# Search Result Explanations Improve Efficiency and Trust

Jerome Ramos
Brown University
Providence, Rhode Island, USA
jerome_ramos@alumni.brown.edu

Carsten Eickhoff
Brown University
Providence, Rhode Island, USA
carsten@brown.edu

## ABSTRACT

Search engines often provide only limited explanation on why results are ranked in a particular order. This lack of transparency prevents users from understanding results and can potentially give rise to biased or unfair systems. Opaque search engines may also hurt user trust in the presented ranking. This paper presents an investigation of system quality when different degrees of explanation are provided on search engine result pages. Our user study demonstrates that the inclusion of even simplistic explanations leads to better transparency, increased user trust and better search efficiency.

**ACM Reference Format:**
Jerome Ramos and Carsten Eickhoff. 2020. Search Result Explanations Improve Efficiency and Trust. In *SIGIR '20: In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20) July 25–30, 2020, Xi'an, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Search engines have become an essential information source. Due to their ubiquitousness, it becomes increasingly vital that they provide unbiased results. "Unfair" search engines can potentially lead to significant negative social and economic impact. Machine-learned models that use large-scale datasets [8] as a basis for learning may end up mirroring or reinforcing existing bias in the data [7]. For example, social media accounts with non-western user names are more likely to be flagged as fraudulent when spam classifiers are predominantly trained on Western names [3]. It is well known that users prefer documents at earlier result list ranks even though factual relevance may be lower [16].

Model complexity aggravates this situation. As rankers increasingly rely on supervised learning from millions of user interactions via highly parametric non-linear models even technology affine users struggle to discern the exact criteria that led to the final ranking.

Finally, such lacking transparency may translate into decreased search efficiency as users struggle to identify relevant material. This is especially true in exploratory search or situations where the searcher holds low domain expertise [11, 20]. In those scenarios the commonly provided document titles and snippets may not be sufficient to determine if the document is truly relevant, forcing the searcher to manually inspect considerable portions of the document.

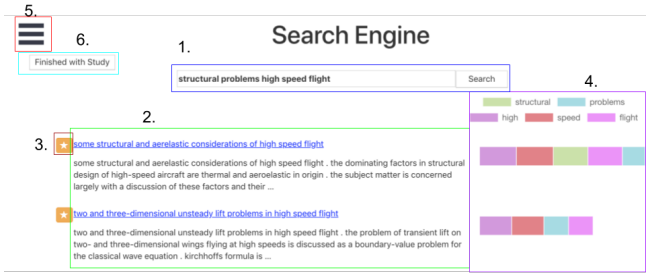


**Figure 1: Transparent search interface. 1) Search bar to issue queries. 2) Search results, ranked from most relevant to least relevant. 3) Star button to bookmark documents. 4) Stacked bar graphs to explain how each query term contributes to the overall score of the document. 5) Hamburger menu to show bookmarked documents and exploration task prompt. 6) Button to redirect users to the quiz after they are done searching.**

In this paper, we make a first step towards measuring the effect that increased search engine explainability has on a) perceived transparency, b) user trust and, c) search efficiency. Our study decomposes the term-specific components of an exact match model and displays them to the searcher, highlighting the overall retrieval model score as well as its breakdown into individual query term contributions.

The remainder of this paper is structured as follows: Section 2 discusses related work on transparent user interfaces and explainable retrieval models. Section 3 introduces our proposed score decomposition approach and search interface. Section 4 introduces the setup of our user study and Section 5 discusses salient findings. Finally, in Section 6, we conclude with an outlook of future directions of research.

## 2 RELATED WORK

Research into search result visualization prominently includes efforts that explore the use of colors and shapes to provide additional term distribution information to searchers. In this family of approaches, Tile-Bars [13] shows users the document length and the density of query terms relative to the documents. HotMaps [14] and HotMaps+WordBars [15] display the term frequency of query terms over all returned documents. These early approaches use Boolean retrieval models rather than score-ranked lists. In addition, their interfaces are significantly different from those found in modern search engines and could potentially have a higher cognitive load for searchers who are exclusively familiar with current commercial search solutions following the 10-blue-links paradigm. Research


*SIGIR '20, July 25–30, 2020, Xi'an, China*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

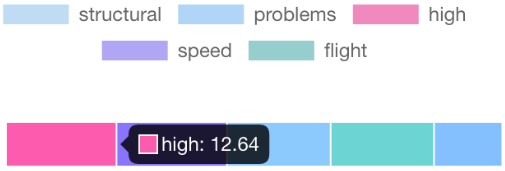


**Figure 2: Hovering over the bar reveals each term's score.**

has broadly explored using alternatives to graphs for visualizing information. However, studies have shown that users prefer bars over numbers for this purpose [19]. URank [5] takes a modern approach to providing visualizations to search engine results. It displays relevance scores for results, noting reduced cognitive load compared to traditional search engines.

Where the focus of the URank studies lay on cognitive load and search strategies, particularly query adjustments [10], this paper is interested in the connection between explainability and user trust/efficiency. As an additional difference to the body of existing work, our study allows participants to issue free-form textual queries rather than letting them select predefined keywords from curated pick lists.

## 3 METHODOLOGY

### 3.1 User Interface

Figure 1 shows our user interface that is designed in minimalist fashion to resemble modern commercial Web search engines. Familiarity and ease of use were important goals in the design process to ensure the measured behavioral traces are consequences of the intervention rather than user confusion. The search bar appears at the top of the screen and allows users to issue their queries. Once a query is submitted, users can see the document title and a 50-word abstract. Each title is a hyperlink that leads to the full article text. Every search result comes with a stacked bar graph that shows the overall relevance score of the document as well as the constituent scores contributed by each query term. Users can hover over one of the bars in the graph (See Figure 2) to see the exact partial score of that term. A legend is provided at the top of the screen to associate a term with its respective color in the graphs.

Finally, every document has a star button next to it that users can click to bookmark documents. They can inspect all bookmarked documents via the hamburger menu in the interface's top left (See Figure 3).

### 3.2 Ranking Score Decomposition

This study relies on the Okapi BM25 retrieval model [18], that ranks matching documents according to their term-wise overlap with a given search query. For every document $D$ and a query $Q$, with terms $q_1, q_2, ..., q_n$, the score of $D$ is given by:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i,D)(k_1+1)}{f(q_i,D)k_1(1-b+b\frac{|D|}{avgdl})} \quad (1)$$

where $f(q_i, D)$ is the frequency of term $i$ in document $D$, $|D|$ is the number of the words in the document, and *avgdl* is the average document length in the collection. $k_1$ is a discounting factor for

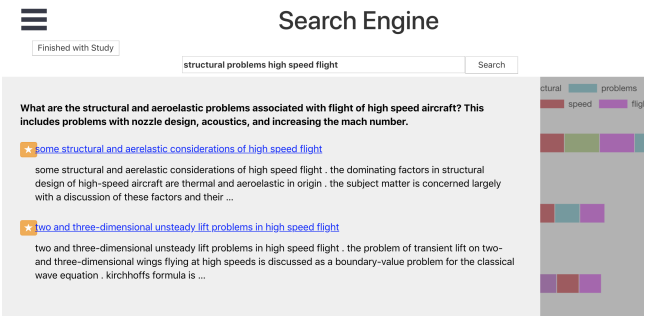


**Figure 3: Viewing the exploration task and the bookmarked documents in the hamburger menu.**

repeated term mentions and $b$ the strength of document length normalization. Additionally, $IDF$ is the inverse document frequency, given by:

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (2)$$

where $N$ is the total number of documents in the collection and $n(q_i)$ is the number of documents containing $q_i$.
The BM25 model was chosen due to its simplicity that allows for easy isolation of query term specific score contributions in the form of the summands of Equation 1, while still yielding competitive ranking performance. All interfaces and experiments were implemented using Javascript, Python, Java, Apache Lucene, and MongoDB[1].

## 4 EXPERIMENTS

Our experiments are conducted on Amazon Mechanical Turk. Users were faced with one of three distinct experimental conditions: a) `T`: A search interface that only shows the titles on the result list (Figure 4); b) `TA`: A search interface that shows the titles and abstracts of the document (Figure 5); c) `TAB`: The full search interface including term contribution bars (Figure 1). All three experimental conditions are served by the same retrieval model and merely show less rich information. Each condition is presented to a total of 50 unique users who are asked to perform open-ended exploratory search tasks. To avoid learning and order effects no searchers are ever exposed to more than one condition or search task.

To ensure a uniform low prior domain expertise of users with the task topic, we choose the Cranfield Collection [2], of scientific articles on aerospace engineering as our experimental corpus. A relatively niche topic was manually chosen to assess the users's ability to quickly gain knowledge on an unfamiliar topic using the three search interface conditions. Users were tasked with an exploratory search where they attempted to answer the question "*What are the structural and aeroelastic problems associated with flight of high speed aircraft?*". The users used their assigned search interfaces to issue queries and learn more about the topic. Once they were satisfied with their research, they were redirected to a multiple choice quiz, where they answered a number of questions related to the prompt.

[1] Our code base is available at https://github.com/bcbi-edu/p_eickhoff_transparent-ir.

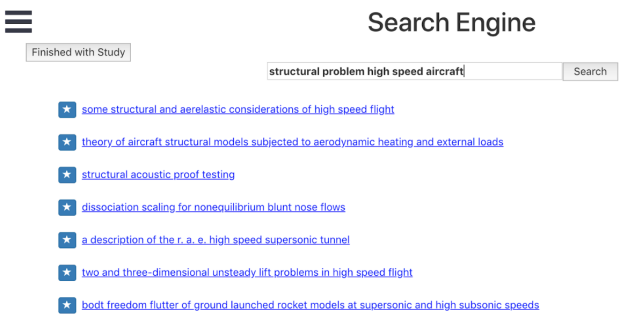


**Figure 4: Search Interface showing only titles.**

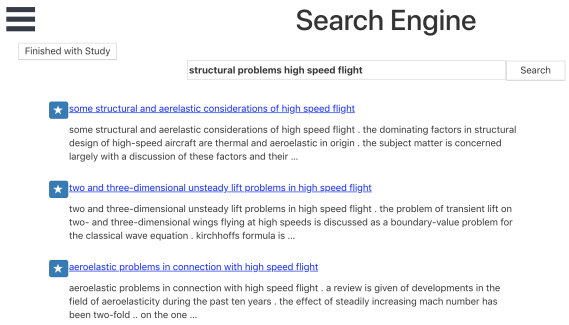


**Figure 5: Search Interface showing titles and abstracts.**

Due to the often noisy nature of labels collected via crowdsourcing [1, 4, 9], this quiz is designed to filter out search sessions that did not make a serious attempt at the experiment. If a user failed the quiz, they were redirected back to the search engine to perform additional research on the topic and are given up to three more attempts to pass the quiz. In this way, users are given another chance to perform more research on the prompt and, in particular, look up information about the questions that they did not know the answer to previously. Once the user passes they quiz, they are given an exit survey to describe their experiences with the search interface. If the user fails the quiz four times, their data is not included in the final results. Similarly, sessions that do not issue any queries, view, or bookmark any documents, were filtered out of the final data.

For each included session, we record issued queries, viewed and bookmarked results as well as the time taken to complete the task. The experiential exit survey asks eight questions about the simplicity, transparency, trustworthiness, and explainability of the search interface. Responses are collected on a five-point Likert scale that range from "strongly disagree" to "strongly agree". The NASA TLX Task Load Index [12] is used to measure the workload of the task. Users were also given the option to provide free-form feedback to comment on their experience with the search interface.

# 5 RESULTS

In this section, we compare the three experimental conditions in terms of perceived system quality, task load, search success and, finally, task efficiency.

## 5.1 Perceived System Quality

Figure 6 plots the answer distribution over the eight exit interview questions for all three search interfaces. For all collected questions, the condition including term component contribution bars scored considerably higher than the other conditions. For all questions in the survey, TAB scored a higher mean (3.785) and median (4.0) when compared to TA (3.0925 mean and 3.0 median) and T (3.2475 mean and 3.0 median). Table 1 shows the results of a Kruskal-Wallis statistical significance test [17] with Bonferroni correction [6] between answer distributions. Despite the comparably small study cohort of only 50 participants per condition, seven out of the eight comparisons between conditions TAB and TA were found to be significant. Only four out of the eight comparisons between TAB and T were found to be significant. Interestingly, both in terms of absolute scores as well as significance tests, condition T, in which only titles are shown, seems to find mildly higher user acceptance than the more informative TA variant.

## 5.2 Task Load

When analyzing the responses to the NASA TLX instrument, users reported significantly less frustration when using TAB compared to TA and T (7.42 vs. 10.42, and 9.30, respectively). Users of TAB also reported mildly lower levels of mental and temporal demand. All three conditions yielded similar levels of physical demand, performance, and effort.

## 5.3 Search Success

To gauge the three experimental conditions' effectiveness, we report the number of turns (searching followed by taking the quiz) the average user needed to eventually pass as well as the average number of correctly answered questions on the quiz. The distribution of scores is tightly matched between all three conditions and no significant differences can be noted.

## 5.4 Task Efficiency

Answers to question F on our survey suggest that the transparent interface TAB was perceived as significantly more efficient than the conventional variants. When inspecting the users' actual time on task, we see this perception justified. TAB sessions were significantly shorter (67.6s) than TA (70.7s) and T (85.9s) sessions. Given that there was no difference in search success between conditions this is a highly encouraging finding.

# 6 CONCLUSION

This paper describes early results in an ongoing effort to measure the effect of retrieval model explainability on system quality. In a crowdsourced user study we showed that search interfaces including even simplistic ranking explanations are perceived as significantly more intuitive to use while simultaneously instilling greater trust in their results. While explanations did not increase the overall likelihood of search success, they led to significant efficiency gains, letting searchers find relevant material faster. These findings are encouraging for future inquiry in this direction. In particular, we aim to study the effects of explainability more broadly as domain and task complexity, result list quality and, system complexity are varied.

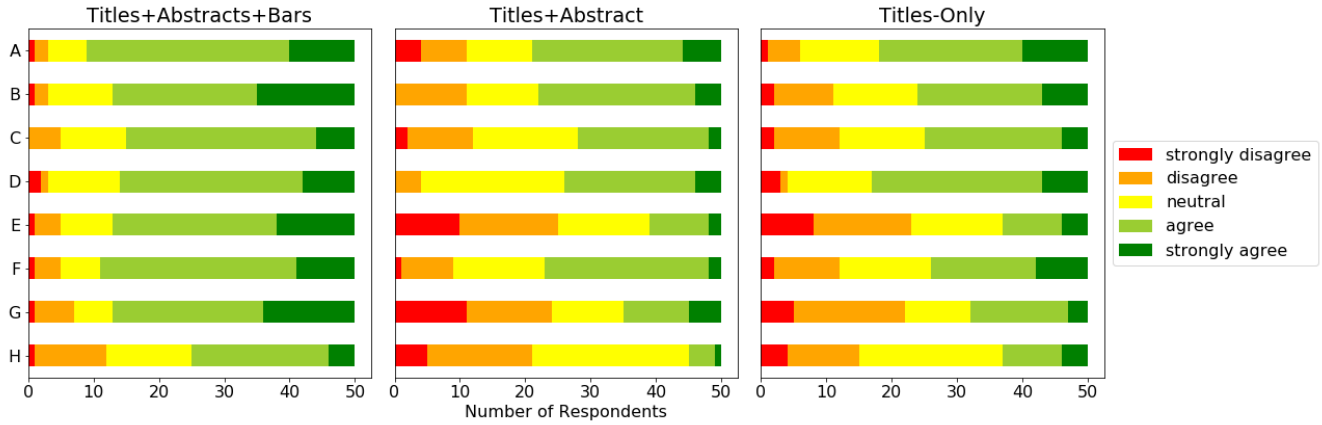


**Figure 6: Exit interview agreement with the statements: A) The search engine was simple to use B) Transparency in search results made it easier to find relevant documents C) Finding relevant results to searches was intuitive D) The results from the search engine are an accurate representation of the truth E) The application clearly explains the rankings of the results F) Transparency in search results helped to find relevant documents quickly G) It is clear why a certain result is ranked higher than another result H) The rankings of the search results corresponds to how trustworthy they were.**

**Table 1: Statistical significance of answer differences. Significant $p$ values ($p < 0.05$) are highlighted in bold.**

| Question | TAB vs. TA | TAB vs. T | TA vs. T |
|---|---|---|---|
| A) The search engine was simple to use | **0.0304** | 0.5359 | 0.679 |
| B) Transparency in search results made it easier to find relevant documents | **0.0116** | **0.0206** | 1.0 |
| C) Finding relevant results to searches was intuitive | **0.0139** | 0.0973 | 1.0 |
| D) The results from the search engine are an accurate representation of the truth | 0.063 | 1.0 | 0.3096 |
| E) The application clearly explains the rankings of the results | **0.00000022** | **0.0000057** | 1.0 |
| F) Transparency in search results helped to find relevant documents quickly | **0.0139** | **0.05** | 1.0 |
| G) It is clear why a certain result is ranked higher than another results | **0.00003** | **0.000087** | 1.0 |
| H) The rankings of the search results corresponds to how trustworthy they were | **0.0007** | 0.1906 | 0.1985 |

## REFERENCES


[1] Piyush Bansal, Carsten Eickhoff, and Thomas Hofmann. 2016. Active content-based crowdsourcing task selection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 529–538.

[2] Cyril W Cleverdon. 1960. The aslib cranfield research project on the comparative efficiency of indexing systems. In *Aslib Proceedings*. MCB UP Ltd.

[3] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*. ACM New York, NY, USA.

[4] Martin Davtyan, Carsten Eickhoff, and Thomas Hofmann. 2015. Exploiting document content for efficient aggregation of crowdsourcing votes. In *Proceedings of the 24th ACM CIKM Conference*.

[5] Cecilia di Sciascio, Vedran Sabol, and Eduardo Veas. 2015. uRank: Visual analytics approach for search result exploration. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 217–218.

[6] Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association* 56, 293 (1961), 52–64.

[7] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 162–170.

[8] Carsten Eickhoff, Kevyn Collins-Thompson, Paul Bennett, and Susan Dumais. 2013. Designing human-readable user profiles for search evaluation. In *European Conference on Information Retrieval*. Springer, 701–705.

[9] Carsten Eickhoff and Arjen P de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval* 16, 2 (2013), 121–137.

[10] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An eye-tracking study of query reformulation. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 13–22.

[11] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 223–232.

[12] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[13] Marti A Hearst. 1995. TileBars: visualization of term distribution information in full text information access. In *SIGCHI*.

[14] Orland Hoeber, Daniel Schroeder, and Michael Brooks. 2009. Real-world user evaluations of a visual and interactive Web search interface. In *2009 13th International Conference Information Visualisation*. IEEE, 119–126.

[15] Orland Hoeber and Xue Dong Yang. 2006. Interactive Web information retrieval using WordBars. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. IEEE, 875–882.

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[17] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[18] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* (1995).

[19] Colin Ware. 2020. *Information visualization: perception for design*. Morgan Kaufmann.

[20] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*. 132–141.