Result Assessment Tool: A Software Toolkit for Conducting Studies Based on Search Results

Sünkler, Sebastian HAW Hamburg, Germany | sebastian.suenkler@haw-hamburg.de
Yagci, Nurce HAW Hamburg, Germany | nurce.yagci@haw-hamburg.de
Sygulla, Daniela HAW Hamburg, Germany | daniela.sygulla@haw-hamburg.de
von Mach, Sonja HAW Hamburg, Germany | sonja.vonmach@haw-hamburg.de
Schultheiß, Sebastian HAW Hamburg, Germany | sebastian.schultheiss@haw-hamburg.de
Lewandowski, Dirk HAW Hamburg, Germany | dirk.lewandowski@haw-hamburg.de

ABSTRACT

The Result Assessment Tool (RAT) is a software toolkit for conducting research using results from commercial search engines and other information retrieval (IR) systems. This software combines modules used for the design and management of studies, the automatic collection of search results through web scraping, and the assessment of search results by jurors using different scales in an assessment interface. Due to the flexibility of RAT, several types of studies can be implemented, for example, classification studies and qualitative content analyses in addition to classic retrieval tests. Therefore, RAT is a versatile tool and useful in various disciplines.

KEYWORDS:
search engines; web scraping; retrieval tests; research software

INTRODUCTION

Studies that rely on results from commercial search engines and other information retrieval systems usually require manual work in designing the test, collecting the search results, finding jurors, gathering their ratings, and analyzing the test results, making it challenging to conduct such studies on a large scale. Jurors are individuals who evaluate collected search results at various levels, e.g., assessing their relevance to a particular search task or performing classification tasks. For instance, a researcher interested in comparing the sources and relevance of the results from Google and Bing on a particular topic would need to design a test, then use the queries to search for results in both search engines, copy the URLs from the result pages, randomize the URL lists, distribute the URLs to the jurors for evaluation, make a list of all domains found in the results and compare them between the two engines. Finally, the researcher must collect and analyze the juror's scores. It is obvious that the described process is cumbersome and cannot be applied on a large scale. These problems have been evident for years, and some software solutions have been developed. However, these are tools primarily designed for one-time use in studies (e.g., Bar-Ilan & Levene, 2011; Tawileh et al., 2010; Trielli & Diakopoulos, 2020), prototypes that have not been further developed (Lingnau et al., 2010; Renaud & Azzopardi, 2012), and software for test collections (Dussin & Ferro, 2008; Koopman, 2014; Ogilvie & Callan, 2001) or for narrowly limited use cases (The Digital Methods Initiative, 2022; Thelwall, 2009). To integrate all steps of the testing process into a complete and sustainable solution, we develop the Result Assessment Tool (RAT). RAT is a software toolkit that enables researchers to conduct large-scale studies based on results from search engines and other IR systems. The software toolkit is unique due to the offered flexibility, automation of tasks such as the scraping of commercial search engines, and the possibility of using the integrated platform to evaluate search results. In addition, RAT provides functions for automatically analyzing the jurors' ratings and determining statistics, such as the overlap of search results between different search services. These features will support the quantitative evaluation of search results to assist manual interpretation by researchers in subsequent studies. In the following, the Result Assessment Tool is presented with all the components currently available.

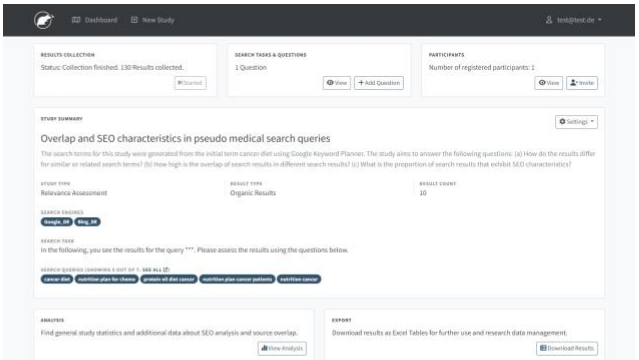
86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom. Author(s) retain copyright, but ASIS&T receives an exclusive publication license.

FUNCTIONALITY

RAT is a flexible web-based software toolkit developed in Python using the database PostgreSQL and Selenium testing suite for web scraping. Researchers can access a web interface to design studies, while participants can simultaneously use this interface to evaluate search results for predefined questions. The toolkit has been designed flexibly, allowing nearly all kinds of studies based on search results to be carried out; in addition to classic studies on IR, classification studies, data analyses, and even qualitative content analyses are possible.
RAT consists of the following six modules:

1. Test design: The test design module is the basic module that researchers use to define the type of study, the result type for the assessment (search results or snippets from search result pages, or both) and the type of access to the assessment interface. Access options include using a single access code (same code for all participants), personal access codes (each participant gets their own code), or group codes (used for group comparisons).
2. Result scraper: This module is used to define search tasks with search queries and select the search engines to be scraped. For instance, a researcher might write their own task descriptions (e.g., "You are searching for information on nuclear energy. How relevant is the following result?"), define a set of queries for their study ("nuclear power", "nuclear energy", "atomic energy", and so on), and define the search engines from which results should be collected (e.g., Google and Bing). Alternatively, lists of URLs can be uploaded to be made available for assessment. The content of the search results or URLs are scraped, and the source code and screenshots are stored in the database. As copies of the results and result documents are generated, all results will be available to jurors in the version when they were scraped; i.e., jurors will not experience any 404 errors or see documents updated in the meantime.
3. Definition of questions: RAT is very flexible in the design of questions. Question types include open-ended questions, Likert scales, sliders, and multiple-choice questions.
4. Assessment interface: In the assessment interface, jurors click through the copies of the results and answer predefined questions (e.g., "How relevant is the result shown here?", "Would you see this result as coming from a reputable source?", "Is this text well-written?").
5. Analysis module: The analysis module offers options for automatically analyzing the scraped results. Examples of such analyses include calculating the overlap of search results (Yagci et al., 2022) or measuring the application of search engine optimization (SEO) on web pages (Lewandowski et al., 2021).
6. Results export: Researchers can download the search results, juror scores, and the results of the analysis modules as tables at any time, enabling further use of all the collected and created data.

Figure 1 shows the overview page for a study in RAT. This view displays the status of the study, and it offers a glance into the search engines, search queries, questions, tasks, and number of participants so far. The study summary provides an overview of all the options specified in the test design process. It shows the study type (e.g., relevance assessment or classification study); result types to be assessed (e.g., organic results or search result snippets); search engines used in the study; and the search queries entered. "Analysis" provides access to the results of the automatic analysis, and "Export" opens a module to download the results in tabular form for further processing. Figure 2 shows an example of an assessment in the assessment interface of the Result Assessment Tool. The jurors see all predefined questions on the left side and a screenshot of a result to be evaluated on the right side.



RESULTS COLLECTION
Status: Collection finished. 130 Results collected.

SEARCH TASKS & QUESTIONS
1 Question

PARTICIPANTS
Number of registered participants: 1

STUDY SUMMARY
Overlap and SEO characteristics in pseudo medical search queries

The search terms for this study were generated from the initial term cancer diet using Google Keyword Planner. The study aims to answer the following questions: (a) How do the results differ for similar or related search terms? (b) How high is the overlap of search results in different search results? (c) What is the proportion of search results that exhibit SEO characteristics?

STUDY TYPE: Relevance Assessment
RESULT TYPE: Organic Results
RESULT COUNT: 10

SEARCH ENGINES: Google_DE, Bing_DE

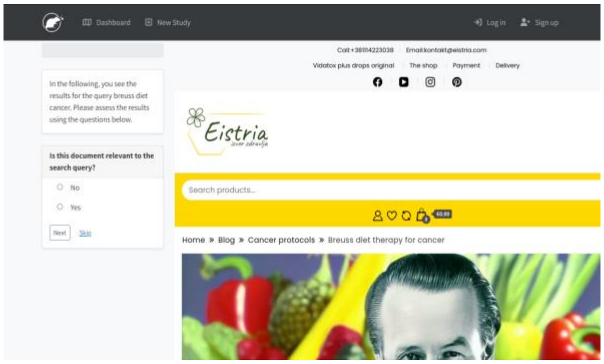
SEARCH TASK
In the following, you see the results for the query "". Please assess the results using the questions below.

SEARCH QUERIES (SHOWING 5 OUT OF 7. SEE ALL)
cancer diet, nutrition plan for chemo, protein oil diet cancer, nutrition plan cancer patients, nutrition cancer

ANALYSIS
Find general study statistics and additional data about SEO analysis and source overlap.

EXPORT
Download results as Excel Tables for further use and research data management.

Figure 1. Dashboard of RAT



In the following, you see the results for the query breuss diet cancer. Please assess the results using the questions below.

Is this document relevant to the search query?
○ No
○ Yes

Figure 2. Assessment interface of RAT

All collected results and assessments can be processed by automatic processes in the analysis module of RAT. The automated analysis for a study in the Result Assessment Tool computes and reports statistics about the study. These statistics provide an overview of the number of search queries, search results to be collected per query, and expected and collected results. These are standard statistics for any study; it is possible to calculate and display other statistics, for instance, the probability of using SEO on a search result or the calculation of the overlap of search results between the search engines used in a study. Both examples already show the potential of automated analyses since no further effort was required for the overlaps and the classification. We designed the analysis module so that researchers can extend it easily. In the future, we will extend this module with standard measures for information retrieval and readability scores, among others.

AVAILABILITY OF SOFTWARE DEMO, SOURCE, AND RESEARCH DATA

To adhere to the Findability, Accessibility, Interoperability, and Reusability (FAIR) principle (Wilkinson et al., 2016), we make the research data on the studies we conducted with RAT (Lewandowski et al., 2023a) and the source code (Lewandowski et al., 2023b) available. The demo is available at https://rat-software.org.

ACKNOWLEDGMENTS

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft (DFG); Grant No. 460676551).

REFERENCES

Bar-Ilan, J., & Levene, M. (2011). A method to assess search engine results. *Online Information Review*, *35*(6), 854–868. https://doi.org/10.1108/14684521111193166

Dussin, M., & Ferro, N. (2008). Design of a Digital Library System for Large-Scale Evaluation Campaigns. In B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, & J. Lippincott (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 400–401). Springer Berlin Heidelberg.

Koopman, B. (2014). Semantic Search as Inference. *ACM SIGIR Forum*. https://doi.org/10.1145/2701583.2701601

Lewandowski, D., Sünkler, S., & Yagci, N. (2021). The influence of search engine optimization on Google's results: A multi-dimensional approach for detecting SEO. 13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom. https://doi.org/10.1145/3447535.3462479

Lewandowski, D., Sünkler, S., Yagci, N., Schultheiß, S., Sygulla, D., & von Mach, S. (2023a). Result Assessment Tool (RAT). https://doi.org/10.17605/OSF.IO/T3HG9

Lewandowski, D., Sünkler, S., & Yagci, N. (2023b). Result Assessment Tool (RAT) [Computer software]. https://github.com/rat-software

Lingnau, A., Ruthven, I., Landoni, M., & van der Sluis, F. (2010). Interactive Search Interfaces for Young Children - The PuppyIR Approach. *2010 10th IEEE International Conference on Advanced Learning Technologies*, 389–390. https://doi.org/10.1109/ICALT.2010.111

Ogilvie, P., & Callan, J. P. (2001). Experiments Using the Lemur Toolkit. Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001.

Renaud, G., & Azzopardi, L. (2012). SCAMP. *Proceedings of the 4th Information Interaction in Context Symposium on - IIIX '12*, 286–289. https://doi.org/10.1145/2362724.2362776

Tawileh, W., Griesbaum, J., & Mandl, T. (2010). Evaluation of five web search engines in Arabic language. In M. Atzmüller, D. Benz, A. Hotho, & G. Stumme (Eds.), *Proceedings of LWA2010* (pp. 1–8).

The Digital Methods Initiative. (2022). *DMI Tools*. https://wiki.digitalmethods.net/Dmi/ToolDatabase

Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*. https://doi.org/10.2200/s00176ed1v01y200903icr004

Trielli, D., & Diakopoulos, N. (2020). Partisan search behavior and Google results in the 2018 U.S. midterm elections. *Information, Communication & Society*, *0*(0), 1–17. https://doi.org/10.1080/1369118X.2020.1764605

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data 3*, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Yagci, N., Sünkler, S., Häußler, H., & Lewandowski, D. (2022). A Comparison of Source Distribution and Result Overlap in Web Search Engines. Proceedings of the Association for Information Science and Technology, 59(1), 346–357. https://doi.org/10.1002/pra2.758