in color and size with the length of the query being constructed. We chose the halo as the feedback mechanism because it is a familiar interface element that is visually unobtrusive and does not compete for attention in text-based query construction tasks.

The initial state of the text box is shown in Figure 1: a soft pink halo with a radius of about 20 pixels surrounds the text box. As the user starts to type a query, the halo becomes progressively less pink. After the query reaches a certain minimum length, the interface settles on a cooler, bluish tone (Figure 4).

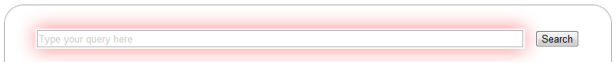

Figure 1. An empty query box has a pink halo.

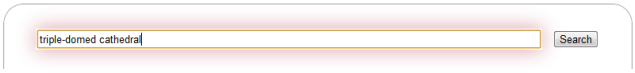

Figure 2. As the person starts to type, red hue starts to fade.

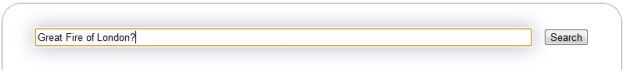

Figure 3. As the query gets longer, the halo becomes progressively bluer.

Figure 4. Long queries are displayed with a blue halo.

In our web-based implementation, the box-shadow CSS property was used to set the color and size of the halo. The color was interpolated between the two extremes, with queries of seven words or longer showing the bluish color.

While it is trivial to compute the query length in words and to set the halo to the associated color, we thought that a mechanistic application of the function might undermine its persuasive quality. Instead, we chose to mask the relationship between query length and halo color by animating the change through a variable duration (0-1000 milliseconds). The goal was to generate a correlated visual

effect that is non-trivial to understand based on casual observation.

EXPERIMENT 1

To investigate the subtle effect of a halo on users' query entry behavior, we built a small search interface on top of the Bing search API and deployed it through Mechanical Turk.

Search tasks

A-Google-a-day (http://agoogleaday.com) search puzzles were used to simulate complex information needs. These puzzles are designed to encourage people to learn how to solve complex information needs, and thus they were good proxies for our experiment. We selected the following three older puzzles (from April 2012) to reduce the chances of our participants recognizing them:

1. How many churches were built by the architect of the world's first triple-domed cathedral (and his office) after the Great Fire of London?
2. What tree does a mammal with fingerprints (primates excluded) rely on for food?
3. What material fuses with lime and soda to create an item on your dinner table that's considered to be an amorphous solid?

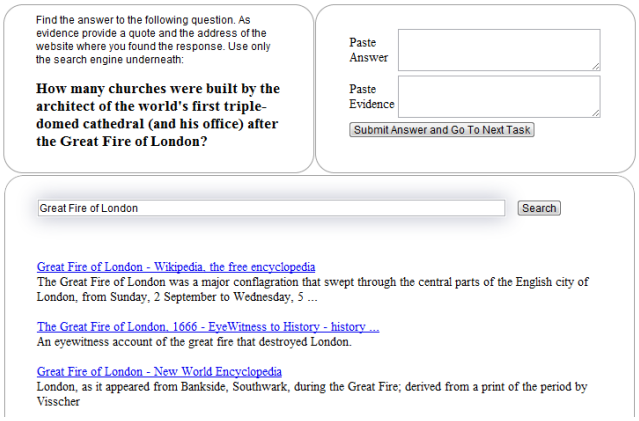

Figure 5. Experimental interface

Experimental software

Our search system included a web-based interface for eliciting queries, for presenting search results, and for collecting the answers to the search tasks (Figure 5). In addition to reporting the answer to the search question, we wanted people to record where or how they found the information. The goal was to discourage people from spamming the experiment and from having them fill in the answer as a guess or based on prior knowledge.

Queries were executed using the Bing search API[5] and results were filtered to remove any reference to the terms 'google' and 'google a day.' We also discouraged people from running the puzzle question as a query because we wanted to elicit query-formulating activity rather than copy-and-paste activity. Participants were shown the following instructions prior to starting the experiment:

We are testing the performance of a new search engine we developed. To test our search engine we will ask you to use it to answer search puzzles. Do not use other search engines or resources other than the ones provided by the search engine results to find the answer to the task.

You can only use the search engine we make available to you in the content of the hit.

Experimental design

The experiment had a 2 x 2 factorial between-subjects design. The factors were the presence or absence of a halo (*Halo* condition) on the search box and the presence or absence of a statement (*Instruction* condition) following the experimental instructions telling participants that "our system performs better with longer queries."

Each subject performed three search tasks in random order in a randomly-assigned experimental condition. Subjects were paid $1.55 through Mechanical Turk upon completion of all tasks. We restricted participants to be based in the United States and required them to have a 98% or better HIT completion rate. Because we were interested in queries people created, 91 queries that were copies of task questions or duplicates submitted within a second of each other were excluded from the dataset; analysis of the remaining 451 queries follows.

Hypotheses

The first experiment was designed to test two hypotheses: 1) the halo visualization around the query box would result in people typing longer queries, and 2) instructions about the effect of longer queries would result in longer queries.

Results

One hundred participants started the experiment, 61 of whom completed it. Table 1 lists the breakdown of participants by experimental condition, the average number of queries by condition, and the average query length per condition. It is worth noting that the average query length for this experiment was considerably longer (5.1 words/query) than those typically reported for web searches [2], which range from two to four words. On average, the participants used 3.2 (SD=1.32) queries to solve the Church task, 2.4 (SD=1.62) to solve the Tree task, 2.4 (SD=1.58) queries to solve the Material task.

A two-way ANOVA test was conducted to assess if the Halo and Instruction variables affect the word length of queries typed by the participants. We found a significant main effect of Halo ($F(1,447) = 5.1, p < 0.05$) indicating that participants type longer queries in the presence of a Halo than in its absence. We also found a nearly-significant main effect of Instruction ($F(1, 447) = 3.4, p = 0.064$). This does not provide enough evidence to conclude that the presence of Instructions results in longer queries. Finally we observed a statistically-significant interaction effect between Halo and Instruction ($F(1,447) = 41.7, p < 0.001$), Figure 6.

Table 1. Performance by condition (Experiment 1).

Condition	N	No. of Queries		Query length	
		Mean	SD	Mean	SD
Total	61	7.39	3.50	5.1	2.75
Halo	30	7.57	4.02	5.4	2.90
No halo	31	7.23	2.99	4.8	2.58
Instruction	32	7.81	3.64	4.9	2.59
No Instruction	29	6.93	3.36	5.4	2.60
Halo, No Instr.	14	6.71	4.10	6.6	3.28
Halo, Instr.	16	8.31	3.92	4.5	2.23
No Halo, No Instr.	15	7.13	2.61	4.2	2.01
No Halo, Instr.	16	7.31	3.38	5.3	2.89

To understand the interaction effect, we performed a Tukey HSD *post hoc* test. It showed that the *Halo with no instruction* condition outperformed all others: its queries had 2.1 words more on average than *Halo with instruction* (p<0.001), 2.3 more words than the *No halo with no instructions* condition (p<0.001), and finally, 1.2 words more than the *No halo with instructions* (p<0.01). These results further strengthen the conclusion from the main effect that a halo results in longer queries. We also found that the *No halo, Instruction* condition on average had 1.1 more words than the *No Halo, No instruction* condition (p<0.01). This result suggests that when no other factor is involved, people will enter longer queries with textual instructions.

EXPERIMENT 2

The second experiment focused on identifying whether the effect of the halo is due to the particular color change from red to blue or due to the interactive aspect of the intervention.

Search tasks

In the second experiment, we changed the tasks and reran the experiment on Mechanical Turk. We chose tasks from the Google-A-Day questions from the months of March to May in 2011 and 2012 that seemed more complex. We piloted the tasks and chose the following:

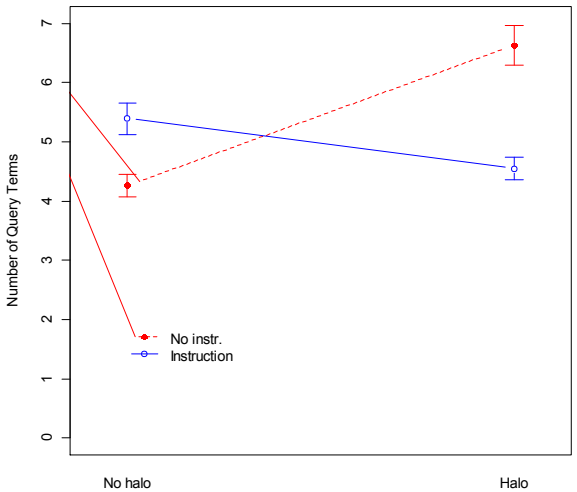

Figure 6. Mean number of queries for the factors Halo and Instruction. Error bars represent ± 1 standard error.

1. I'm a serious piece of music known for comedy. I was played by a famous cat. I've been played by a woodpecker, dueling ducks and two vaudevillian brothers. Who composed me?
2. There may be 39 signatures on the U.S. Constitution, but there were only 38 signers. Which state's absent delegate had his name signed by a colleague?
3. You are standing in the farthest west U.S. town with a population of one person. What is the speed limit?

Experimental design

Our experiment had a one-factor between-subjects design. The factor had four levels: red fading to blue halo (*Halo*), blue fading to red halo (*Inverted*), static blue halo (*Static*), and no halo (*Control*).

Each subject performed three search tasks in random order in a randomly-assigned experimental condition. Subjects were paid $1.03 through Mechanical Turk upon completion of all tasks. We restricted participants to be based in the United States and required them to have a 98% or better HIT completion rate.

Hypothesis

This experiment was designed to test the hypothesis that the change in color was responsible for people's continued attention. Thus we expected the *Halo* and *Inverted* conditions to out-perform the *Static* and *Control* conditions.

Results

Ninety-two people participated in Experiment 2. Three of the participants were discarded because they were able to see the tasks several times before performing them, and five others were excluded because they spent less than a minute finding the answer and they provided random answers. Data from the remaining 84 participants were used in the analysis. Because we were interested in queries the searchers' created, we removed 83 queries that were close to or exact copies of task questions; analysis of the remaining 1077 queries follows.

The average query length for this experiment was considerably longer (6.0 words/query) than that of experiment 1 (5.1 words/query).

To test the hypothesis that the interactive halo would cause people to type longer queries, we performed a one-way ANOVA with the word length of the queries as the dependent variable and our experimental conditions as independent variables. We found a statistically significant main effect (F(3,1073) = 3.876, p<0.05). A Tukey HSD *post hoc*-test confirmed that the Halo condition outperformed the others with an average of 0.77 more words than the control and 0.89 more words than the Static Halo condition (p<0.05). There was no significant difference between the Halo and the Inverted Halo. We also did not find a significant difference between the Inverted Halo and the rest of the conditions.

Table 2. Performance by condition in Experiment 2

Condition	N	No. of Queries		Query length	
		Mean	SD	Mean	SD
Total	84	12.82	6.16	6.01	3.406
Control	20	13.2	7.06809	5.753	3.0189
Constant Halo	23	11.869	5.2336	5.637	3.196
Halo	22	12.909	6.553	6.528	3.9852
Halo Inverted	19	13.473	6.01265	6.105	3.240

DISCUSSION

What does all this mean? Our results suggest that the *Red fading to Blue* halo was effective at eliciting longer queries as we had hypothesized. Unfortunately, the story is not as clear-cut as expected. We saw an interesting interaction with instructions in Experiment 1, and did not find a statistically-significant increase in query length due to the *Blue fading to Red* halo, although that interface condition showed a trend in the right direction. We attribute some of the difference to cultural norms; the message of the Blue fading to Red halo was by design more ambiguous. Some people may be more sensitive to the ambiguous message while others only react to the dynamic behavior of the halo.

The magnitude of the difference between halo and control in the first experiment and the *Red fading to Blue* halo and control in the second experiment is curious. One possible explanation is that the search puzzles in the latter experiments were more difficult than in the first experiment, as suggested by the increase in the average query length for tasks in the second experiment over that in the first (6.0 vs. 5.1 words). In addition, we compared the task correctness of answers between the two experiments and found that in the second study searchers had a significantly lower correctness score (0.39 vs. 0.69; $t(143)=6.8078$, $p<.001$) than searchers in the first study.

Variability in performance of search systems and users due to search topic is well-described in the literature [10, 11]. We would need to run this experiment over many more tasks to understand how robust these effects are. One challenge that needs to be addressed is obtaining a large-enough pool of participants. One way to address this issue is to pre-test a larger number of topics from which a subset with desirable properties (e.g., consistency, diversity, etc.) would be selected for specific experiments.

While we used query length as a proxy for query quality, the halo can represent other metrics, such as diversity of results, novelty, etc. These other metrics may represent more directly desirable outcomes from the perspective of a multi-query search session.

CONCLUSION

This research was motivated by the desire to shape searchers' behavior toward more constructive outcomes using techniques from persuasive computing. As an initial exploration of this space, we created a novel interaction technique to encourage people to create longer keyword queries, and evaluated it with a Mechanical Turk experiment. The encouraging results of our evaluation suggest that this is a promising area for further exploration.

ACKNOWLEDGMENTS

We thank Tony Dunnigan for his suggestions about the design of the halo, and Thea Turner for help with some of the mechanics of data analysis and for her comments on the draft.

REFERENCES

1. Aula, A., Rehan M. K., and Zhiwei G. (2010) How does search behavior change as search becomes more difficult? In *Proc. CHI '10*, pp. 35-44, ACM Press.
2. Bailey, P., White, R.W., Liu, H., and Kumaran, G. (2010) Mining historic query trails to label long and rare search engine queries. *ACM Trans. Web*, 4:15:1–15:27, 2010
3. Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., Yuan, X.-J. (2002) Query Length in Interactive Information Retrieval. In *Proc. SIGIR 2003* (Toronto, Ont, Canada). ACM Press.
4. Belkin, N.J., Cool, C., Jeng, J., Keller, A., Kelly, D. Kim, J., Lee, H.-J., Tang, M.-C., Yuan, X.-J. (2002) Rutgers' TREC 2001 Interactive Track Experience. In Proc. *TREC 2001*, pp.465-472. Washington, DC: GPO.
5. Bing API. Available at https://datamarket.azure.com/dataset/8818F55E-2FE5-4CE3-A617-0B8BA8419F65
6. Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002) Predicting query performance. In Proc. *SIGIR '02*, pp. 299–306, New York, NY, USA. ACM Press.
7. Fogg, B.J. (1998) Persuasive Computers: Perspectives and Research Directions. In Proc. *CHI 1998* (Los Angeles, CA). ACM Press.
8. He, J., Larson, M., and de Rijke, M. (2008) Using coherence-based measures to predict query difficulty. In *Proc. ECIR 2008*, pp. 689–694. Springer.
9. Kumaran, G. and Carvalho, V. R. (2009) Reducing Long Queries Using Query Quality Predictors. *In Proc. SIGIR 2009*, (Boston, MA), ACM Press.
10. Voorhees, E.M. (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M*, 36:697-716.
11. Voorhees, E.M. and Buckley, C. (2002) The effect of topic set size on retrieval experiment error. In *Proc SIGIR '02*, ACM, New York, NY, USA, 316-323.