Understanding the Interpretability of Search Result Summaries

Siyu Mi
Department of Computer Science, Virginia Tech
siyu6@vt.edu

Jiepu Jiang
Department of Computer Science, Virginia Tech
jiepu@vt.edu

ABSTRACT

We examine the interpretability of search results in current web search engines through a lab user study. Particularly, we evaluate search result summary as an interpretable technique that informs users why the system retrieves a result and to which extent the result is useful. We collected judgments about 1,252 search results from 40 users in 160 sessions. Experimental results indicate that the interpretability of a search result summary is a salient factor influencing users' click decisions. Users are less likely to click on a result link if they do not understand why it was retrieved (low *transparency*) or cannot assess if the result would be useful based on the summary (low *assessability*). Our findings suggest it is crucial to improve the interpretability of search result summaries and develop better techniques to explain retrieved results to search engine users.

KEYWORDS

Interpretability; search result summary; click behavior.

ACM Reference Format:
Siyu Mi and Jiepu Jiang. 2019. Understanding the Interpretability of Search Result Summaries. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331306

1 INTRODUCTION

Machine learning techniques are becoming increasingly powerful today, but they are also more and more sophisticated and difficult to understand for human beings. What we need is not only accurate models but also models that are explainable to us. Understanding how the model works may also help the user make better decisions and further improve the model.

Despite many discussions of explainable AI and machine learning recently [3, 4, 9, 11, 13], few previous work explicitly examined the interpretability of search results. Some latest studies [14, 15] applied explainable techniques such as LIME [12] to interpret search result ranking. However, it remains unclear how helpful these techniques are in terms of helping search engine users. Also, we believe current search engines do have already offered some interpretable functionalities, but no previous work examined them in such a way.

We note that information retrieval, among many research fields with extensive use of machine learning, is in fact one of the earliest to offer interpretations for system decisions and outputs. Particularly, a query-biased search result summary [16] delivers two important information to search engine users:

- *Why the system retrieves a search result* — we use **transparency** to refer to the ability of a summary to interpret this information. Through selecting sentences with high coverage of query terms and highlighting keywords in URLs and snippets, search result summaries inform users keyword matching and term frequency are important criterion for retrieving and ranking search results.
- *To which extent a result would be useful* — we use **assessability** to refer to the ability of a summary to explain search result relevance. We believe assessability is a unique aspect of interpretability offered by search result summary, as many other machine learning applications directly present system outputs to end users.

We report results from a lab user study for evaluating the interpretability of search result summaries in existing web search engines. We recruit participants to work on different search tasks using an experimental search system, where the results and summaries came from the API of a commercial web search engine. We collect their judgments regarding both the interpretability and the (expected) usefulness of search result summaries after each session.

- Participants have high transparency and assessability ratings for current search engine's summaries.
- The summaries' transparency and assessability judgments positively correlate with each other and usefulness ratings.
- Results of a regression analysis suggests that the transparency and assessability of summaries have significant effects on users' click decisions when they browse a SERP.

2 USER STUDY

2.1 Experimental Design

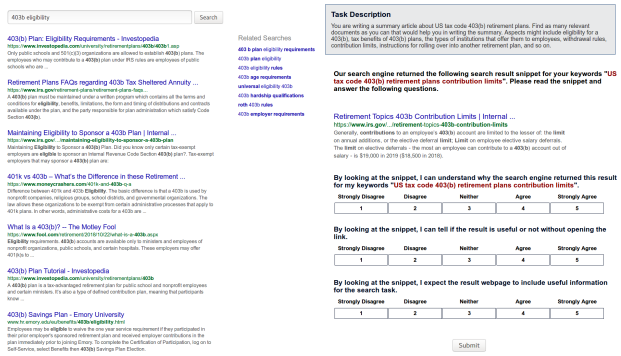
We conducted a lab user study to evaluate the interpretability of search results in web search engines. We instructed participants to work on assigned search tasks in an experimental system and make judgments about the retrieved results afterward. We recorded users' search behavior and collected search result judgments.

Our study used a 2×2 within-subject design to balance different types of search tasks. The tasks come from the TREC session tracks [1] and were categorized into four types by the targeted task product and goal based on Li and Belkin's faceted classification framework [8]. The targeted task product is either *factual* (to locate facts) or *intellectual* (to enhance understanding about a topic). The goal of a task is either *specific* (clear and fully developed) or *amorphous* (an ill-defined or unclear goal). We divided participants into groups of four. Participants in the same group worked on the same four tasks (one task for each type) but using a different sequence (rotated using a Latin square). We assigned different tasks to each group to increase task diversity. We also included a training task at the beginning to help users understand the whole procedure.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331306

Figure 1: Screenshots for search and judgment pages.



The experimental system sends user requests to Bing API and returns Bing search results and query suggestions. Figure 1 (left) shows a SERP from our system. We displayed search results in the same look as Google did at the time of our study, including the font size, weight, and color for title, URL, and snippet. Our SERP only showed result abstracts and related searches. We also highlighted words[1] in URLs and snippets as Bing did at the time of the study.

For each task, we asked participants to collect information using our experimental search system to address the problem stated in the task description. We instructed the participants that they could issue different queries and click on multiple result links. The participants could request to finish a session if they believe they had finished the task requirements. Our system would also terminate a session automatically after 10 minutes. On average, the participants spent 262 seconds in a search session.

2.2 Search Result Judgments

We collected two types of judgments after each session:

Summary Judgments. We asked participants to evaluate search result summaries retrieved during the search session. Table 1 shows the summary judgment questions[2]. We included two interpretability questions (*transparency* and *assessability* as defined in Section 1). We also asked participants to assess the expected usefulness of the result based on the summary (*usefulness*). The judgment page customized some of the questions based on the query retrieved the result summary, where $q is the actual query string. Participants responded to the three judgment questions using a five point Likert-scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

Figure 1 (right) shows an example page for collecting summary judgments. We displayed a summary in the same look we showed it on the SERP, except that we disabled the URL link (we hope users to make judgments based on the summary itself). We also showed task description and the search query retrieved the result summary on the page to help users make judgments.

Result judgments (not reported in this paper). After judging a summary, users needed to read the result web page and respond to a few judgment questions regarding the result document. Here we only focus on summary judgments and do not report result judgments.

[1] Bing search API returned highlighted URLs and snippets. The highlighted words may have not appeared in the search query.

[2] During the training session, we instructed participants that they needed to answer the questions based on the whole search result abstract (although the wording of the questions used "snippet").

Table 1: Search result summary judgment questions.

Transparency	By looking at the snippet, I can understand why the search engine returned this result for my keywords "$q".
Assessability	By looking at the snippet, I can tell if the result is useful or not without opening the link.
Usefulness	By looking at the snippet, I expect the result webpage to include useful information for the search task.

Table 2: Summary of collected result judgments.

Priority of Summary Judgments		Judged/Total (%)
1	Clicked results	546/588 (92.9%)
2	Not clicked, possibly viewed	273/578 (47.2%)
3	Not clicked, possibly not viewed	433/3,056 (14.2%)

Priority of Judgments. A user could issue multiple queries and retrieve a large number of results during a session. It is impractical to require participants to judge all the retrieved results due to the time constraints of a lab study. Thus, we generated a priority list of judgments as in Table 2. We gave clicked results the highest priority because they are connected with more search behaviors (e.g., dwell time) than other results. We divided the unclicked results into "possibly viewed" and "possible not viewed" ones and gave the former a higher priority. We consider an unclicked result summary as "possibly viewed" if the result is at a higher rank than at least one clicked results on the same SERP—previous eye-tracking studies showed that users have high chances to have viewed the summaries at a higher rank than a clicked result [7].

Participants judged results in the priority list one after another until they spent 10 minutes on judgments. We shuffled the sequence of results within each priority group. We did not set any time limits for judging a result to ensure participants have enough time.

2.3 Collected Data

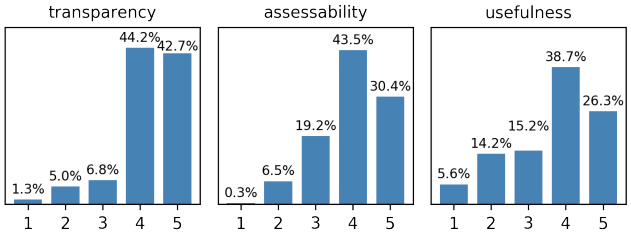
The user study included 40 participants (20 are female) from a university in the United States. We recruited participants through fliers posted to the main buildings of the campus and a Facebook group targeting the whole university. We compensated each participant $20 for their time (about 1.5 hours). We required all participants to be English native speakers and older than 18. The participants included 33 undergraduate students, 6 graduate students, and one staff. About 2/3 of the participants were studying STEM majors. Their average age was 21.95 (SD = 6.07).

In total, we collected user behavior and result judgments from 160 sessions (excluding training sessions). On average, participants issued 2.63 queries, clicked on 3.67 results, and judged 7.83 results (including both summaries and result web pages) per session.

3 ANALYSIS

We examine the collected search result summary judgments in this section. We particularly focus on the following research questions:

Figure 2: Distribution of search result summary ratings.



- RQ1: Are search result summaries from current search engines transparent and assessable to users? (Section 3.1)
- RQ2: How do transparency and assessability relate to each other and usefulness? (Section 3.2)
- RQ3: Does the interpretability of search result summary influence click decisions? (Section 3.3)

When reporting results, we use ns for "not significant at 0.05 level" and $*$, $**$, and $***$ for $p < 0.05$, 0.01, and 0.001, respectively.

3.1 Interpretability and Search Task

Participants rated the search result summaries in our experiments (returned by Bing API) as highly transparent and assessable.

Figure 2 plots the distribution of the collected judgments. 87% of the judged summaries received a transparency rating as high as 4 (44.2%) or 5 (42.7%). 74% of the judged summaries have an assessability rating of 4 (43.5%) or 5 (30.4%). This shows that users considered top-ranked result summaries from current web search engines as highly transparent and assessable. However, it remains unclear if users' perceptions of the two properties are accurate.

3.2 Interpretability, Relevance, and Usefulness

The collected transparency and assessability judgments positively correlate with each other. They also positively correlate with usefulness, but with different strengths.

Table 3 reports the Spearman's correlation of the three judgments among all the assessed search result summaries ($N = 1,252$). We observed a moderate positive correlation between transparency and assessability ($\rho = 0.483, p < 0.001$), suggesting possible connections between the two aspects of interpretability. Both transparency and assessability are also positively correlated with usefulness (but transparency has a relatively stronger correlation), suggesting possible connections between interpretability and search result quality.

Figure 3 and Figure 4 further disclose some details of the correlation between each pair of judgments. We divided the judged summaries into groups based on transparency ratings and assessability ratings ("low" for ≤ 3, "med" for 4, and "high" for 5). Figure 3 and Figure 4 report the mean and standard error of ratings for different group of summaries.

As Figure 4 shows, the differences of the "high" assessability group with the other two ("med" and "low") are much larger than the differences between "low" and "med" groups (although all the differences are statistically significant at 0.001 level). In contrast, we observed consistent clear differences in usefulness ratings between summaries with "high", "med", and "low" transparency ratings in Figure 3. This further explains the stronger correlation between transparency and usefulness comparing to other pairs of judgments.

Table 3: Spearman's correlation of judgments ($N = 1,252$).

	Transparency	Assessability	Usefulness
Transparency	1.0	-	-
Assessability	0.483	1.0	-
Usefulness	0.617	0.476	1.0

All the correlations are statistically significant at 0.001 level.

Figure 3: Comparison of summaries with low ($N = 163$), medium ($N = 554$), and high ($N = 535$) transparency levels.

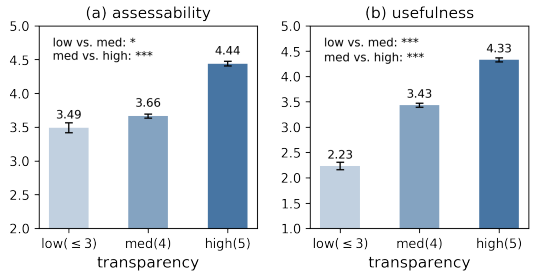
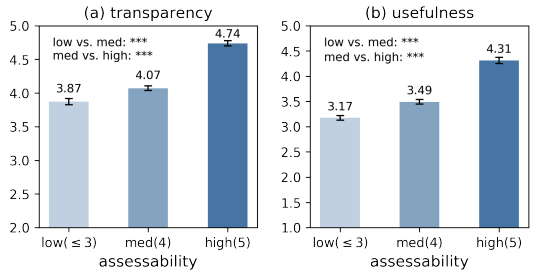


Figure 4: Comparison of summaries with low ($N = 326$), medium ($N = 545$), and high ($N = 381$) assessability levels.



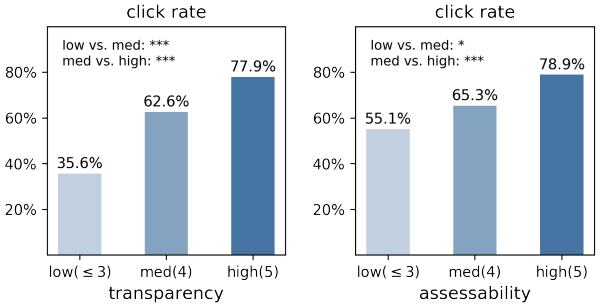
3.3 Interpretability and Click Behavior

We continue to examine the relationship between interpretability and click decisions. We only consider summary judgments for clicked results and the "possibly viewed" results ($N = 819$).

Figure 5 plots the "click rates" for summaries with different levels of transparency and assessability. Here the click rate is calculated as: # judged and clicked summaries / # judged summaries. Figure 5 suggests some possible connections between the two interpretability measures and click decisions. However, it remains unclear whether the different click rates are simply because the two interpretability measures are correlated with usefulness (a widely examined factor for click decision).

We further performed a logistic regression analysis for users' click decisions among the clicked and "possibly viewed" results. We used a multilevel regression model because the data violates the independence assumption for regular logistic regression—we have multiple judgments within a session and a user can perform

Figure 5: Click rates for summaries with different levels of transparency and assessability (only consider clicked results and "possibly viewed" results).



multiple sessions. We model user and session as random effects and examine the list of variables in Table 5 as fix effects. The list of independent variables of interests included:

- Interpretability judgments – transparency and assessability.
- Usefulness [5, 6, 10] – it is widely assumed that users would click on a result link if the summary looks useful. This is also the fundamental basis for using click as implicit feedback.
- The rank of the summary on the original SERP – Joachim et al. [7] hypothesized that users are more likely to click on top-ranked results regardless of relevance/usefulness due to a trust bias.
- The number of query terms matched in the summary – Both Yue et al. [17] and Clarke et al. [2] examined the attractiveness bias of click behavior. Here we use keyword matching as measures for attractiveness. We separately look into the title, URL, and snippet of summaries. We examined results using different text preprocessing methods and found they do not much influence the conclusions. The reported results used the Krovetz stemming and removed stop words (the INQUERY stop word list).

Table 4 reports the results of the regression analysis. Unsurprisingly, we observed a significant positive effect of usefulness on click decisions ($b = 0.547$, $p < 0.001$). Among the three variables for measuring attractiveness of a summary, the number of query terms in URL also shows a significant positive effect on click ($b = 0.186$, $p < 0.05$). We did not observe any significant effect of rank on click decisions, probably because of our selection of results.

In addition to a list of widely examined factors, we have also observed significant positive effects of both transparency ($b = 0.419$, $p < 0.01$) and assessability ($b = 0.388$, $p < 0.01$) on click decisions. This further confirms that transparency and assessability are salient factors influencing search engine users' click decisions. It also helps clarify that the differences of click rates observed in Figure 5 are less likely due to the correlation of transparency and assessability with usefulness.

4 CONCLUSION

The interpretability of artificial intelligence systems has attracted a lot of attention these days. We believe IR system is one of the earliest to provide interpretable techniques to users—search result summary informs users why a result was retrieved and to which extent the result would be useful (after opening the link). We explicitly evaluate this classic, important, yet *neglected* interpretable technique in IR systems through a lab user study. Our findings are illuminating in several different ways:

First, our results suggest that search result summary plays an important role in explaining system's decisions (the retrieval of a particular result) and outputs (the usefulness of a result), as showed from the high transparency and assessability ratings by users.

Second, our results disclose a new salient factor—the interpretability of search result summary—influencing users' click decisions. This suggests search engines can improve the interpretability of search results to optimize users' click decisions. Another important implication is that click models may also need to take into account interpretability to better model click data.

Third, we also recommend that new explainable techniques for search engines and search results should be fully compared with existing query-biased search result summary.

Table 4: Multilevel regression: click as dependent variable.

Independent Variables	Estimate	Std. Error	Sig.
Rank on the SERP	0.046	0.04	
Transparency	0.419	0.16	**
Assessability	0.388	0.12	**
Usefulness	0.547	0.11	***
# query terms in title	0.065	0.07	
# query terms in URL	0.186	0.08	*
# query terms in snipepet	−0.051	0.03	

REFERENCES

[1] B. Carterette, P. Clough, M. Hall, E. Kanoulas, and M. Sanderson. Evaluating retrieval over sessions: The TREC session track 2011-2014. In *SIGIR '16*, pages 685–688, 2016.
[2] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *SIGIR '07*, pages 135–142, 2007.
[3] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2019.
[5] J. Jiang, D. He, and J. Allan. Comparing in situ and multidimensional relevance judgments. In *SIGIR '17*, pages 405–414, 2017.
[6] J. Jiang, D. He, D. Kelly, and J. Allan. Understanding ephemeral state of relevance. In *CHIIR '17*, pages 137–146, 2017.
[7] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05*, pages 154–161, 2005.
[8] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.
[9] Z. C. Lipton. The mythos of model interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning*, pages 96–100, 2016.
[10] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *SIGIR '16*, pages 463–472, 2016.
[11] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
[12] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD '16*, pages 1135–1144, 2016.
[13] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI '18*, pages 1527–1535, 2018.
[14] J. Singh and A. Anand. Posthoc interpretability of learning to rank models using secondary training data. In *2018 SIGIR Workshop on ExplainAble Recommendation and Search*, 2018.
[15] J. Singh and A. Anand. EXS: Explainable search using local model agnostic interpretability. In *WSDM '19*, pages 770–773, 2019.
[16] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR '98*, pages 2–10, 1998.
[17] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *WWW '10*, pages 1011–1018, 2010.