This is a preprint of the following article:
Sünkler, S., Yagci, N., Schultheiß, S., Von Mach, S., & Lewandowski, D. (2024). Result Assessment Tool: Software to Support Studies Based on Data from Search Engines. *Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science*, *14612*, 206–211. https://doi.org/10.1007/978-3-031-56069-9_19

Result Assessment Tool: Software to support studies based on data from search engines

Sebastian Sünkler[1], Nurce Yagci[1], Sebastian Schultheiß[1], Sonja von Mach[1], and Dirk Lewandowski[1,2]

[1] Hamburg University of Applied Sciences, Department Information, Media and Communication, Finkenau 35, 22081 Hamburg, Germany
firstname.lastname@haw-hamburg.de
[2] University of Duisburg-Essen, Campus Duisburg, 47057 Duisburg, Germany
firstname.lastname@uni-due.de

Abstract. The Result Assessment Tool (RAT) is a software toolkit for conducting research with results from commercial search engines and other information retrieval (IR) systems. The software integrates modules for study design and management, automatic collection of search results via web scraping, and evaluation of search results in an assessment interface using different question types. RAT can be used for conducting a wide range of studies, including retrieval effectiveness studies, classification studies, and content analyses.

Keywords: search engine evaluation, web scraping, retrieval tests.

1 Introduction

Conducting research using search engine data is challenging. Data from commercial search engines is not publicly available, and only a few search engines offer an API for accessing search results. In addition, recruiting jurors to evaluate search results is challenging, making it difficult to conduct studies on a large scale.

The information retrieval (IR) community has been developing software tools to conduct retrieval effectiveness studies for decades. However, the tools primarily consist of one-time use tools (e.g., Bar-Ilan & Levene, 2011; Tawileh et al., 2010; Trielli & Diakopoulos, 2020), prototypes that have not been developed further (Lingnau et al., 2010; Renaud & Azzopardi, 2012), and software to be

used with test collections instead of real-world data (Dussin & Ferro, 2008; Koopman, 2014; Ogilvie & Callan, 2001) or specific use cases (Digitalmethods, 2023; Thelwall, 2009). Therefore, we propose the Result Assessment Tool (RAT) as a sustainable solution that integrates all the necessary steps to conduct studies based on search engine data.

2 Significance of the Result Assessment Tool

On the one hand, the significance of the RAT derives from the need for a sustainable solution within the IR community to conduct large-scale studies based on search engine data. On the other hand, RAT is not limited to IR. In the field of health, for instance, health experts evaluated the quality (e.g., Janssen et al., 2018) or manually coded the content (e.g., Rachul et al., 2020) of health-related search results. In addition, researchers in the field of media and communication science classified search results e.g. based on content types and ideological biases (e.g., Ballatore, 2015). Since such studies usually rely on small data sets that are manually collected, evaluated, and analyzed, we see great potential for RAT, as its scalability can improve the studies by providing larger data sets and making jurors' work easier. Due to its modular construction, RAT possesses a high level of adaptability. Thus, its functionality can be expanded based on the needs of its users by adding new modules.

3 RAT use cases

Since evaluating the quality of information retrieval systems is an everyday use case when analyzing search engine data, the Result Assessment Tool was initially developed to conduct retrieval effectiveness studies. Even though quality is a multifaceted concept, the retrieval efficiency of search engines remains the foundation of all comprehensive quality evaluations. The retrieval effectiveness of Google and Bing was the subject of a study conducted with this early RAT version (Lewandowski, 2015). The jurors evaluated search results for 1,000 informational and navigational queries using a crowdsourcing strategy. Conducting classification studies is another use case. RAT supports both manual (i.e., by hand) and automatic (i.e., algorithmic) classifications of search results. The researchers or judges can perform the manual classification through the assessment interface. Automatic classifications can be accomplished by utilizing existing classifiers or by adding one's own classifiers. We refer to the study conducted with RAT by Hinz et al. concerning automatic result classification. The authors analyzed whether candidates use search engine optimization (SEO) on their personal websites for the 2021 federal election[1].

RAT is also capable of facilitating content analysis based on search results. An example is the work by Haider et al. (2023). For the Swedish term for wind

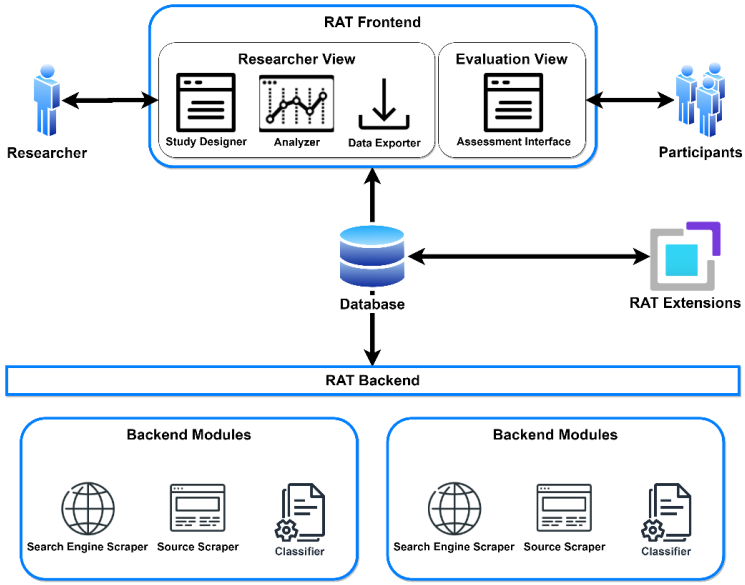
[1] The SEO-classifier implementation is described at (Lewandowski et al., 2021)

power (vindkraft), the query sampler extension (Schultheiß et al., 2023) generated 252 queries, for which the RAT scraped 5,710 search results[2].

The software also supports domain analyses, such as comparisons between search engines and countries. Yagci et al. (2022) analyzed the source diversity of Google and alternative search engines and the degree to which their root domains overlap. The top 10 search results from Google, Bing, DuckDuckGo, and MetaGer were scraped for 3,537 queries, resulting in 141,480 results.

4 Functionality of RAT

The Result Assessment Tool is an adaptable web-based software toolkit built with Python, the PostgreSQL database, and Selenium for web scraping. Researchers can use a web interface to design studies, while participants evaluate search results for predefined questions using the same interface. In addition to traditional IR studies, classification studies, and data analyses, qualitative content analyses are also possible due to the modular design of the toolkit. We develop the software based on the principles of user-centered design (UCD). The evaluation of the usability of the software is an integral part of the UCD process (Abras et al., 2004; International Organization for Standardization, n.d.), which is why continuous usability tests and heuristic evaluations are conducted. Fig. 1 shows the software architecture of RAT with its applications and modules.



[2] We developed a script that generates search queries based on keyword suggestions generated by the Google Ads API: https://developers.google.com/google-ads/api/

Fig. 1. Overview of the applications and their modules in RAT

The software consists of two applications that can be installed on separate computers and are linked through the database. This allows researchers to share the resources required for time-consuming and computationally intensive processes. The backend application provides scraping processes and classification tasks, while the frontend application provides a graphical interface for researchers to design studies and for study participants to evaluate search results. In addition, we provide an infrastructure for developers and researchers to create extensions for RAT that will be connected to RAT through the database.

The RAT Frontend application is a Flask GUI designed for researchers and study participants. It includes a Researcher View for designing studies and analyzing results and an Evaluation View for collecting participant assessments. The Study Designer is the basic module researchers use to define the study type, the assessment result type (search results and/or snippets from search result pages), and the type of access to the assessment interface. In the Researcher View, researchers can invite participants and define questions using Likert scales, open-ended questions, sliders, and multiple-choice questions. The Analyzer module computes and reports statistics about the study, including search queries, the expected number of results, and evaluation statistics. The Evaluation View in the RAT Frontend allows participants to register using a link provided by researchers. This approach enables anonymous access to participation and does not collect identifiable data. Answers are stored in a database, accessible using the Data Exporter to download the results.

The RAT Backend application processes inputs from the Researcher View in the RAT Frontend. The main module in the backend is the result scraper. During the scraping of results, metadata, source code, and a screenshot of the result are collected. The result scraper is based on Selenium, a test suite for web applications. A framework for automatically adding classifiers to analyze search results has also been implemented. The RAT Backend's architecture is based on a job management system using the Advanced Python Scheduler library (APScheduler). Jobs for all modules are created through inputs in the Researcher View at the RAT Frontend, and search results are collected by the Search Engine Scraper. Alternatively, lists of uniform resource locators (URLs) can be uploaded to be made available for assessment. Researchers can use scrapers we already provide (Google, Bing, DuckDuckGo) or add their search engine scrapers. The classification module allows automatic classifications based on the collected data. RAT provides the possibility of adding any classifier using templates and the database.

Availability of software demo, source code, and research data

To adhere to the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles (Wilkinson et al., 2016), we make the research data on the studies we conducted with the RAT available via the Open Science Framework (OSF)[3], provide a software demo[4], and provide the source code[5]. This is part of our sustainability strategy, which also serves as a marketing measure for building an international community of researchers and developers. Registering an account through the launch demo button is necessary for the software demo. The user is supported in creating the study within the tool.

Acknowledgments

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

References

Abras, C., Maloney-krichmar, D., & Preece, J. (2004). User-Centered Design. In W. Bainbridge (Ed.), *Encyclopedia of Human-Computer Interaction* (pp. 445–456). Sage Publications.

Ballatore, A. (2015). Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday*, *20*(7).

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

Digitalmethods. (2023). *DMI Tools*. Tool Database. https://wiki.digitalmethods.net/Dmi/ToolDatabase

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.

Haider, J., Ekström, B., Wallin, E. T., Lorentzen, D. G., Rödl, M., & Söderberg, N. (2023). *Tracing online information about wind power in Sweden: An exploratory quantitative study of broader trends*. http://dx.doi.org/10.13140/RG.2.2.27914.13766

[3] Repository for the research data generated with RAT: https://osf.io/t3hg9/

[4] The RAT software demo is available at https://rat-software.org/

[5] Repository for the source code: https://github.com/rat-software/

International Organization for Standardization. (n.d.). *ISO 9241-210:2019*. ISO. Retrieved October 10, 2023, from https://www.iso.org/standard/77520.html

Janssen, S., Käsmann, L., Fahlbusch, F. B., Rades, D., & Vordermark, D. (2018). Side effects of radiotherapy in breast cancer patients: The Internet as an information source. *Strahlentherapie Und Onkologie: Organ Der Deutschen Rontgengesellschaft ... [et Al]*, *194*(2), 136–142. https://doi.org/10.1007/s00066-017-1197-7

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

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, *66*(9), 1763–1775. https://doi.org/10.1002/asi.23304

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

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.*

Rachul, C., Marcon, A. R., Collins, B., & Caulfield, T. (2020). COVID-19 and 'immune boosting' on the internet: A content analysis of Google search results. *BMJ Open*, *10*(10), e040989. https://doi.org/10.1136/bmjopen-2020-040989

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

Schultheiß, S., Lewandowski, D., Von Mach, S., & Yagci, N. (2023). Query sampler: Generating query sets for analyzing search engines using keyword research tools. *PeerJ Computer Science*, *9*, e1421. https://doi.org/10.7717/peerj-cs.1421

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).

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. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. 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