

Result Assessment Tool

Software-Toolkit für die Durchführung von Studien auf der Grundlage von Suchergebnissen

*Sebastian Sünkler¹, Nurce Yagci¹, Daniela Sygulla¹, Sonja
von Mach¹, Sebastian Schultheiß¹, Dirk Lewandowski^{1,2}*

¹ Hochschule für Angewandte Wissenschaften Hamburg

² Universität Duisburg-Essen

{sebastian.suenkler, nurce.yagci, daniela.sygulla, sonja.vonmach,
sebastian.schultheiss, dirk.lewandowski}@haw-hamburg.de

Abstract

Das Result Assessment Tool (RAT) ist ein Software-Toolkit für die Durchführung von Studien mit Ergebnissen aus kommerziellen Suchmaschinen und anderen Information-Retrieval- (IR-) Systemen. Dabei bietet die Software Module für das Design und die Durchführung solcher Studien. Dazu zählen eine Komponente zum Studiendesign, eine automatisierte Sammlung von Suchergebnissen durch Web Scraping und ein Interface, in dem Juroren und Jurorinnen anhand verschiedener Skalen Bewertungen zu den Suchergebnissen abgeben können. Aufgrund der Flexibilität des Result Assessment Tool können verschiedene Studientypen damit realisiert werden. So lassen sich neben klassischen Retrievaltests auch Klassifikationsstudien und qualitative Inhaltsanalysen umsetzen. Damit ist das RAT vielseitig und fächerübergreifend einsetzbar.

Keywords: Suchmaschinen, Suchergebnisse, Web Scraping, Retrievaltest, Informationsqualität, Klassifikation, Software, Forschungssoftware

1 Einleitung

Die Durchführung umfangreicher und reproduzierbarer Studien, die sich auf die Ergebnisse kommerzieller Suchmaschinen und anderer Information-Retrieval- (IR-) Systeme stützen, ist aufgrund verschiedener Aspekte schwie-

rig. Zu diesen Aspekten zählen insbesondere die Sammlung von Suchergebnissen, die Suche nach Jurorinnen und Juroren sowie die Erfassung ihrer Bewertungen zu den Suchergebnissen und die anschließenden Datenanalysen. Dies erfordert in der Regel zeitintensive und aufwendige manuelle Maßnahmen und Schritte. Um diese Probleme anzugehen und um Studien mit Suchergebnissen durchzuführen, werden in der IR-Community seit Jahren Software-Tools entwickelt. Dabei handelt es sich jedoch in erster Linie um Tools für den einmaligen Gebrauch (Bar-Ilan/Levene, 2011; Tawileh et al., 2010; Trielli/Diakopoulos, 2020), um Prototypen, die nicht weiterentwickelt wurden (Lingnau et al., 2010; Renaud & Azzopardi, 2012), und um Software für Testsammlungen (Dussin/Ferro, 2008; Koopman, 2014) oder eng beschränkte Anwendungsfälle (Digitalmethods, 2023; Thelwall, 2009). Zusammenfassend fehlt es bisher an einer nachhaltigen Software-Lösung, um alle Schritte für solche Studien zu integrieren.

Mit dem Result Assessment Tool (RAT) entwickeln wir ein Software-Toolkit, das die genannten Schritte für die Durchführung von groß angelegten Studien mit Suchergebnissen vereinfachen soll. In dem folgenden Abschnitt stellen wir die Funktionsweise aller bisherigen Komponenten vor.

2 Funktionsweise des RAT

Das RAT ist eine flexible webbasierte Software, die in Python mit einer PostgreSQL-Datenbank sowie Selenium als Technologie für das Web Scraping entwickelt wird. Die Architektur der Software teilt sich dabei in ein Web-Frontend, das in Flask entwickelt wird, und ein Server-Backend auf, in dem alle Prozesse zur Datenerfassung (vor allem das Scraping der Suchmaschinen), Prozesse für automatisierte Analysen (z. B. Klassifikationen von Suchergebnissen) sowie die Datenbankoperationen durchgeführt werden.

Bei dem Web-Frontend stehen zwei Anwendungen zur Verfügung. In der einen Anwendung haben Forschende Zugang zu einem Interface, um Studien zu entwerfen, während die andere Anwendung den Zugang für Juroren und Jurorinnen bereitstellt, in dem Suchergebnisse bewertet werden. Dabei ist das Tool so flexibel gestaltet, dass praktisch alle Studien durchführbar sind, bei denen Suchergebnisse die Grundlage bilden. Neben klassischen Retrievalstudien sind somit Klassifikationsstudien und Datenanalysen oder auch qualitative Inhaltsanalysen möglich. Insgesamt setzt sich das Result Assessment

Tool aus sechs Modulen zusammen, die die Durchführung der Studien ermöglichen:

1. *Studien-Design*: Dieses Modul ist das Basismodul, mit dem Forschende die Art der Studie, den Ergebnistyp für die Bewertung (Suchergebnisse oder Ausschnitte aus Suchergebnisseiten oder beides) und die Art des Zugriffs auf das Bewertungsinterface festlegen.
2. *Suchmaschinen-Scraper*: Dieses Modul wird verwendet, um Suchaufgaben mit Suchanfragen zu definieren und die Suchmaschinen auszuwählen, die durchsucht werden sollen. So kann eine Forscherin ihre eigenen Aufgabenbeschreibungen und eine Reihe von Suchanfragen für ihre Studie definieren und anschließend die Suchmaschinen festlegen, von denen Ergebnisse gesammelt werden sollen (z. B. Google und Bing). Alternativ können auch Listen von URLs hochgeladen werden, um sie für die Auswertung zur Verfügung zu stellen. Der Inhalt der Suchergebnisse bzw. URLs wird gescraped und in der Datenbank gespeichert. Dieses Modul ist so gestaltet, dass Forschende eigene Scraper anhand von Skripten hinzufügen können.
3. *Gestaltung der Fragebögen*: Das RAT ist sehr flexibel bei der Gestaltung der Fragen. Zu den Fragetypen gehören offene Fragen, Likert-Skalen, Schieberegler und Multiple-Choice-Fragen.
4. *Bewertungs-Interface*: In diesem Interface klicken sich die Juroren durch die Screenshots der Ergebnisse und beantworten die vorab gestellten Fragen (z. B. „Wie relevant ist das hier gezeigte Ergebnis?“).
5. *Analysemodul*: Das Analysemodul bietet Optionen für die automatische Analyse der gescrapten Ergebnisse. Beispiele für solche Analysen sind die Berechnung der Überschneidung von Suchergebnissen (Yagci et al., 2022) oder die Messung der Anwendung von Suchmaschinenoptimierung (SEO) auf Webseiten (Lewandowski et al., 2021).
6. *Export der Ergebnisse*: Forschende können die Suchergebnisse, die Bewertungen der Juroren und Jurorinnen und die Ergebnisse der Analysemodule jederzeit als Tabellen herunterladen und so alle Daten weiterverwenden.

Abbildung 1 zeigt die Übersichtsseite in der Anwendung für die Forschenden im Web-Frontend. Diese Ansicht zeigt den aktuellen Status der Studie und bietet eine Übersicht zu den Suchmaschinen, Suchanfragen, Aufgabenstellungen und zur Anzahl der bisher Teilnehmenden. Die Studienzusammenfassung gibt einen Überblick über alle Optionen, die im Testdesignprozess festgelegt wurden, wie die Suchanfragen und Suchmaschinen.

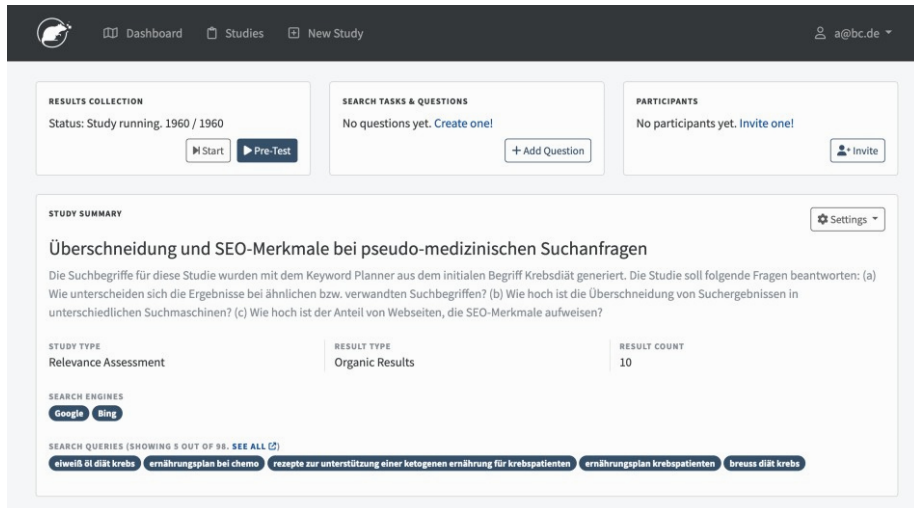


Abb. 1 Übersichtsseite zu einer Studie im RAT

Abbildung 2 zeigt ein Beispiel für eine Bewertung in der Bewertungsfläche des Result Assessment Tool. Die Juroren und Jurorinnen sehen auf der linken Seite unter der jeweiligen Suchanfrage alle festgelegten Fragen und auf der rechten Seite einen Screenshot eines zu bewertenden Ergebnisses.

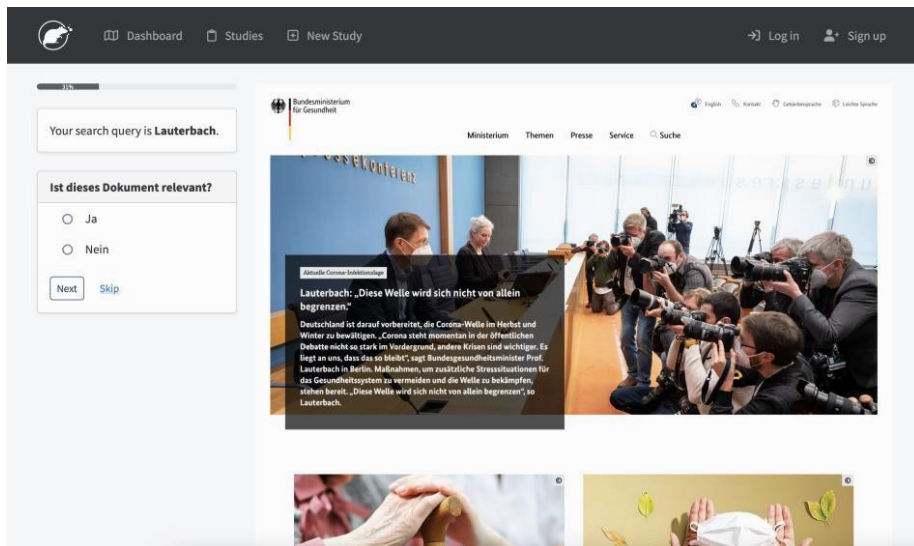


Abb. 2 Bewertungs-Interface des RAT

Alle gesammelten Ergebnisse und Bewertungen lassen sich durch automatische Analyseprozesse im RAT weiterverarbeiten. Dabei werden Statistiken über die Studie berechnet und berichtet. Diese Statistiken geben einen Überblick über die Anzahl der Suchanfragen, die zu sammelnden Suchergebnisse pro Anfrage und die erwarteten und gesammelten Ergebnisse, die suchmaschinenbedingt von der gewünschten Suchergebnismenge abweichen können. Zusätzlich lassen sich weitere optionale automatisierte Analysen durchführen, z. B. die Berechnung der Wahrscheinlichkeit der Verwendung von SEO für ein Suchergebnis oder die Berechnung der Überschneidung der Suchergebnisse zwischen den in einer Studie verwendeten Suchmaschinen. Diese beiden Beispiele zeigen bereits das Potenzial der automatisierten Analysen. In Abbildung 3 ist die Übersichtsseite einer Studie zu sehen, die die genannten Statistiken zeigt.

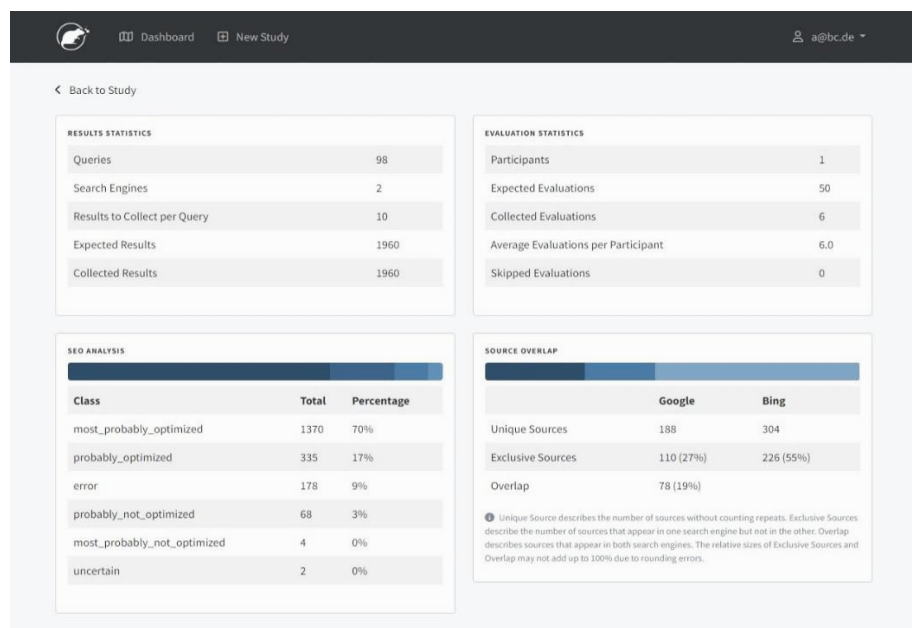


Abb. 3 Übersicht zu automatisch durchgeführten Analysen im RAT

Wir haben das Analysemodul so konzipiert, dass Forschende es leicht erweitern können. In Zukunft werden wir dieses Modul u. a. um Standardmaße für Information Retrieval und Readability Scores erweitern.

3 Evaluierung des Result Assessment Tool

Die Entwicklung der Software erfolgt gemäß den Standards der menschenzentrierten Gestaltung (User-Centered Design, UCD) (Abrás et al., 2004). Die Bewertung der Benutzerfreundlichkeit (Usability) der Software ist ein wesentlicher Bestandteil des UCD-Prozesses. Aus diesem Grund werden kontinuierliche Usability-Tests und heuristische Evaluationen verwendet.

4 Verfügbarkeit der Software-Demo, Quellcode und Forschungsdaten

Eine Demo der RAT kann von Forschenden unter <https://rat-software.org/> genutzt werden. Wir folgen bei der Veröffentlichung der Forschungsdaten, der Studien, die wir selbst mit dem RAT durchführen, und bei der Veröffentlichung des Quellcodes der Software den FAIR-Prinzipien (Wilkinson et al., 2016). Die Forschungsdaten können unter <https://osf.io/t3hg9/> und der Quelltext unter <https://github.com/rat-software/> eingesehen werden.

Danksagung

Die Entwicklung des Result Assessment Tools erfolgt im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts Relevance Assessment Tool (RAT), Förderkennzeichen 460676551.

Literatur

- Abrás, C.; Maloney-Krichmar, D.; Preece, J. (2004): User-Centered Design. In W. Bainbridge (Hrsg.): *Encyclopedia of Human-Computer Interaction* (S. 445–456). Thousand Oaks: Sage Publications.
- Bar-Ilan, J.; Levene, M. (2011): A method to assess search engine results. In: *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 (Hrsg.): *Research and Advanced Technology for Digital Libraries* (S. 400–401). Berlin, Heidelberg: Springer.
- Koopman, B. (2014): Semantic Search as Inference. In: *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. In: *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. In: *2010 10th IEEE International Conference on Advanced Learning Technologies*, S. 389–390. <https://doi.org/10.1109/ICALT.2010.111>
- Renaud, G.; Azzopardi, L. (2012): SCAMP. In: *Proceedings of the 4th Information Interaction in Context Symposium on – IIX '12*, S. 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 (Hrsg.): *Proceedings of LWA2010*.
- Thelwall, M. (2009): *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*. Cham: Springer Nature Switzerland. <https://doi.org/10.2200/s00176ed1v01y200903icr004>
- Trielli, D.; Diakopoulos, N. (2020): Partisan search behavior and Google results in the 2018 U.S. midterm elections. In: *Information, Communication & Society*, 0(0), 1–17. <https://doi.org/10.1080/1369118X.2020.1764605>
- Wilkinson, M. D.; Dumontier, M.; Aalbersberg, IJ. 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. In: *Scientific Data*, 3 (1), Article 1. <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. In: *Proceedings of the Association for Information Science and Technology*, 59 (1), 346–357. <https://doi.org/10.1002/praz.758>

In: W. Semar (Hrsg.): Nachhaltige Information – Information für Nachhaltigkeit. Tagungsband des 17. Internationalen Symposiums für Informationswissenschaft (ISI 2023), Chur, Schweiz, 7.–9. November 2023. Glückstadt: Verlag Werner Hülsbusch, S. 438–444. DOI: <https://dx.doi.org/10.5281/zenodo.10009338>