Alleviating Information Overload of Document Retrieval for TU Delft Digital Collections

Information Retrieval Group 20

JUSTIN DE HAAN, Delft University of Technology, the Netherlands
SÉRÉNIC MONTÉ, Delft University of Technology, the Netherlands
KEVIN NANHEKHAN, Delft University of Technology, the Netherlands

In this article, we explore how to improve the user experience of the TU Delft Repository. This library is an important place for students to find inspiration for their own thesis or to keep abreast of the latest progress on interesting topics. We found that it can be difficult to explore the result of a query. We observed that query result exploration can be challenging, and addressed this issue by incorporating a keyword-based filtering method into the existing ranking algorithm. This improvement benefited users who were uncertain about their specific search objectives. However, in some cases, the search query scope was too narrow, leading to insufficient information being delivered to the user. Additionally, we discovered that the keywords were not standardized, causing the filter to overlook relevant documents with similar topics but slightly different keywords.

CCS Concepts: • **Information systems → Information retrieval**.

Additional Key Words and Phrases: Ranking Visualisation Strategy

ACM Reference Format:
Justin de Haan, Sérénic Monté, and Kevin Nanhekhan. 2023. Alleviating Information Overload of Document Retrieval for TU Delft Digital Collections: Information Retrieval Group 20. *J. ACM* , (March 2023), 9 pages.

1 INTRODUCTION

The TU Delft repository comprises tens of thousands of diverse documents, serving as a vital resource for students seeking inspiration for their theses or staying up-to-date with the latest advancements in interesting topics. Navigating this extensive database to find valuable information without being overwhelmed by the sheer volume of documents can be challenging. It is therefore important that a clear user interface is available. This interface should also be able to filter and group the information properly so that more useful documents are returned and the information overload on the user is reduced. The current user interface is limited herein, because some key aspects are not included in the current design. For example, it is difficult for the user to find relevant keywords in the returned set of documents. If the user has found information that is interesting and wants to see more of the same topic, this cannot be done immediately on the current page, for this the user has to enter a new search query. This is certainly a problem for searchers who are exploring new topics. The research question we aim to address is as follows: **How can we improve the visualization of the TU Delft repository search page, such that information overload is reduced?**

Authors' addresses: Justin de Haan, Delft University of Technology, Delft, the Netherlands; Sérénic Monté, Delft University of Technology, Delft, the Netherlands; Kevin Nanhekhan, Delft University of Technology, Delft, the Netherlands.

In this article, we present an enhanced document ranking system and interface for the TU Delft repository, aiming to reduce information overload and add keyword filtering. These modifications facilitate more efficient qualitative data retrieval. The changes done consist of two steps. The first step is to change the front-end by allowing the user to select keywords of a document on the search page. The second step is a document filter system that filters the retrieved documents by the selected keywords and then ranks them on these keywords. This reduces user information overload by making information easier to filter and more relevant. The implemented code for this has been made available on Github [1].

One of the key lessons learned are the benefits of keyword filters in exploring various topics within the search scope. However, the search query scope can occasionally be too narrow, resulting in insufficient information delivery. Additionally, the lack of standardized keywords limits the method's full potential. Improvements could be achieved, for example, by employing natural language processing to detect and merge keywords that are semantically equivalent.

2 RELATED WORK

Searchers employ a variety of search strategies while looking for information within digital libraries, depending on the task type (simple or complex) and the features of the search system [1]. The success of a search process depends on how well the system supports these search strategies and how well it enables users to understand the impact of what they have done[1].

Information overload is a common problem faced by users while searching for information, as they may have to deal with large volumes of data and complex relationships between different pieces of information, resulting in a high degree of complexity [4]. Search tasks that are classified as knowledge discovery may be considered complex search tasks, due to their open-ended and multi-faceted nature, and under-specification of the information need [10]. Exploratory search has been proposed as a useful approach for tackling complex search tasks [10]. The goal of exploratory browsing is to develop an understanding of the information that is available for the current information need.

Visually linked or semantically linked keywords and user-driven visualization techniques can be employed to mitigate information overload, as they can help users to more effectively explore, navigate, and understand the search space more. By presenting information in a visual and interactive manner, these approaches can help users identify relevant information more easily, reducing the cognitive load and making the search process more efficient.

Sabol & Vaes [3] developed a graphical representation of keywords by placing each keyword in its own visualization. The keywords can then be selected, which will highlight the visual in a distinct color. Semantically related keywords are accentuated using a complementary color to visually demonstrate the relationship between keywords. This enables users to evaluate related keywords, which in turn aids them in narrowing down or expanding their search.

Hoebler & Shukla [5] proposed a *visually linked keywords* approach to support discovery and learning, making it easier for searchers to identify relationships between keywords of interest and individual search results. Meanwhile, Ma & Ma [7] evaluated various graphical methodologies that capture keywords and their relationships, such as word clouds, color coding, and iconography, and conducted a comparison between them.

The user interaction with the ranking system also makes for an important topic to consider. The work by Jabbari et al. [6] explores the issue of the false drop and its effect on human-machine interactions. Here it highlights the complex problem of the human-computer interface environment not being well understood due to the difficulty in detecting human behaviour and motives during their search. The identified causes of this false drop are that similar words have

[1] https://github.com/IPconfig/IR-Project

different meanings in the documents and query, incorrect matching of terms outside their context and most importantly the incorrect assignment of index terms causing irrelevant documents to be retrieved. As to mitigate these causes to some extent, the paper highlights the importance of following the user interface design principles such as providing informative feedback, reducing working memory load, providing alternative interface environments for novice and skilled users and visualizing information.

3 METHODOLOGY

In this section, we outline the methodology for creating a new interface for the document retrieval system of the TU Delft online library. Figure 1 provides an overview of the process.

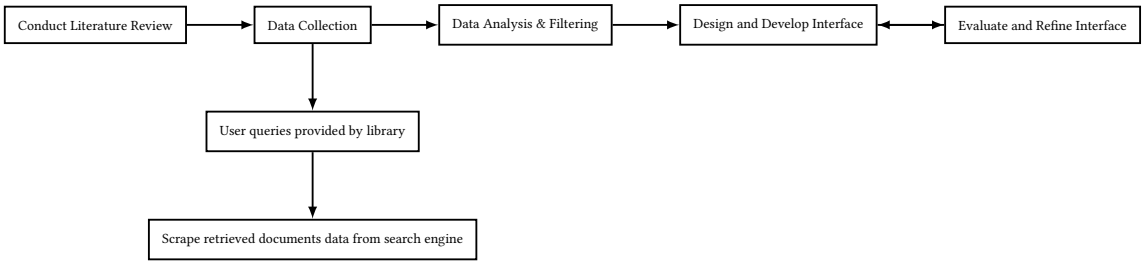

Fig. 1. Flow Diagram outlining the methodology

3.0.1 Literature Review. A literature review was conducted to identify existing research and approaches addressing similar research questions. Relevant literature was retrieved from the Web of Science library using the following query:

> (ALL=(document retrieval interface or exploratory search)) AND (TMSO==("4.48 Knowledge Engineering & Representation") AND TMIC==("4.48.1215 Information Visualization"))

This step informed the design of our study and ensured that the project was grounded in established research.

3.0.2 Data Collection. User queries provided by the TU library were used to retrieve the ranking of the resulting document set and their metadata from the search engine. A Python script was utilized to scrape the results and acquire the dataset.

3.0.3 Data Analysis and Filtering. The acquired dataset was analyzed and filtered to retain only interesting user queries. It turns out that not all documents contain metadata we could exploit. We collected a selection of search queries that returned a lot of documents with full metadata. However, after evaluating these queries, we decided not to use them because they were too generic, rendering even exploratory search inadequate.

3.0.4 Interface Development and Testing. Informed by the data analysis and literature review, the new interface was developed and tested.

3.0.5 Interface Evaluation and Refinement. The interface was evaluated based on any issues or problems that emerged during testing. As necessary, the interface was refined and modified to improve its effectiveness and usability.

3.1 System Design

In this section, we present the design of our proposed system, which integrates visually linked keywords with user-driven visualization techniques to enhance exploratory search and mitigate information overload. We detail the key components of our system, their interactions, and the design choices made to create a seamless and intuitive search experience for users.

3.1.1 System Architecture. The artifact employs containerized microservices to achieve modularity and scalability. We developed the artifact using the Python Flask framework, primarily due to the researchers' familiarity with it. Gunicorn serves as our WSGI server, and NGINX is employed as a reverse proxy service.

Owing to the unavailability of an API for the repository.tudelft.nl service, we forward our search requests to the site and scrape the results. To access the metadata, we export the data to a CSV file, which is then read by our service. This export process limits our search queries to a maximum of 3000 documents at a time and is the primary reason for slow document retrieval when a large number of documents are returned.

3.1.2 Keyword Representation. We employ the visually linked keywords approach as described by Hoebler & Shukla [5] and extend their work by adding a keyword frequency count in parentheses behind the keywords. The frequency count allows users to quickly ascertain how many times a particular keyword appears in the document set.

When a searcher selects a search result they believe to be relevant, the same keywords in all other search results are also marked in a visually distinct color within circle markers. This enables the searcher to visually scan the keywords associated with the other search results to discover those that use the highlighted keywords. If a searcher decides that a particular keyword is no longer of interest, another click removes the highlight.

Using perceptually distinct colors for keyword highlighting, a relationship between the same keywords used in different search results will be perceived [9]. However, when many keywords are selected in this manner, the ability to visually match the colours is diminished. Given the visual constraints, we restrict the number of highlighted colours to a maximum number of nine.

3.1.3 Keyword Selection. There is a lot of literature about keyword selection, as keywords are pivotal in a variety of retrieval tasks. Studies in this domain allow users to create their own keywords, crowd-source a common set of keywords, extract keywords using language models [2], or simply use the author-provided keywords. In our work, the keywords are extracted from the metadata of each search result and provided in a list beside the typical paper details. This option is chosen as it is the easiest to implement given the short amount of time.

4 EVALUATION

In this section, we discuss the evaluation of our project. The evaluation method we have chosen is AB testing. We begin by explaining what A/B testing is, why it is suitable for our project, and finally, the tests we have conducted.

4.1 AB testing

The purpose of AB testing is to compare user response to different versions of a product, in our case the TU Delft repository search page. By visualizing the web page differently, AB testing can investigate which alternative works better. AB testing can be divided into five steps. The first step is to define success, step two is to identify problems, step three is to specify a hypothesis, step four is to prioritize the identified problems, and the last step is to test the different versions [8].

We modified these steps slightly to better fit the limited scope of our research. Instead of naming and ranking all problems on the TU Delft Repository site, we focused on a main problem. Our adapted approach includes the following steps:

(1) Definition of success
(2) Identification of main problem
(3) Hypothesis definition
(4) The AB Test
(5) Evaluation

4.2 AB steps

Step one: Definition of success. The first step is to define what we want to achieve with our website. What is the purpose of the website and what is the ultimate goal of the website. The website is the TU Delft Repository of the TU Delft library and has as mission to upgrade student experience in interacting with collections. In this paper we aim to improve the experience by making it easier for students to explore search queries and find interesting topics. And make this process more efficiently and easier. **Our definition of success is that users find relevant papers more easily for a topic in the TU Delft Repository.**

Step two: Identification of problems. The problem encountered on the TU Delft Repository site is that it is difficult to find papers with similar topics. Most documents in the repository contain keywords, however it is not possible to see them in the search page. This limits the user in that it is sometimes difficult to quickly find the topic of a document. It is possible to search on keywords with the "subject" tag, however student who are exploring a subject do not know beforehand what relevant keywords are. **The problem we identified was the lack of easy keyword searching on the TU Delft repository site in the exploratory stage of literature research.**

Step three: Hypothesis. By adding the keywords to each entry and facilitating the keyword search, we hope that users can more easily find relevant papers on topics the user finds interesting. The hypothesis we assume is: **By showing keywords per document and sorting by chosen keywords, the number of relevant papers goes up for the user compared to the current website**

Step four: The AB Test. We asked two students who are both interested in nature in the living environment to perform a search in the TU Delft Repository, however we gave one of the students the real website and the other our own environment. In this section we describe the journey the two students undertook. Because of the limited time and test population we gave both students the same starting query.

The first student used the real repository website. The search query: "nature living" resulted in 252 documents of diverse topics, see figure 2. At the third paper, the student found something that appeals to him, namely the paper: *Towards a more nature-inclusive and climate resilient built environment: A framework and tool for the economic valuation of the costs and benefits associated with the implementation of vertical greening systems on buildings.* To find other papers on the same topic, the student went through the other retrieved documents. The student could not find any more interesting papers in the initial selection, since checking most papers for relevance takes too much time. However, when the student was finished with his search we tipped the student that you can select a relevant keyword in the selected paper as new search query. This new search query returned more relevant documents for this student. However, the scope of the initial query was removed, and thus some less relevant documents were present.

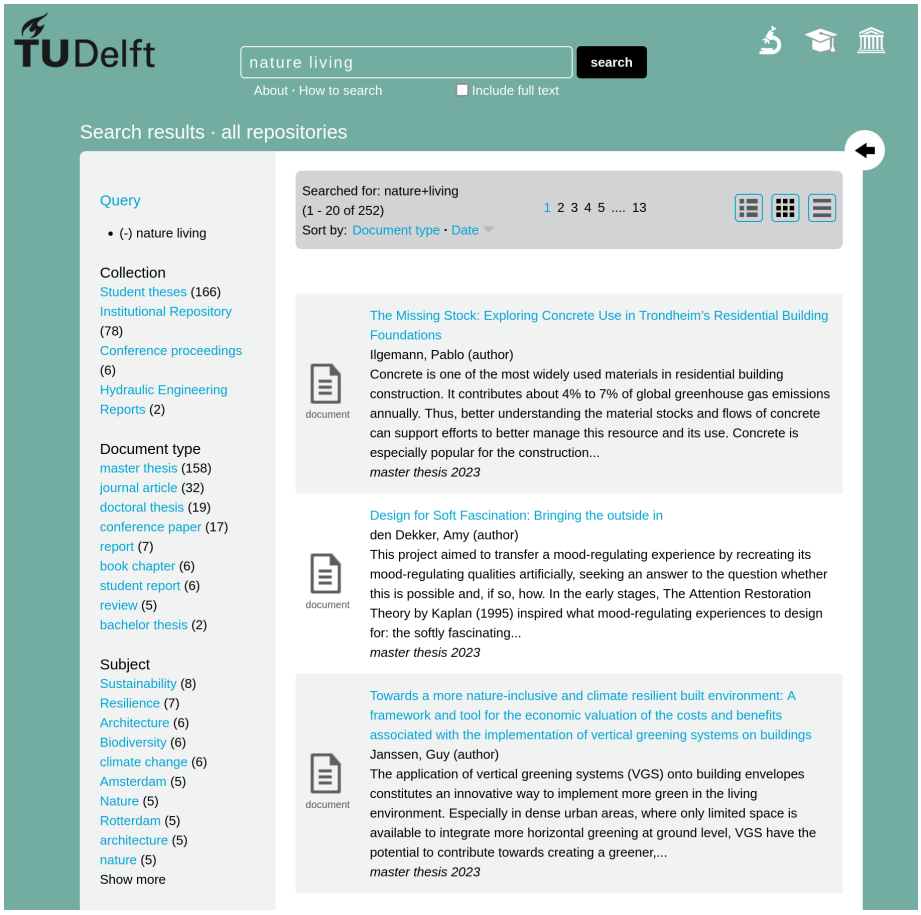

Fig. 2. Results for the query: "nature living" on the current TU Delft repository

The other student was given the experimental environment developed in this paper, the student entered the same initial search query and got the same initial results, see the left side of figure 3. Only now the student can click on the relevant keywords for this paper directly on the search screen. The keyword he presses is "Nature Based Solutions". Then the student is shown 3 relevant papers. One of the papers contains another interesting keyword, namely "Nature Inclusive Design". The student was able to relatively quickly narrow down on relevant information, whilst starting with a broad search query, see right side of figure 3. Because the initial search query was still presented the keyword filter filtered inside the initial return scope.

Step five: Evaluation of the results. In this section, we are going to evaluate the results of the AB test. First, we deal with the quality of the returned data. Second, we will cover the user experience.

The original site and the new site both return the same results for the same search query, which is to be expected since this study queries the database similarly to the original site. There is however a difference in the experience of the user. Because the new version displays keywords, the user can in most cases more quickly find out what the topic of the paper is and if it is relevant to the user. If a relevant topic is found by the user, the user can immediately filter the

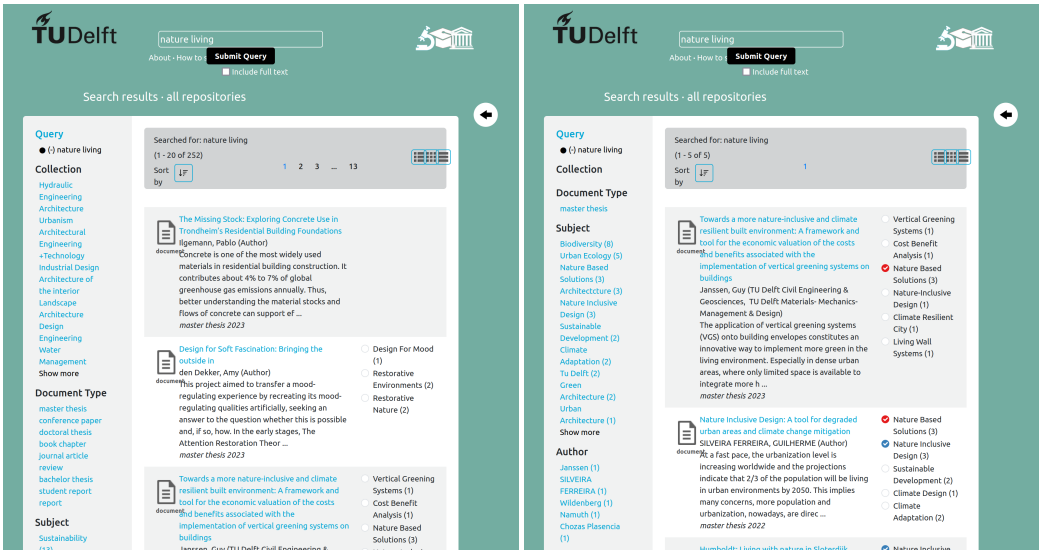

Fig. 3. Left unfiltered results for the query: "nature living". Right the filtered results with selected keywords: "Nature Based Solutions" and "Nature Inclusive Design".

documents on this keyword without losing the scope of the original query. In this filter, the user can quickly narrow the scope and find the relevant results. On the original site, this is not possible to preserve the scope of the query by clicking on the keyword. The new results can therefore contain documents that are not relevant to the original query.

The difference in user experience mainly comes out during the student's search for relevant information. On the original website, the student has to put more effort into deciding whether a document is relevant or not. Since the student cannot always decide if the document is relevant based on the title and first couple of sentences of the abstract. On the new site, the student also gets information on the keywords and can use this extra information to more quickly decide if the document is relevant. The filtered results of the new repository were received positively. Since the users perceived the original site the query reformulation by clicking on the keyword would lead to more results that are not always relevant to their initial query. An important side note is that we had a very small population to test this, and it may well be that in some cases this is not useful and the user prefers to have a larger scope by leaving out the original search terms in the new search query.

4.3 Conclusion of the evaluation

The AB test shows that the addition of the keywords to the retrieved documents and the possibility to filter them improved the user experience. The students were able to filter information faster and arrive at more specific search queries. However, not everything is positive about the developed software. Because the keyword search always falls within the scope of the original search query. This can cause the student to miss relevant information since it was not included in the original scope. Another drawback is the fact that not all papers use the same keywords, so two keywords are about the same topic but are spelled slightly differently, and those documents are excluded by the filter, like: "*nature inclusive design*" and "*nature-inclusive design*". Lastly, due to the limited number of tests and the small test population, it is not possible to draw a general conclusion from these results and more tests are needed for a conclusive answer.

5 CONCLUSIONS

In this article, we investigated ways to enhance the user experience of the TU Delft Repository by addressing limitations in its interface. We discovered that exploring the results of a query can be challenging and addressed this issue by incorporating a keyword-based filtering method into the existing ranking algorithm. This improvement benefits users who are uncertain about what they are looking for, as the provided keywords offer insight into the main topics of the papers and help users better understand the subdomains within their query.

However, the implemented method also has drawbacks. Filtering keywords within the scope of the given query can sometimes result in an information bubble that is too narrow to find enough relevant information. This can be solved by doing a new search through all documents with the selected keywords. Another challenge we faced is the lack of uniformity in keywords; many documents are lost due to slight variations in keywords with the same meaning, such as "*nature inclusive design*" and "*nature-inclusive design*". This problem can be addressed in the future by employing natural language processing (NLP) libraries that capture the context of different keywords, enabling the creation of a list of similar keywords that can be merged. Consequently, the number of keywords is reduced, and the keywords have a broader scope, making the chosen search less restrictive.

6 TEAM CONTRIBUTIONS

In this section, we outline the significant contributions made by the team members to the project. Each team member has contributed their skills, effort and knowledge to achieve the project objectives.

6.1 Justin

Literature review about search interfaces, contributed to the research direction by selecting a goal and doing a thorough literature review on exploratory search. Developed a web scraper to collect documents and their metadata for further analysis by Sérénic, created a skeleton framework for the project. implemented functionality from selected literature, merge the functionality into the interface created by Kevin and developed the interface further by creating paginated search results and connecting it to the TU Delft library repository for online retrieval and filtering. Written the *methodology*, and parts of the *related work* section of the report.

6.2 Sérénic

Literature review of exploratory search methods and information retrieval interfaces, collaborated with the group members to determine the research goal. Researched the search queries that are frequently used by the users of the TU Delft Repository. Participated in programming our framework. Researched evaluation methods and supervised the AB testing. Wrote the abstract, introduction, evaluation, conclusion.

6.3 Kevin

Literature review on the search interfaces for document retrieval and also collaborated with the other group members in determining the research goal. Mostly worked on the user interface by recreating the TU Delft library repository as much as was possible. This from its layout (results page and the individual document page) to also adding existing functionality of the faceted search and sorting documents. Also wrote parts of the *related work* section of the report.

REFERENCES

[1] Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications* 9, 3 (1995), 379–395. https://doi.org/10.1016/0957-4174(95)00011-W

[2] Oleg Borisov, Mohammad Aliannejadi, and Fabio Crestani. 2021. Keyword Extraction for Improved Document Retrieval in Conversational Search. *CoRR* abs/2109.05979 (2021). arXiv:2109.05979 https://arxiv.org/abs/2109.05979

[3] Cecilia Di Sciascio, Vedran Sabol, and Eduardo Veas. 2017. Supporting Exploratory Search with a Visual User-Driven Approach. *ACM TRANSACTIONS ON INTERACTIVE INTELLIGENT SYSTEMS* 7, 4, SI (Dec. 2017). https://doi.org/10.1145/3009976

[4] Orland Hoeber, Dolinkumar Patel, and Dale Storie. 2019. A Study of Academic Search Scenarios and Information Seeking Behaviour. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (Glasgow, Scotland UK) *(CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 231–235. https://doi.org/10.1145/3295750.3298943

[5] O Hoeber and S Shukla. 2022. A study of visually linked keywords to support exploratory browsing in academic search. *JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY* 73, 8 (Aug. 2022), 1171–1191. https://doi.org/10.1002/asi.24623

[6] Leila Jabbari, Hassan Mantegh, and Mila Malekolkalami. 2021. An Explanation of the False Drop in Information Retrieval in Human-Computer Interaction. *American Journal of Information Science and Technology* 5 (01 2021), 80. https://doi.org/10.11648/j.ajist.20210503.14

[7] XY Ma and H Ma. 2020. Comparative study of graphic-based tag clouds: theory and experimental evaluation for information search. *ONLINE INFORMATION REVIEW* 44, 5 (Sept. 2020), 1135–1160. https://doi.org/10.1108/OIR-12-2019-0372

[8] Dan Siroker and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers* (1st ed.). Wiley Publishing.

[9] M.O. Ward, G. Grinstein, and D. Keim. 2015. *Interactive Data Visualization: Foundations, Techniques, and Applications, Second Edition.* CRC Press. https://books.google.nl/books?id=XHZ3CAAAQBAJ

[10] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query—Response Paradigm.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02260-9