Text Selections as Implicit Relevance Feedback

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

Georg Buscher
Microsoft Bing
Bellevue, WA 98004
georgbu@microsoft.com

ABSTRACT

Users' search activity has been used as implicit feedback to model search interests and improve the performance of search systems. In search engines, this behavior usually takes the form of queries and result clicks. However, richer data on how people engage with search results can now be captured at scale, creating new opportunities to enhance search. In this poster we focus on one type of newly-observable behavior: text selection events on search-result captions. We show that we can use text selections as implicit feedback to significantly improve search result relevance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, selection process.*

Keywords

Text selection; implicit relevance feedback.

1. INTRODUCTION

Searchers express their needs to search engines via query statements. These queries can be augmented by information from prior search activity, such as previous search engine result page (SERP) clicks within the session or across multiple sessions. Indeed, previous research has shown that prior clicks mined from session histories can be effective in modeling search interests [6]. However, search interaction with SERPs and landing pages is richer than simply queries and clicks [4]. A variety of methods have been proposed for gathering implicit feedback based on document examination and retention behaviors [5], but generally such information is not available to search engines at scale. More recently, methods have been proposed to capture richer interactions with SERPs such as cursor movements, scrolls, and text selections [2]. Monitoring these events on the SERPs of search engines affords a range of possibilities to improve retrieval effectiveness by modeling user interests. In this poster, we study the value of feedback gleaned implicitly from one of these behaviors, SERP text selections, where users mark blocks of text for further manipulation.

2. TEXT SELECTION AS FEEDBACK

The source of our implicit feedback is text selection events on SERPs, specifically blocks of text selected from result captions. We refer to the text within which the selection occurs as the *container*. Figure 1 shows an example of a selection event in a result caption, where the user has selected text in the caption snippet.

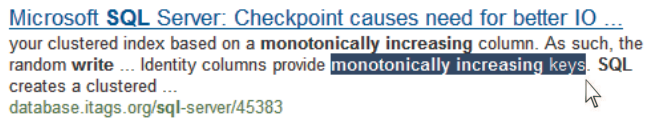

Figure 1: Text selection (blue background) on SERP snippet.

Searchers may select text for a number of reasons, including to copy-and-paste to another application, as a query to a search engine (e.g., via direct functionality available in some browsers or the browser search box), or as a reading aid. Using browser-specific JavaScript functionality similar to that in [2], we can identify when text selections occur and can also determine the bounding box of the immediately surrounding HTML element (e.g., the result snippet).

We believe that text selection events provide valuable information about searcher interests that could be used to improve search performance on future queries for that user. Specifically, we were interested in determining whether there was value from using text selection events as implicit feedback and using the text associated with a text-selection event (the title, the snippet, or the title plus snippet) to model search interests. For example, in Figure 1, although only some words in the snippet are selected, the full snippet (from "your clustered index" to "creates a clustered") could be used to construct a model of the user's search interests.

We describe a log-based study of one application of such interest models: re-ranking the search results for near-future queries based on their similarity to captions that contained text selections.

3. STUDY

3.1 Data

To perform the analysis required for our study, we used 928 users' queries and SERP interactions over a four-week period in late 2011. Data were gathered as part of a related study of SERP interaction behavior. Participants were employees of Microsoft Corporation and volunteered to participate in the study. Participation involved installing a plugin for the Internet Explorer browser that recorded the Web pages visited. When users were on the SERP of the Microsoft Bing search engine, the plugin injected JavaScript into the page, allowing for it to record cursor movements, scrolling, and text selections using a method similar to that used in [2]. Scalability considerations meant that the actual content or precise position of the selected text were not recorded using this method. However, the full text of every caption on a SERP (title, snippet, URL) could be obtained by mining the Bing search logs, and those captions containing a text selection could be identified.

Participants performed 9,433 search sessions comprising a total of 39,606 queries over the course of the study. 754 (1.9%) of the queries logged had text selections, including in the search box (reformulations) and on inline answers. Since we were focused on re-ranking the organic search results, we restricted the set of queries to the 389 queries *A* (1.0%) that had at least one text selection on the snippet of an organic result and had a follow-on query *B* with at least one term in common with the current query *A*, to help ensure consistency in user information needs. We will be re-ranking the results for query *B* directly after the query *A* which contained a text selection, since that gives the search engine an opportunity to include the implicit feedback in the result ranking.

Although 389 queries may seem like a small set by search log analysis standards, there is still sufficient volume to study the value of using text selections in our study. Also, in terms of application, any significant improvement in search performance for 1% of queries to a search engine is still potentially impactful given the amount of investment engines make in improving their performance and the large the number of queries that they process.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

3.2 Relevance Judgments and Measures

For evaluation, we required a relevance judgment for each result. Obtaining many relevance judgments directly from end users is impractical and there is no known approach to train expert judges to provide reliable judgments that accurately reflect individual user preferences. Hence we obtained these judgments using a log-based methodology similar to that adopted by [1]. Specifically, we assign a positive judgment to each of the top 10 results for which there was a satisfied (SAT) SERP click. We defined a satisfied click in a similar way to previous work [3], as either a click followed by no further clicks for 30 seconds or more, or the last result click in the session. URLs without a click received a negative judgment. Positive judgments are used to evaluate retrieval performance before and after re-ranking using average precision (AP) across the top 10 results and reporting mean average precision (MAP) over all queries in our set. Queries without a positive judgment for any result in the top 10 are excluded from the evaluation since we needed at least one click to evaluate our method.

3.3 Systems

We compared the performance of three systems:

Original ranking (baseline): The search engine result ranking.

QuerySimilarity: The original search engine results are re-ranked by: (1) Selecting snippets from the directly preceding query *A* with at least one term in common with the current query *B*. Note that stopwords such as "and" or "the" are removed prior to this comparison. (2) Building a term-based model of the user's search interests comprising the non-stopword terms in the snippets and the frequency with which they appear in the snippets. (3) Re-ranking the top-10 results for the current query *B* by the cosine similarity between the interest model and each result snippet. Note that this system serves as an alternative baseline in our study.

SelectionSimilarity: The original search engine results are re-ranked using the same cosine-similarity function as in *QuerySimilarity*. However, rather than building the interest models based on all snippets from the previous query *A* with at least one shared term, we build models based on the non-stopword terms appearing in snippets with a text selection event from the previous query.

3.4 Method

To ensure that we were comparing the performance of the systems on the same queries, we focused on the 252 queries from 92 users where an interest model could be generated for both *QuerySimilarity* and *SelectionSimilarity* and the next query *B* had a SERP click (for evaluation). For each query in the set, we built the interest models for each method using the preceding captions and text selections as appropriate, and re-ranked the top 10 search results (as described in the previous section). Once we had the new ranking, we computed AP for that list using the click-based judgments and averaged across all queries to get the MAP value on which we compare system performance. We also retained the individual AP values for each query and used them for statistical significance testing, performed using parametric tests at α=0.05.

4. FINDINGS

4.1 Overall Performance

The performance statistics for the original Bing ranker are proprietary and cannot be shared. However, this is a strong baseline and we can present the percentage change in MAP obtained by re-ranking using our two similarity methods with respect to the original ranking. Figure 2a shows the percentage change in MAP for each method. Error bars denote the standard error of the mean (SEM). The findings suggest that re-ranking based on *Selection-Similarity* leads to strong gains of around 6% in MAP over the baseline, while those based on *QuerySimilarity* perform much worse than the original ranking. A one-way analysis of variance (ANOVA) revealed that there were significant differences between the three systems ($F(2,753)=5.85$, $p=.003$), and Tukey post-hoc tests showed all differences were significant at $p<.02$. *QuerySimilarity* may underperform because of the added noise from snippets included to increase result diversity, but which in our case hurt retrieval effectiveness.

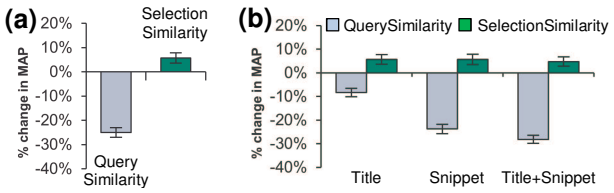

Figure 2: Change in MAP for methods vs. baseline (a) and impact of container on MAP change (b) (± SEM).

4.2 Effect of Container Type

The results presented in the previous section are for snippet only, since that is the container where the user selected the text of interest. However, the title of the result may also be valuable. We varied the container type, using the title, snippet, or title plus snippet to represent the matching results in query *A* and all results in query *B*. Figure 2b illustrates the performance of each method using different container types. We can see that *SelectionSimilarity* is largely unaffected by the container type, but *QuerySimilarity* performs much better when title-only is used (perhaps due to reduced noise). A two-way ANOVA, with ranking method and container type as factors, revealed significant differences between the container types for *QuerySimilarity* ($F(1,1506)=9.58$, $p=.001$), with title-only outperforming the others (Tukey post-hoc tests, $p<.01$).

5. CONCLUSIONS AND FUTURE WORK

We presented a novel method for using text selections on SERPs as implicit feedback. Our results are promising and show that we can significantly improve search performance for queries where text selection events are available on immediately-preceding and related SERPs. Although we saw relevance gains from our method, the use of cosine similarity for re-ranking is somewhat simplistic. Features of text highlighting such as the length of the selection and the number of blocks of text selected could be learned to improve our models. Also, recording the selected text, as well as what users did with it post-selection could yield more accurate feedback and larger gains. User studies may also help better understand SERP highlighting and how to use this behavior.

REFERENCES

[1] Bennett, P. et al. (2011). Inferring and using location metadata to personalize web search. *Proc. SIGIR*, 135–144.

[2] Buscher, G. et al. (2011). Large-scale analysis of individual and task differences in search result page examination strategies. *Proc. WSDM*, 373–382.

[3] Fox, S. et al. (2005). Evaluating implicit measures to improve the search experience. *ACM TOIS*, 23(2): 147–168.

[4] Joachims, T. et al. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM TOIS*, 25(2).

[5] Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference. *SIGIR Forum*, 37(2): 18–28.

[6] White, R.W., Bennett, P.N., and Dumais, S.T. (2010). Predicting short-term interests using activity-based search context. *Proc. CIKM*, 1009–1018.