Evaluating the Impact of Snippet Highlighting in Search

Tereza Iofciu
L3S Research Center
iofciu@L3S.de

Nick Craswell and Milad Shokouhi
Microsoft Bing Search
{nickcr,milads}@microsoft.com

ABSTRACT

When viewing a list of search results, users see a snippet of text from each document. For an ambiguous query, the list may contain some documents that match the user's interpretation of the query and some that correspond to a completely different interpretation. We hypothesize that selectively highlighting important words in snippets may help users scan the list for relevant documents. This paper presents a lab experiment where we show the user a top-10 list, instructing them to look for a particular interpretation of an ambiguous query, and track the speed and accuracy of their clicks. We find that under certain conditions, the additional highlighting improves the time to click without decreasing the user's ability to identify relevant documents.

1. INTRODUCTION

When users view a list of search results they see 'snippets' of text from the retrieved documents. A snippet helps the user decide whether to click, view and potentially make use of a document. A good snippet gives an indication of whether a document seems relevant, deserving click.

This paper evaluates *lists of* snippets, in the context of ambiguous queries. For ambiguous queries, a user may be faced with some results that are completely off-topic. For example, when users type the query 'house', they may be looking for information on the US House of Representatives, the TV series House or real estate. When users type 'microsoft' they may be looking for investment information, products to buy or technical support. There are multiple interpretations of the query, and it is unlikely that a user wants all of them. Therefore snippets should allow users to quickly reject results that are completely off topic, and scan towards those that are valuable. Therefore our experiments involve scanning a results lists of ambiguous queries.

In particular we consider two types of highlighting for the words in snippets. Our baseline approach is similar to the typical interfaces of the current web search engines, where the user's query keywords are highlighted in bold. Our other method highlights additional words (in yellow), that are not query words but are important for that particular document. The baseline method always highlights the same words in each snippet, while the new approach highlights the *differences* between snippets.

For example, for the query "Cornwall England", where the query intent is not very clear, a search engine retrieves general information pages, like Wikipedia pages, but also pages with tourist information. The baseline highlighting puts only the words 'cornwall' and 'england' in bold. Our new method, in addition, highlights 'tourist', 'Wikipedia' and 'pictures'. This potentially allows, for example, a user who is ready to book their holiday to find travel booking sites more easily. In one experiment the additional highlighting is automatic, in the other it is manual. In both cases the hypothesis is that users will be able to scan towards relevant documents more quickly with the additional highlighting.

2. RELATED WORK

There are many studies in literature focusing on different aspects of document representation and summarization in the context of information retrieval. Some approaches are evaluated in a task-oriented manner where speed and accuracy are compared for different search result representations. A recent example of 'extrinsic' evaluation, with references to past studies, is [1].

Alternatively snippet evaluation can be intrinsic: For example measuring whether the summary contains important n-grams from the document. These measures, such as DUC's ROUGE[1], are correlated with extrinsic measures, and have the advantage of being reusable. The present study is non-standard, so we can not repeat any existing intrinsic or extrinsic method. Ours is an extrinsic evaluation concerned with *lists* of summaries.

Our study is similar to the one presented in [5] and later in [4], where the importance of query biased summaries for web search result representation was demonstrated. A task-oriented evaluation was conducted, similar to [1], where the participants had to fulfill different types of search tasks. In the task-oriented studies the users were free to build their own queries in order to solve the tasks. Similar to our experimental setup, in [3] the queries, TREC topics in this

[1] `http://berouge.com/`

case, and their search results, have been fixed thoughout the experiment.

3. USER STUDY SETUP

This paper describes two rounds of experiments. The main difference between the two is the highlighting method (manual vs automatic) and the method for selecting ambiguous queries. However, we made a number of general improvements in our second experiment.

In both experiments our experimental subjects followed a similar procedure. The user is shown an ambiguous query, along with a 'topic description' of how the query should be interpreted. For example, the query 'house' and the description 'information on the TV show'. Then, the user clicks a link to indicate that they are ready, and we show the top-10 list for the query (taken from the Microsoft Web search engine). The user's task is to identify and click a document that fits the topic description, and then the move on to the next query-topic description. The top-10 results and snippets are always the same for each query, and query words are always highlighted in bold. We only vary whether there is additional highlighting, in yellow, of non-query words.

3.1 Manual Experiment Setup

Our pilot experiment used manual highlighting rather than any realistic method for automatically highlighting extra words in snippets. We describe the manual experiment, although the 'automatic highlighting' experiment improves on it in a number of dimensions.

Selecting the queries. If a query has most of its clicks on a single URL, it is probably not an ambiguous query. It is more likely to be navigational [2]. To select ambiguous queries we first select queries with skewness smaller than 0.5, from the 'torso' of the query distribution (not a head query, not a tail query). We manually inspected the top-10 list for 100 of these queries, to identify 50 that seemed to have results that cover more than one topic, and used these as our manual experiment queryset.

Query intent. For each of the selected 50 queries, we developed a topic description. The topic was selected to describe some aspect of the query's top-10 results. We also judged the relevance of each result to the topic, and made a second pass where topics and judgments were checked by a second assessor.

Highlighting. Three assessors each viewed the top-10 result snippets and selected 'important' words for highlighting. The result snippets were shown in the order they were retrieved by the search engine. They did so without knowing the query's topic description, to avoid any bias towards that interpretation. In our experiment, we then highlighted any word or phrase that was selected by two or more assessors.

3.2 Automatic Experiment Setup

After the manual experiment, we noticed that some queries were not really ambiguous (for example 'comet 17p holmes'). This is a problem because it led to the development of a contrived topic, which was confusing to our users and unlikely to agree with our highlighting. In our second experiment, we improved our method for selecting ambiguous queries and introduced an automatic highlighting method.

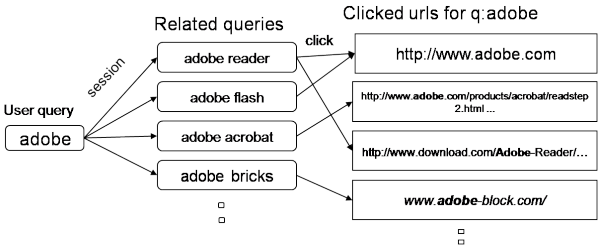

Figure 1: Ambiguous query and intent selection.

Selecting the queries and query intents. To help us identify ambiguous queries, we developed a distinctiveness measure for search results based on information from search logs. Session information connects query q and query q' if query q tends to be followed by q' within user sessions. Click information connects query q and URL u if we have observed users clicking on search result u for query q.

To calculate our distinctiveness score for a query, such as 'adobe' in Figure 1, we assign queries to the top-10 URLs. The assignment is according to click data, however we only include queries that are also connected to the original query in session data. The query 'adobe bricks' has a click connection with one URL, and a session connection with 'adobe', so it is associated with the URL.

The distinctiveness of a URL is the proportion of its associated queries that were not assigned to any other URL. The output of our process is a set of query-URL pairs with distinctiveness of 0.5 or greater.

For the automatic experiment, 40 pairs of query and distinct URL were manually selected from 700 candidates. The query's 'topic description' was 5 of the associated click/session queries, preferring queries with greater numbers of clicks.

Highlighting. We used three approaches for automatic highlighting:

- Top query phrase. Using click data only (not session data) we highlighted the most popular click query that occurred in the snippet, if any.
- Top URL anchor phrase. If no query phrase was highlighted, we highlighted the most popular incoming anchor phrase that occurred in the snippet. Anchor information came from a large Web search engine.
- Wikipedia disambiguation terms. Where a Wikipedia disambiguation page existed for a given query, such as "Cornwall (disambiguation)"[2], then all the disambiguating entity names were highlighted in the query result page.

[2]`http://en.wikipedia.org/wiki/Cornwall_(disambiguation)`

Figure 2: Automatic highlighting for the query "Cornwall".

The first two approaches can highlight differently for each result in the top-10, since each URL has different click data and incoming anchor text. The third approach was applied globally to the search results.

Figure 2 shows an example of automatic highlighting. As always, the additional highlighting gives the highlighted word/phrase a yellow background.

4. EXPERIMENT RESULTS

In both experiments, each user saw all queries. Half the users saw additional highlighting on the odd numbered queries. The other users saw it on even numbered queries. At the end of the experiments the participants were asked to answer a questionnaire.

4.1 Manual Experiment

The manual experiment had 16 participants who each processed 50 queries. We manually judged the relevance of each top-10 result with respect to the chosen interpretation (topic). The same top-10 was also used for topic development (i.e. assigning the desired topic to a query), so upon judging the top-10 there were always one or more relevant documents found for the assigned topic. Figure 3 shows that relevant documents were distributed evenly over ranks, but users tended to click documents near the top of the list. This is consistent with our instructions to click the first relevant document found. It also matches the ranks of the 'shallowest relevant document' for each query, i.e. the first relevant document to be found in the top-10 retrieved.

Results indicate that manual highlighting was not useful. Table 1 shows that users were slower when faced with the new highlighting, and users delayed longer in cases where they eventually clicked an irrelevant document. We then divided our observations into two groups, fast and slow, based on the time to click. We show the accuracy of clicks in Table 2. This again indicates that a delay in the manual highlighting case is associated with making more mistakes.

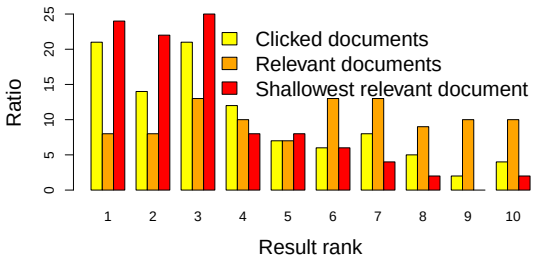

Figure 3: Relevant results vs. clicked results

Table 1: Average time until click

Highlighting	Time (sec) when relevant	Time (sec) when not relevant
baseline	20.83	19.19
manual	23.24	27.38

Table 2: Probability of clicking a relevant result

Highlighting	Relevance (when fast)	Relevance (when slow)
baseline	0.76	0.79
manual	0.78	**0.67**

4.2 Automatic Experiment

The automatic experiment had 8 users who each processed 40 queries. Having identified a number of problems in the manual experiment, we made a number of changes in the automatic experiment. Of course we employed an automatic highlighting method and used a new method for identifying potentially ambiguous queries (see Section 3.2). For each query users now click the topic description itself to indicate that they are ready to see the top-10. This was intended to reduce the chances of a user ignoring a topic. We also precomputed and optimized the HTML of top-10 lists, to make the top-10 lists render on the screen more quickly.

Highlighting had a much smaller effect in the automatic experiment than in the manual experiment. In particular, automatic highlighting did not cause users to become both slow and inaccurate for some queries. For example, adding automatic highlighting did not change the click distribution over ranks (Figure 4). The automatic method highlighted fewer words than the manual method, and may have been more consistent.

In the automatic experiment click accuracy was 0.9, compared to 0.75 for the manual experiment. In the automatic experiment, this level of accuracy was maintained with and without the additional highlighting. A breakdown of accuracy differences per-query is presented in Figure 5.

Within the automatic experiment, the main effect we observed was the time taken to click. The baseline highlighting had a time till click of 13.5 seconds, while the time for automatic highlighting was 11.2 seconds. Figure 6 shows the

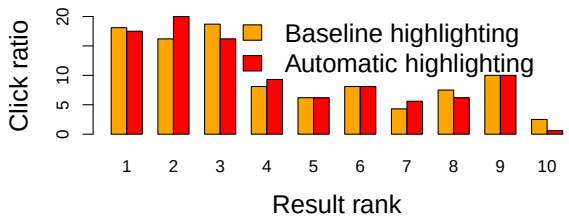

Figure 4: Click histogram highlighting vs. baseline highlighting

difference in average time on a per-query basis.

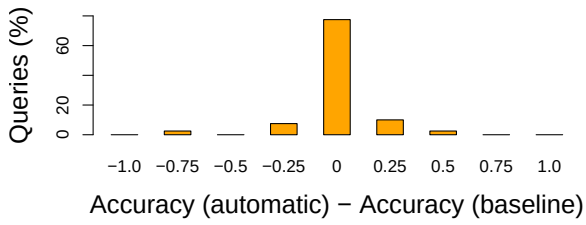

Figure 5: Accuracy of automatic vs. baseline highlighting

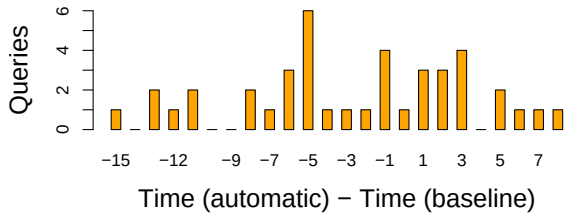

Figure 6: Time taken for automatic vs. baseline highlighting

4.3 Questionnaire Results

At the end of the experiment the participants had to fill in a questionnaire about the search tasks and their experience with the experiment. In the manual experiment users were more likely to say that there was too much yellow highlighting (the additional highlighting was always yellow).

In both setups more than 60% of the participants have reported to having been sometimes familiar with the search topics and more than 70% found the connection between the query and the selected intent often understandable.

5. CONCLUSION AND FUTURE WORK

This paper described our experiments in highlighting the important words in the search snippets for ambiguous queries. Unlike many summarization experiments, we tested how easy it was to scan a top-10 list of snippets, rather than the quality of individual snippets.

Our manual experiment was set up with a lot of human effort: Manual topic development, manual highlighting of the snippet words selected by two out of three assessors, and full relevance judgments of the top-10s. However, we suspect that some topic descriptions were somewhat 'contrived', having been developed for queries that were not really ambiguous. This may have been confusing our users, who also reported in the post-experiment questionnaire that there was too much highlighting. Overall, showing manual highlighting was associated with slower and less accurate clicks.

Our automatic experiment used a log analysis method to identify queries that seem ambiguous, because they have one distinctive URL in the top-10. Although this set of query-URL pairs still required manual vetting, we believe it was a much cleaner set of ambiguous queries. We also introduced an automatic highlighting method based on click logs, anchors and Wikipedia disambiguation pages. Finally we made two changes to the experimental interface, by speeding up the software and increasing the focus on topic descriptions by forcing them to click the description before proceeding. In combination, these changes led to us no longer seeing slow and inaccurate click behavior in the presence of highlighting. Instead, click accuracy was maintained, while speed improved by 17%, to about 11.2 seconds per query.

One drawback of our experiments is that we only used ambiguous queries, and there was always a manual vetting procedure during query selection. Therefore we have not studied the influence of highlighting in general. In future work we would like to understand the influence of query type on our experiments, and improve our automatic techniques for discovering ambiguous queries, since it may be desirable to highlight differently for different query types. We also intend to experiment with eye-tracking tools, to measure more directly the influence of highlighting on user attention.

6. REFERENCES

[1] Hideo Joho and Joemon M. Jose. Effectiveness of additional representations for the search result presentation on the web. *Inf. Process. Manage.*, 44(1):226–241, 2008.

[2] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *WWW '05*, New York, USA, 2005. ACM.

[3] Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR '98*, New York,USA, 1998. ACM.

[4] Ryen White, Joemon M. Jose, and Ian Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Inf. Process. Manage.*, 39(5):707–733, 2003.

[5] Ryen White, Ian Ruthven, and Joemon M. Jose. Web document summarisation: A task-oriented evaluation. In *DEXA '01*, Washington, DC, USA, 2001.