Summary Attributes and Perceived Search Quality

Daniel E. Rose
A9.com Inc.
130 Lytton Ave., Suite 300
Palo Alto, CA 94301
danrose@a9.com

David Orr
Yahoo! Inc.
701 First Ave.
Sunnyvale, CA 94089
dmorr@yahoo-inc.com

Raj Gopal Prasad Kantamneni
Yahoo! Inc.
701 First Ave.
Sunnyvale, CA 94089
kprasad@yahoo-inc.com

ABSTRACT

We conducted a series of experiments in which surveyed web search users answered questions about the quality of search results on the basis of the result summaries. Summaries shown to different groups of users were editorially constructed so that they differed in only one attribute, such as length. Some attributes had no effect on users' quality judgments, while in other cases, changing an attribute had a "halo effect" which caused seemingly unrelated dimensions of result quality to be rated higher by users.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces— *evaluation/methodology.*

General Terms

Measurement, Experimentation, Human Factors

Keywords

Web search, query-biased summaries, user behavior

1. INTRODUCTION

What do users values in web search engine result summaries? To investigate, we conducted a series of experiments in which surveyed users answered questions about summaries that had been deliberately manipulated to exhibit certain characteristics.

We use the term *summary* to refer to the title, abstract, and URL displayed for each web search result. Abstracts in web search summaries are usually *query-biased* [3]; that is, the abstract for a document depends on the query used to retrieve it.

2. METHODOLOGY

We conducted a series of experiments in which online surveys were given to randomly selected Yahoo! Search users. In each survey, approximately 2500 users were randomly assigned to two or more experimental groups. Each user was presented with a hypothetical search situation and a query that might be used in that situation. Users were then shown the summary of a search result for that query and asked to answer questions about it.

Rather than using actual search engine summaries, we presented users with editorially-written summaries that varied only by a particular attribute. Each time we manipulated one attribute, other attributes were kept constant. To construct the abstracts, editors typically began with a web page retrieved for a particular query, then manually identified snippets[1] to be included, modifying them to create the desired test condition. Editorial creation of abstracts enabled us to control which parts of the document were included and insure consistency. Titles were held constant across all conditions, and consisted of the queries themselves. Table 1 shows an example of the variant abstracts created for one dimension of abstract quality, the "choppiness" of the text.

Twelve survey questions were grouped into three areas: Nature of page/readability (e.g. "The information is presented in an uncluttered manner"); relevance ("Would you click on this result") and trust ("How likely are you to trust the information on the resulting web page"). In order to keep the survey short, each participant saw only the questions from one of the areas. Responses to the Relevance questions were Yes or No; the other responses were answered by an agreement scale of 1 to 5. Each participant also answered questions for two scenarios. The result was that on average, each question was answered by about 1000 users for each condition. All scenarios described an informational search goal [2], and all queries were two words long.

3. RESULTS

We ran six experiments, each modifying one of the following summary attributes and holding all others constant:

- *Text Choppiness.* Conditions tested: (1) all snippets complete sentences; (2) incomplete sentences, but with well-chosen breakpoints; (3) incomplete sentences, but with deliberately bad breakpoints.
- *Snippet Truncation.* Conditions tested: (1) complete sentences; (2) beginning of sentence removed; (3) end of sentence removed. In the latter two cases, good breakpoints were chosen.
- *Query Term Presence.* Conditions tested: (1) both query terms present in the first snippet, both in the

[1]A *snippet* is a contiguous piece of text extracted from the document, though some search engines have used this term to refer to the entire abstract.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Table 1: Abstracts varying in "choppiness." Query was `honda accord`.

Complete Sentences	Interested in a full-blown overview of the **Honda Accord** from its birth to the current version? ... Initially available only as a hatchback, the **Accord** rode a 93.7-inch wheelbase and sported a clean, uncluttered body style.
Incomplete Sentences, Good Breakpoints	... a full-blown overview of the **Honda Accord** from its birth to the current ... available only as a hatchback, the **Accord** rode a 93.7-inch wheelbase, weighed about 2,000 pounds and sported a clean, uncluttered ...
Incomplete Sentences, Bad Breakpoints	... blown overview of the **Honda Accord** from its birth to the current version? Then you'll ... to park. Initially available only in two-door hatchback form, the **Accord** rode a 93.7-inch wheelbase, weighed ...

Table 2: Overview of Results.

Attribute	Direct Effect	Indirect Effect	Best Condition
Text Choppiness	readability, clutter	—	complete sentences
Snippet Truncation	—	trust, understand why retrieved, perceived relevance	end removed
Query Term Presence	—	—	n/a
Query Term Density	—	—	n/a
Abstract Length	—	clutter	short
Genre	conveys genre	trust, likely to have information	include genre cues

second snippet, and one in the third; (2) both query terms in the first snippet, one in the second, and none in the third.

- *Query Term Density ("Spamminess").* Conditions tested: (1) for a two-snippet abstract, both query terms shown once in the first snippet only; (2) both terms shown in the first snippet and one in the second; (3) both terms shown repeatedly, for a total of eight occurrences.

- *Abstract Length.* Conditions tested: (1) Approximately four lines of text ("long"), assuming typical font size and window dimensions; (2) three lines of text ("medium"); (3) two lines of text ("short").

- *Genre.* Conditions tested: (1) abstracts contained genre cues (such as "official site"); (2) abstracts did not contain genre cues.

Table 2 shows an overview of the results. Each row refers to an experiment in which only the specified attributed was manipulated. An entry in the "Direct Effect" column indicates that there was a statistically significant difference (at 95% confidence) among the treatment conditions in user responses on the topic(s) that were directly related to the manipulation. For example, the first row shows that increasing the "choppiness" of the text led to different user responses on two survey questions: one asking about readability and one assessing clutter. An entry in the "Indirect Effect" column indicates that there was a difference in user response on an aspect of quality not immediately related to the manipulated attribute. The final column shows which condition was preferred.

We observed two interesting phenomena. First, four of the experiments found no differences in the directly related questions. For example, based on our previous field study, we expected that a very high density of query term hits would be perceived as "spammy," and lead to lower scores on questions related to trust. However, this did not occur.

Second, there were three experiments in which users perceived differences in a dimension not directly related to the attribute being manipulated. For example, in the genre experiment, we expected that the presence of genre cues would cause higher ratings for the statement "I know what kind of page to expect if I were to click on the result," and it did. But this condition also received higher ratings for how likely the user is to trust the result, and whether the user believes that the page will have the information s/he is looking for.

4. CONCLUSIONS

By manipulating particular attributes of search result summaries while asking users to indicate how well the results meet certain quality criteria, we have started to understand what users value in summaries.

We found that neither abstract length nor query term density had a significant effect on perceptions of quality when users read summaries, though we expect these factors play a role when users scan result pages, as eyetracking research suggests [1]. On the other hand, text choppiness and truncation affected user ratings of several factors, not only readability (as one might expect), but also trust in the results and understanding why the page was retrieved.

Perhaps our most interesting finding was that providing genre cues in abstracts caused users to have a more favorable impression not only of how well the abstract conveyed the type of page expected but several other attributes as well. This "genre halo effect" increased how much users trusted the result and how likely they felt it would contain the information they were seeking.

5. REFERENCES

[1] G. Hotchkiss, S. Alston, and G. Edwards. Eye tracking study: An in depth look at interactions with Google using eye tracking methodology. Technical report, Enquiro, EyeTools, and Did-It, Jun 2005.

[2] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04*, pages 13–19, New York, NY, USA, 2004.

[3] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR '98*, pages 2–10, New York, NY, USA, 1998.