Search Studies an der HAW Hamburg

Dirk Lewandowski* und Sebastian Sünkler

Das Relevance Assessment Tool

Eine modulare Software zur Unterstützung bei der Durchführung vielfältiger Studien mit Suchmaschinen

<https://doi.org/10.1515/iwp-2019-0007>

Zusammenfassung: In diesem Artikel stellen wir eine Software vor, mit der sich Studien zu Such- und Informationssystemen realisieren lassen. Das Relevance Assessment Tool (RAT) soll umfangreiche Untersuchungen mit Daten von kommerziellen Suchmaschinen unterstützen. Die Software ist modular und webbasiert. Es lassen sich damit automatisiert Daten von Suchmaschinen erfassen. Dazu können Studien mit Fragen und Skalen flexibel gestaltet und die Informationsobjekte anhand der Fragen durch Juroren bewertet werden. Durch die Modularität lassen sich die einzelnen Komponenten für eine Vielzahl von Studien nutzen, die sich auf Web-Inhalte beziehen. So kann die Software auch für qualitative Inhaltsanalysen eingesetzt werden oder durch das automatisierte Scraping eine große Datenbasis an Web-Dokumenten liefern, die sich quantitativ in empirischen Studien analysieren lassen.

Deskriptoren: Software, Suchmaschine, Daten, Empirische Untersuchung, Information Retrieval System

The Relevance Assessment Tool

A modular software to support the implementation of various studies with search engines

Abstract: In this article, we present a software that can be used to carry out studies on search and information systems. The goal of the Relevance Assessment Tool (RAT) is to support extensive research using data from commercial search engines. The software is a modular web-based software with which data from search engines can be recorded automatically, studies with questions and scales can be

flexibly designed and the information objects can be evaluated by jurors on the basis of the questions. Due to the modularity the individual components can be used for a multitude of studies relating to web content. Thus, the software can also be used for qualitative content analyses or, through automated scraping, can provide a large database of web documents that can be analysed quantitatively in empirical studies.

Descriptors: Software, Search Engine, Data, Empirical, Studies, Information Retrieval System

L'outil d'évaluation de la pertinence

Un logiciel modulaire pour le support de multiples études avec des moteurs de recherche

Résumé: Dans cet article, nous présentons un logiciel qui permet de mener des études sur les systèmes de recherche et d'information. L'objectif du Relevance Assessment Tool (outil d'évaluation de la pertinence, RAT) est de soutenir des recherches vastes à l'aide de données provenant de moteurs de recherche commerciaux. Le logiciel est un logiciel modulaire basé sur le web qui permet de collecter automatiquement des données de moteurs de recherche, de présenter des études construites à l'aide de questionnaires et d'échelles de façon flexible et d'évaluer les objets d'information à l'aide de questions. Grâce à cette modularité, les composantes individuelles peuvent être utilisées pour une multitude d'études relatives à du contenu web. Ainsi, le logiciel peut également être utilisé pour une analyse de contenu qualitative ou, par le biais de la récupération automatique de données, fournir une grande base de données de documents Web qui peuvent être analysés quantitativement dans des études empiriques.

Deskriptoren: Logiciels, Moteur de recherche, Données, Études, Système de recherche d'information

*Kontaktperson: Prof. Dr. Dirk Lewandowski, Hochschule für Angewandte Wissenschaften Hamburg (HAW Hamburg), Fakultät Design, Medien und Information, Department Information, Finkenau 35, 22081 Hamburg, E-Mail: dirk.lewandowski@haw-hamburg.de
Sebastian Sünkler, Hochschule für Angewandte Wissenschaften Hamburg (HAW Hamburg), Fakultät Design, Medien und Information, Department Information, Finkenau 35, 22081 Hamburg, E-Mail: sebastian.suenkler@haw-hamburg.de

Einleitung

Die Durchführung von Studien zu Such- und Informationssystemen hat eine lange Tradition in der Informationswissenschaft; seien es klassische systemorientierte Evaluierungen wie Retrieval-Tests zur Ermittlung der Retrieval-Effektivität oder Studien zum Informations- und Suchverhalten der Nutzerinnen und Nutzer solcher Systeme.

Suchmaschinen sind der wichtigste Zugang zu Informationen im Web und haben eine hohe gesellschaftliche Relevanz (Purcell 2012; Koch & Frees, 2016). Daher ist auch das Interesse an Studien, die auf Daten aus dem Web basieren, und insbesondere auch auf Daten kommerzieller Suchmaschinen, sehr groß. Beispiele für Bereiche, in denen solche Studien mit Suchmaschinen durchgeführt werden, sind:

1. Studien zur Relevanz von Suchergebnissen (z. B. Lewandowski, 2008)
2. Studien zu Verzerrungen (*bias*) in Suchmaschinen (z. B. White & Horvitz, 2009; Ballatore, 2015; Ottenbacher, Bates & Clough, 2017)
3. Studien zur Messung der Überschneidungen der Suchergebnisse zwischen verschiedenen Suchmaschinen (z. B. Spink, Jansen, Blakely, & Koshman, 2006)

Eine Software, die für die Realisierung der genannten Studien mit Suchmaschinen flexibel eingesetzt werden kann, ist das Relevance Assessment Tool (RAT), das in der Forschungsgruppe Search Studies an der Hochschule für Angewandte Wissenschaften in Hamburg 2009 programmiert und kontinuierlich weiterentwickelt wurde. Die Software dient dazu, umfangreiche Studien mit Daten von kommerziellen Suchmaschinen zu unterstützen. Sie war für die Durchführung von Retrieval-Tests im Rahmen der Forschung zur Information-Retrieval-Evaluierung geplant; im Lauf der Jahre haben sich aber zahlreiche andere Anwendungsfälle aufgetan.

In diesem Artikel wird das RAT mit seiner Funktionalität und der technischen Implementierung dargestellt. Dabei werden zunächst die Anwendungsbereiche für das RAT erläutert. Dann wird dargestellt, welche Funktionen das Tool bietet, um Studien in den genannten Anwendungsbereichen durchzuführen und wie diese technisch umgesetzt wurden. Anschließend werden einige Forschungsarbeiten beschrieben, in denen das RAT eingesetzt wurde. Der Artikel endet mit einem Ausblick möglicher zukünftiger Weiterentwicklungen.

Anwendungsbereiche

Die Durchführung von Retrieval-Studien erfordert in der Regel eine Vielzahl zeitintensiver manueller Arbeitsschritte: die Konzeption der Studie, die Zusammenstellung von Suchergebnissen, das Finden von Jurorinnen und Juroren für die Bewertung der Suchergebnisse sowie die Erfassung und Analyse der Relevanzurteile. Dies führte zu der Idee, eine Software zu entwickeln, die durch einen modularen Aufbau ein Framework bietet, um solche Studien effizienter durchführen zu können und dabei auch große Dokumentensammlungen zu erstellen. Gleichzeitig sollte das Studiendesign wiederverwendbar sein. Außerdem wurde eine Funktionalität implementiert, die den Zugang zu den Studien für Jurorinnen und Juroren einfach gestaltet. Durch einen Zugangscode zu den Studien können in kurzer Zeit viele Bewertungen über Crowdsourcing gesammelt werden, was eine Akquise stark vereinfacht. Dadurch unterscheidet sich das RAT von Software, die jeweils nur für einzelne Studien entwickelt wurde (z. B. Fox, Karnawat, Mydland, Dumais, & White, 2005; Machill, Neuberger, Schweiger, & Wirth, 2003; Pan et al., 2007) und von Tools, die sich eher für den Einsatz vereinzelter Aufgaben eignen und einen breiteren Fokus haben. Dazu zählen u. a. die Digital Methods Tools¹ oder Crawling Software wie SocSci-Bot² und Webometric Analyst³ für Webanalysen oder Information-Retrieval-Evaluierungstools, die auf Testkollektionen basieren, wie das Lemur Toolkit⁴ und Revelation⁵.

Information-Retrieval-Evaluierung

Evaluierung ist ein elementarer Bestandteil in der Information-Retrieval-Forschung. Ergebnisse dieser Evaluierungen werden sowohl von Forschenden als auch von Entwicklenden genutzt, um Systeme zu verbessern. Da bei derartigen Studien die Systeme im Vordergrund stehen, werden sie auch als systemorientierte Studien bezeichnet. Sie unterliegen der Methodik des viel kritisierten Cranfield-Paradigmas, das trotz der bekannten Schwächen ein Standardvorgehen in der IR-Evaluierung darstellt (Voorhees, 2009). So ist dieses Paradigma etwa die Basis für die TREC-Evaluierungen, bei denen Systementwickler oder Forschende anhand von Dokumentensammlungen die Retrieval-Effektivität ihrer Systeme messen können.

¹ <https://wiki.digitalmethods.net/Dmi/ToolDatabase> [1.12.2018].

² <http://socscibot.wlv.ac.uk> [1.12.2018].

³ <http://lexiurl.wlv.ac.uk> [1.12.2018].

⁴ <https://www.lemurproject.org/lemur.php> [1.12.2018].

⁵ <http://ielab.github.io/revelation/> [1.12.2018].

Die Kritik an dem Paradigma beruht auf der Annahme eines konstruierten Nutzungsverhaltens, das davon ausgeht, dass Suchende alle Dokumente zu einer Anfrage nach und nach ansehen und die Relevanz jedes einzelnen Dokuments einschätzen. Davon unterscheidet sich jedoch elementar das tatsächliche Nutzungsverhalten. In der Mehrzahl der bekannten Anwendungsfälle werden nur wenige Suchergebnisse, vor allem auf den Top-Positionen, angeschaut. Das Konzept des RAT basiert auf diesem Paradigma, unterstützt darüber hinaus aber auch das interaktive Paradigma (s. u.). Dabei fußen die Konzeptionsmöglichkeiten im RAT auf wissenschaftlich fundierten Handlungsempfehlungen für IR-Evaluierungsstudien allgemein (Tague-Sutcliffe, 1992) und insbesondere für die Evaluierung von Websuchmaschinen (Gordon & Pathak, 1999; Hawking et al., 2001; Lewandowski, 2012).

Information-Retrieval-Evaluierung basiert traditionell auf Dokumentkollektionen, die sowohl Suchanfragen, Suchergebnisse als auch Bewertungen durch Jurorinnen und Juroren enthalten. Diese Kollektionen können Systementwickler und Forschende als Datenbasis für ihre Systeme nutzen, um die Retrieval-Effektivität zu messen. Das RAT arbeitet anders und ist für Studien mit kommerziellen Websuchmaschinen entwickelt worden, bei denen kein Zugang zu den Systemen besteht. Es funktioniert unabhängig von diesen Systemen und erstellt praktisch selbst die Dokumentenkollektionen.

Interactive Information Retrieval

Information Retrieval-Evaluierung hat sich in den letzten Jahren von systemzentrierten zu nutzungszentrierten Studien entwickelt (Kelly, 2009), da in diesen das tatsächliche Nutzungsverhalten und dadurch auch prozessorientierte Kennzahlen einbezogen werden können (White, 2016, S. 309ff.), die große Vorteile bei der Verbesserung von Suchsystemen bringen. Dabei ist es auch sinnvoll, Suchanfragen nicht isoliert, sondern für die Berücksichtigung des realen Verhaltens in einer ganzen Suchsession zu betrachten. Im Rahmen des RAT wurde dafür eine Erweiterung als Browser-Plugin geschaffen, mit dem sämtliche Interaktionen (von der Eingabe der Suchanfragen sowie die Klicks auf die Suchergebnisse) bei der Bearbeitung von Suchaufgaben auf-

gezeichnet werden. Die innerhalb der Session gesehenen Suchergebnisse können dann an das RAT übergeben und von der Nutzerin oder dem Nutzer bewertet werden.

Weitere Anwendungsbereiche neben der IR-Evaluierung

Neben der IR-Evaluierung lassen sich weitere Anwendungsbereiche identifizieren, in denen die Module aus dem RAT eingesetzt werden können. Dazu zählen etwa:

- Klassifikationsaufgaben von Informationsobjekten: Das RAT kann durch die flexible Gestaltung der Bewertungsskalen auch dazu genutzt werden, Informationsobjekte durch Jurorinnen und Juroren klassifizieren zu lassen. Voraussetzung ist, dass die Objekte durch eine URL aufrufbar sind.
- Zusammenstellung von Suchergebnissen und Suchinteraktionen für weitere Analysen: Da das RAT Suchergebnisse von kommerziellen Suchmaschinen automatisch erfassen kann (sowohl Trefferbeschreibungen auf den Suchergebnisseiten als auch den Quelltext der Webseiten selbst), können diese Inhalte für empirische Datenanalysen oder auch im Kontext von Machine Learning eingesetzt werden. Weitere Analysemöglichkeiten bestehen auch im Rahmen der qualitativen Inhaltsanalyse, da besonders die Quelltexte der Webseiten für solche Aufgaben nutzbar sind.

Es zeigt sich, dass das RAT zwar in erster Linie im Kontext der IR-Evaluierung für systemorientierte und nutungsorientierte Studien eingesetzt werden kann, aber dass sich vielfältige Möglichkeiten darüber hinaus ergeben. Die flexible Gestaltung der Bewertungsskalen, die skalierbare Datenerfassung sowie der Crowdsourcing-Ansatz zur Sammlung von Bewertungsurteilen können die zeitintensive manuelle Arbeit solcher Studien verringern und damit die Effizienz für Forschende steigern.

Beschreibung der RAT-Software

Das RAT wird seit 2009 entwickelt und besteht aus vier Basismodulen (Abb. 1), die durch ein Plugin zur Erfassung

Abbildung 1: Basismodule im RAT.

Abbildung 2: Schritte für die Testgestaltung im RAT.

von Suchinteraktionen ergänzt werden. Die Module setzen sich aus der Testgestaltung, dem Erfassen der Suchergebnisse, dem Bewertungsinterface zur Erhebung der Relevanzurteile und der Downloadmöglichkeit der Suchergebnisse mit den Urteilen zusammen.

Testgestaltung

Die Testgestaltung im RAT orientiert sich an den Anforderungen für die Durchführung von Retrievalstudien. Abbildung 2 zeigt die einzelnen Schritte, die durchlaufen werden müssen, damit die Studie für die Bewertung freigegeben werden kann. Dabei werden im ersten Schritt ein Name und eine Beschreibung festgelegt sowie Skalen und Fragestellungen für das Bewertungsinterface erstellt. Damit kann das RAT auch international eingesetzt werden. Im Schritt drei lässt sich der Zugang zum Bewertungsinterface definieren. Das bezieht sich darauf, ob ein Zugangscode für alle Jurorinnen bzw. Juroren oder Jurorengruppen mit individuellen Codes erstellt werden. Dies ist sinnvoll, wenn Nutzergruppen miteinander verglichen werden. Im vierten Schritt wird ausgewählt, welche Ergebnistypen in die Bewertung einfließen. Die Optionen beziehen sich auf die Suchergebnisse und/oder die Trefferbeschreibungen. Mit der Auswahl der Suchmaschinen werden die Systeme festgelegt, die evaluiert werden. Anschließend in Schritt sechs werden Fragen und Skalen definiert, bevor die Suchanfragen erstellt und an die Suchmaschinen geschickt werden⁶.

Suchaufgaben

Alle Studien basieren auf Suchaufgaben, für die die Suchergebnisse und/oder Trefferbeschreibungen aus den Suchmaschinen zusammengestellt werden. Dabei setzen sich die einzelnen Aufgaben aus einem Titel, der Suchanfrage und einer genaueren Beschreibung zusammen, bei denen auch festgelegt werden kann, welche Dokumente relevant sind. Diese Elemente basieren auf den typischen Beschreibungen von Suchaufgaben aus der TREC-Initiative (Har-

man, 2011, S. 30). Das RAT bietet die Möglichkeit, solche Suchaufgaben entweder einzeln oder auch eine Vielzahl solcher Aufgaben durch den Upload einer Excel-Tabelle im System zu speichern.

Bewertungsskalen und Fragestellungen

Das RAT stellt viele Optionen bereit, um Fragen zu erfassen, die in Dokumenten zu stellen (s. Abb. 3). Diese können im vorgegebenen Rahmen in ihrer Form frei gestaltet werden. Dabei werden folgende Skalen unterstützt:

1. Freie Skalen: Bei diesen Skalen kann die Jurorin oder der Juror eine Zahl in einem definierten Intervall eingeben (z.B. sinnvoll bei der Eingabe eines Prozentwerts zwischen 0 % und 100 %).
2. Geschlossene Skalen: Diese Skala wird verwendet, wenn die Anzahl wählbarer Werte überschaubar ist (z.B. eine 5-Punkte-Skala). Im Bewertungsinterface wird diese Skala in Radiobuttons dargestellt, aus denen die Jurorin oder der Juror einen Wert auswählen kann.
3. Checkboxes: Mit Checkboxes werden Fragen unterstützt, die im Gegensatz zu den geschlossenen Skalen mehrere Antwortmöglichkeiten zulassen.
4. Kommentarfelder: Sind freie Textfelder für offene Fragestellungen.
5. Schieberegler: Mit Schieberegler kann die Jurorin ihre bzw. der Juror seine Eingabe visuell tätigen. Dabei werden der Mindest- und Maximalwert angezeigt, aber nicht der ausgewählte Wert. Nur die Position auf dem Schieberegler ist erkennbar. Der Vorteil dieser Skala liegt darin, dass sie, im Gegensatz zu dem ordinalen Skalenniveau auf den geschlossenen Skalen, auf einem metrischen Niveau liegt und damit anspruchsvollere statistische Auswertungen ermöglicht.

Automatische Erfassung von Suchergebnissen

Die Erfassung von Suchergebnissen für das RAT basiert auf dem so genannten Screen Scraping. Screen Scraping

⁶ Eine ausführliche Darstellung der einzelnen Schritte findet sich in Lewandowski & Sünkler (2013).

Question Types

Question name	Question type	Options
<div> <div>+ Add new question</div> <div>+ Sort questions</div> </div>		
<p>Choose question type</p> <p>Please fill out all fields. More options will be displayed in the next step.</p> <p>Choose a question type.</p> <div> ☒ Closed Scale ☐ Slider ☐ Comment Box ☒ Checkbox </div> <p>Scale name:</p> <input type="text" value="mySurvey_Scale_1"/> <p>Question:</p> <input type="text" value="Did you see the logo in the bottom left corner?"/> <p>Should the options be arranged below each other or side-by-side?:</p> <p> ☐ Below each other ☐ Side by side <small>For Comment Boxes and Slider choose Below each other.</small> </p> <p>For what result type the scale should be displayed?:</p> <p> ☐ Description and Resultpage ☐ Description only ☐ Resultpage only </p> <div> <div>Next Step</div> <div>Cancel</div> </div>		

Abbildung 3: Definition der Fragen und Skalen im RAT.

bezeichnet eine technische Methode, um Inhalte aus Dokumenten zu extrahieren. Für das RAT bedeutet dies, dass Informationen wie die Trefferbeschreibung, die URL zum Suchergebnis, die Trefferposition usw. aus dem HTML-Quelltext der Suchergebnisseite (*search engine result page*, SERP) ausgelesen und gespeichert werden. Dafür muss definiert werden, welche HTML-Tags die gewünschten Informationen einschließen. Diese Definitionen werden in gesonderten Scrapern festgehalten.

Die Methode des Scrapings erlaubt es Forschenden, die gleichen Informationen zu gewinnen, die eine Suchmaschinennutzende nach dem Absenden einer Anfrage von der Suchmaschine erhält (Höchstötter & Lewandowski, 2009). Der Nachteil besteht in der Abhängigkeit vom HTML-Quelltext. Wenn die Suchmaschinen diesen verändern, muss auch der Scraper angepasst werden. Allerdings ist die Methode das Mittel der Wahl, da beispielsweise die Abfrage über Application Programmin Interfaces (APIs) zu verfälschten Ergebnissen führen kann⁷ und ein

Anbieter wie Google bspw. gar keine API für die Websuche mehr anbietet.

Momentan unterstützt das RAT folgende Scraper:

- Google (kommerzielle Suchmaschine)
- Bing (kommerzielle Suchmaschine)
- Econbiz (Bibliothekssystem)
- Otto (Produktsuche)
- eBay (Produktsuche)

Bewertungsinterface

Der Prozess im Bewertungsinterface wird in Abbildung 4 illustriert und besteht aus fünf Schritten. Die Jurorinnen bzw. Juroren bekommen über eine Web-Adresse die Zugangsseite zum Bewertungsinterface und werden auf-

⁷ Die APIs von Suchmaschinen sind sehr limitiert, so wird immer nur eine begrenzte Anzahl der Ergebnisse zurückgeliefert und die Anzahl der Suchanfragen ist auch begrenzt. Darüber hinaus entsprechen die

zurückgegebenen Suchergebnisse nicht den Ergebnissen, die bei direkten Anfragen angezeigt werden. (Tosques & Mayr, 2008)

Abbildung 4: Schritte im Bewertungsprozess des RAT.

Relevance Assessment Tool

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied SciencesFortschritt: 0% 100%

Suchaufgabe:

Finden Sie Informationen über die HAW Hamburg

Suchbegriffe:

haw hamburg

Bitte Bewerten Sie die Relevanz des Dokuments in Bezug auf die oben genannte Suchaufgabe auf einer Skala von 0 - 4?

- ☐ 0 - nicht relevant
- ☐ 1 - kaum relevant
- ☐ 2 - teilweise relevant
- ☐ 3 - relevant
- ☐ 4 - sehr relevant

Ist das HAW-Logo auf jeder Seite korrekt dargestellt?

- ☐ ja
- ☐ nein

Hier können Sie die Relevanz kommentieren:

Eingabe:

Nächstes Ergebnis

Überspringen

HAW Hamburg

Deutsch English Français Espagnol
(language available when flag highlighted)

English
Aktuell
Studiengänge

Bewerbung

- Bewerbungsfristen
- Deutsche, EU-Bewerber/innen
- Voraussetzungen
- Bewerbung
- Studienplatzvergabe
- Hochschulwechsel
- Studiengangswechsel
- Gaststudierende
- Nicht EU Bewerber/innen
- International Guest Students
- Mastersstudiengänge
- Vorkurse

International
Formales
Beratung
Campusleben

Studium

Design, Medien und Information Life Sciences Technik und Informatik Wirtschaft und Soziales

Deutsch > Bewerbung > Deutsche, EU-Bewerber/innen > Bewerbung

Online Bewerbung

An der HAW Hamburg bewerben sich Staatsangehörige aus Deutschland, der EU & EWR sowie Ausländer/innen und Staatenlose, die ihre Hochschulzugangsberechtigung in Deutschland erworben haben, innerhalb der Bewerbungsfristen online für ihre Studiengänge. Folgende Schritte sind (in den meisten Fällen) erforderlich:

- 1. Für einige Studiengänge ist die Bewerbung nur zum Winter- oder nur zum Sommersemester möglich. Informieren Sie sich auf der [Studiengangsübersichtsseite!](#)
- 2. Beginnen Sie mit dem [Selbsttest](#), der Ihnen bei der Entscheidung für ein Studienfach helfen soll. Er ist bis auf zwei Ausnahmen für die Bewerbung aller Studiengänge obligatorisch. [Übersicht](#)

Für die Onlinebewerbung ist eine Registrierung beim

Bewerbungsfristen 2011/2012

Wintersemester 2011/12: 1. Juni - 15. Juli
Sommersemester 2012: 1. Dezember - 15. Januar

Ausnahmen:

Information Engineering (Bachelor):
1. April – 15. Juli (Bewerbungen nur zum Wintersemester möglich)

Public Health (Master):
1. April - 30. September (Bewerbungen jeweils für das kommende Sommersemester möglich)

Renewable Energy Systems (Master):
1. April - 15. Juli (Bewerbungen nur zum Wintersemester möglich)

Health Sciences (Master):
01. April - 15. Juli (Bewerbungen nur zum Wintersemester möglich)

Abbildung 5: Bewertungsinterface im RAT.

gefordert, einen Zugangscode einzugeben. Nachdem sie ein Formular mit demografischen Fragen zum Alter und Geschlecht ausgefüllt haben, erhalten sie die Einweisung für die Studie.

Anschließend werden die einzelnen Informationsobjekte mit den definierten Fragen und Skalen angezeigt (Abb. 5). Die Jurorinnen und Juroren beantworten die Fragen, bis alle Informationsobjekte bewertet sind. Das RAT bietet die Option, Gutscheincodes zu hinterlegen, die im Anschluss an die Studie als Anreiz an die Jurorinnen und Juroren versendet werden können.

Download der Ergebnisse

Mit diesem Modul können die Forschenden die Daten (Suchergebnisse und Jurorenbewertungen) in Form einer Excel-Tabelle herunterladen. Dafür werden die gespeicherten Informationen aus der Datenbank ausgelesen, zusammengestellt und die Zuordnung der Suchergebnisse zu den Ergebnisbewertungen vorgenommen.

Abbildung 6: Systemarchitektur der interaktiven Komponente des RAT.

Interaktive Komponente für die Erfassung der Suchinteraktionen

Das Ziel der interaktiven Komponente des RAT ist die Verknüpfung von tatsächlich angefallenen Suchinteraktionen während der Bearbeitung von Suchaufgaben und eine direkt anschließende Bewertung der Suchergebnisse nach Abschluss einer Suchsession. Die Komponente für das RAT basiert auf dem Search-Logger-Framework, einem Browser-Plugin für Mozilla Firefox (Singer, Norbirsath, Vainikko, Kikkas, & Lewandowski, 2011). So können alle Interaktionsdaten (Auswahl der Suchmaschine, gestellte Suchanfragen, angeklickte Suchergebnisse) erfasst werden.

Dabei bearbeiten die Testpersonen definierte Suchaufgaben mit der Suchmaschine ihrer Wahl. Anschließend werden die geloggten Daten durch die interaktive Komponente im RAT ausgewertet. Es folgt eine Extraktion der gestellten Suchanfragen und anschließend werden diese an die Scraper übergeben, um die Suchergebnisse für die Jurorenbewertung zusammenzustellen. Die Testpersonen bekommen dann diese Ergebnisse zurück. Der große Vorteil liegt darin, dass die Nutzerinnen bzw. Nutzer tatsächlich die Suchergebnisse bewerten, die sie durch ihre eigenen Suchanfragen gefunden haben. Damit werden Zufriedenheitsmaße bei Suchsessions aus dem Interactive Information Retrieval mit klassischen dokumentbasierten Relevanzbewertungen kombiniert (White, 2016, S. 324). Die Systemarchitektur für diese Komponente ist in Abbildung 6 dargestellt.

Technische Implementierung der Software

Das Relevance Assessment Tool ist eine dynamische web-basierte Software, die hauptsächlich in PHP entwickelt wurde. Denn PHP kann man sehr gut mit anderen Webtechnologien wie HTML und Javascript nutzen und Datenbanken anbinden. Ein weiterer wichtiger Grund ist die sehr gute Implementierung der Screen-Scraping-Bibliothek cURL⁸.

Insgesamt besteht die Software aus mehreren Modulen, die in den vorherigen Abschnitten näher erläutert wurden. Für die Software-Architektur wurde auf das bewährte MVC-Prinzip (Model-View-Controller) zurückgegriffen, das eine Software auf drei Ebenen so einteilt, dass eine Wartung der Software mit geringem Aufwand möglich ist. Dazu können einzelne Komponenten und Module durch andere Technologien ausgetauscht werden, ohne das gesamte System anpassen zu müssen. Im einfachsten Fall könnte dies z. B. ein Wechsel des Datenbankmanagementsystems sein.

Die Softwarearchitektur des RAT wird in Abbildung 7 dargestellt. Sie zeigt die gängigen Schichten bei einer MVC-Architektur mit den dort verorteten Modulen. Bei einer MVC-Architektur werden die Komponenten auf drei Schichten verteilt. Die Präsentationsschicht enthält die Module, die Nutzende direkt sehen. Auf dieser Ebene erfolgen die Formulareingaben für die Testgestaltung in einer eigenen Anwendung sowie die Erfassung der Jurorenurteile im Bewertungsinterface. Darunter liegt die Applikationsschicht, in der die Eingaben verarbeitet werden. Die Module auf dieser Schicht geben auch Rückmeldungen an

⁸ <https://curl.haxx.se> [1.12.2018].

Abbildung 7: Softwarearchitektur des RAT.

die Präsentationsschicht. Das Fundament der Anwendung bildet eine MySQL-Datenbank. Diese hält die Metadaten vor und befindet sich auf der Datenschicht. Durch die Darstellung der Architektur wird auch die Kommunikation der Schichten untereinander deutlich. Kein Modul agiert unabhängig, da an bestimmten Stellen in der Software ein Datenaustausch stattfinden muss.

Das RAT in der Forschung

Das Relevance Assessment Tool wurde in den letzten Jahren bereits in einigen Forschungsarbeiten eingesetzt, die ganz verschiedene Forschungsthemen behandelten. Im

Folgenden wird eine Auswahl dieser Arbeiten präsentiert, um die Flexibilität des Tools zu verdeutlichen. Insgesamt wurden bereits 21 Forschungsarbeiten (Stand: 2018) realisiert einschließlich Abschlussarbeiten von Studierenden und Seminararbeiten⁹. Dies zeigt, dass das RAT bereits mehrfach erfolgreich in der Lehre eingesetzt wurde.

⁹ Es können hier nicht alle Arbeiten benannt werden, da es sich zum Teil um Projektarbeiten mit Unternehmen handelte, zu denen keine wissenschaftlichen Publikationen verfügbar sind.

Einsatzbereich	Beispiele
Studien zur Messung der Suchqualität	Schaer et al. 2016, Behnert & Plassmeier 2016, Lewandowski 2014, Podgajnik 2013, Günther 2012
Datenerfassung für Inhaltsanalysen	Günther 2016 Gather 2013 Siehe auch den Artikel zu AAPVL in diesem Heft
Untersuchung der Suchqualität im Kontext des Interactive Information Retrieval	Sünkler & Lewandowski 2017
Index-Vergleich	Desktop vs. Mobile im Rahmen einer Lehrveranstaltung
Vergleich von Suchvorschlägen	Studie für die Deutsche Telekom im Rahmen einer Lehrveranstaltung
Relevanzkriterien in Bibliotheken	Siehe Artikel zum LibRank-Projekt in diesem Heft
Bewertungsinterface für die Klassifikation von Inhalten	Bachelorarbeit (unveröffentlicht)
Gruppenvergleiche anhand von Interaktionsdaten	Bachelorarbeit (unveröffentlicht)

Studien zur Messung der Suchqualität

Die Retrievaltests, die in den letzten Jahren mit dem RAT, ursprünglich zur Unterstützung von Retrievalstudien entwickelt, umgesetzt wurden, machten deutlich, dass sich der Zeitaufwand in der Datenerhebung durch das Scraping von Suchmaschinen und bei der Erhebung von Relevanzurteilen erheblich reduziert. So zeigt Lewandowski (2015), wie mit Hilfe des RAT eine groß angelegte Retrievalstudie mit 1.000 Suchanfragen realisiert werden kann. Weitere Arbeiten beschäftigten sich mit dem Vergleich der Ergebnisqualität von Suchmaschinen und darüber hinaus gehenden Fragestellungen. So untersuchten Schaer et al. (2016) in einer Studie, wie die Relevanz von Suchtreffern auf den Top-Positionen im Gegensatz zu Suchergebnissen ausfällt, die weiter hinten in den SERPs auftauchen. Dafür wurde die Suchmaschine Millionshort (ein Service von Microsoft Bing) gescraped. Dort kann man auch Suchergebnisse ab der Trefferposition 1.000.000 ansehen. Eine studentische Projektarbeit nutze das RAT auch zur Messung der Retrievaleffektivität von Musikdiensten wie Apple iTunes und Amazon Music.

Datenerfassung für Inhaltsanalysen

Das RAT benötigt nicht immer ein eigenes Studiendesign. Bei einigen Fragestellungen kann es auch einfach zur Unterstützung bei der Erfassung von Web-Daten mit Hilfe des Scrapers eingesetzt werden. So wurden in drei Masterarbeiten Datenkollektionen mit dem RAT erstellt, die nicht im Kontext einer Retrievalstudie evaluiert wurden. Bei Gather (2013) wurden die Trefferbeschreibungen von Google und Bing mit dem RAT gespeichert und anhand von Besonderheiten in der deutschen Sprache untersucht. Günther (2016) überprüfte anhand einer Auswertung der Suchergebnisse in Google USA und Google Deutschland zu aktuellen Themen, ob Unterschiede in der Informationsvermittlung bestehen.

Des Weiteren wurden die Technologien des RAT in dem Forschungsprojekt AAPVL (Entwicklung von Automatisierten Analyseverfahren zur Identifizierung und Bewertung von nicht verkehrsfähigen Produkten des Virtuellen Lebensmittelmarktes) verwendet, um die Datenbasis für automatisierte Inhaltsanalysen von Suchergebnissen zu schaffen (Krewinkel et al., 2016). Dabei wurden sowohl Klassifikationen der Suchergebnisse als auch Inhaltsanalysen durchgeführt, um Informationen zu den Lebensmittelhändlern zu speichern.

Vergleich von Rankingalgorithmen in einem Bibliothekssystem

In bibliothekarischen Informationssystemen werden Rankingalgorithmen auf verschiedene Bibliotheksobjekte wie Bücher, wissenschaftliche Artikel und Arbeitspapiere angewendet. Manche dieser Materialien sind als Volltexte verfügbar, bei anderen werden nur bibliografische Merkmale und Abstracts zur Verfügung gestellt. Da normalerweise auch in solchen Systemen nur textbasierte Rankingfaktoren verwendet werden, gibt es Probleme, wenn nur bibliografische Merkmale in dem System abrufbar sind. Im Rahmen des LibRank-Projekts sollten diese Probleme analysiert und intelligente Rankingverfahren gefunden werden, um die Recherchequalität zu steigern (Behnert & Plassmeier, 2016).

Das RAT wurde hierbei eingesetzt, um Suchergebnisse zu verschiedenen Rankingalgorithmen aus dem EconBiz-System der Zentralbibliothek für Wirtschaft (ZBW)¹⁰ zusammenzustellen und die Ergebnisbewertungen im RAT

¹⁰ <https://www.econbiz.de/> [1.12.2018].

vorzunehmen. Dabei wurden diese von verschiedenen Nutzergruppen erhoben.

Interactive Information Retrieval

Die interaktive Komponente mit allen Funktionalitäten wurde in der Studie von Sünkler & Lewandowski (2017) getestet. Dabei wurden die Suchanfragen der Testpersonen erfasst und im Hintergrund an die Scraper gesendet, um die Suchergebnisse zusammenzustellen und im Bewertungsinterface bewerten zu lassen.

Ziel war ein Vergleich der Suchqualität von Google und Bing. 64 Probanden bekamen je zwei Suchaufgaben und konnten die Suchmaschine für die Aufgabenlösung frei wählen. In einer Vor- und Nachbefragung wurde der Schwierigkeitsgrad der Aufgabe bewertet. Alle Testpersonen wählten Google für ihre Aufgabenbearbeitung, und nachdem sie eine Aufgabe mit ihren eigenen Suchanfragen bearbeitet hatten, wurden diese Suchanfragen an Google und Bing geschickt. Dann bewerteten die Probanden die Suchergebnisse im RAT-Bewertungsinterface. Dafür beurteilten sie die Suchergebnisse, die sie angeklickt hatten sowie die restlichen Suchergebnisse bis zur zehnten Trefferposition zu den selbst gestellten Anfragen.

Fazit und Ausblick zur Weiterentwicklung des RAT

Das RAT bietet ein flexibles Toolkit zur Durchführung verschiedener Studien zu Such- und Informationssystemen. Durch das Angebot aller notwendigen Module für die Gestaltung von Tests, von der Datenerhebung bis hin zum Design von Fragestellungen zu den Informationsobjekten und den Komponenten zum Loggen von Suchinteraktionen, eröffnen sich viele Möglichkeiten für wissenschaftliche Untersuchungen. Eine besondere Stärke ist auch, dass durch die Skalierbarkeit der Software großangelegte, repräsentative Untersuchungen umsetzbar sind. Mit dem RAT können alle Arten von Studien realisiert werden, in denen Informationsobjekte bewertet bzw. durch Testpersonen oder Jurorinnen und Juroren klassifiziert oder in denen qualitative oder quantitative Analysen von Webinhalten durchgeführt werden sollen. Die Zugänglichkeit zum Tool ist einfach, da es als Webanwendung in jedem herkömmlichen Webbrowser nutzbar ist und durch Login-Codes auch einen Crowdsourcing-Ansatz unterstützt. Dies ermöglicht eine einfache Rekrutierung von Jurorinnen und Juroren bzw. Testpersonen.

Das RAT befindet sich momentan in dem Status eines Software-Prototyps und wird laufend, je nach den Anforderungen der durchzuführenden Studien, erweitert. Dadurch wurden nach und nach neue Funktionen und Module entwickelt, die zum vorgestellten Status Quo führten. In Zukunft ist geplant, das Tool weiter auszubauen und den Prototypenstatus zu beenden. Dafür ist angedacht, Funktionen wie einen „Scraper-Shop“ anzubieten, bei dem jeder Entwickler Scraper herunterladen und eigene Scraper anbieten kann, um die bisher unterstützten Scraper fortlaufend und durch Inhalte aus der Informationswissenschaft zu erweitern. Ferner soll die Software mit weiteren Features zur Datenanalyse im Tool selbst und für die Gestaltung von Online-Experimenten ergänzt werden. Dazu ist geplant, Verfahren zur Manipulation von Suchergebnissen zu implementieren. Außerdem möchten wir die Software als Open Source anbieten, damit eine rege Weiterentwicklung gefördert wird.

Literatur

- Ballatore, A. (2015). Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday*, 20 (7). <http://www.firstmonday.org/ojs/index.php/fm/article/view/5597/4652> [1.12.2018]
- Behnert, C., & Plassmeier, K. (2016). Results of Evaluation Runs and Data Analysis in the LibRank project.
- Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 147–168. <https://doi.org/10.1145/1059981.1059982>
- Gather, A. (2013). Suchmaschinen und Sprache: eine Studie über den Umgang von Google und BING mit den Besonderheiten der deutschen Sprache. [Masterarbeit] Hamburg: Hochschule für Angewandte Wissenschaften.
- Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35(2), 141–180.
- Günther, M. (2012). Evaluierung von Suchmaschinen: Qualitätsvergleich von Google- und Bing-Suchergebnissen unter besonderer Berücksichtigung von Universal-Search-Resultaten. [Bachelorarbeit] Hamburg: Hochschule für Angewandte Wissenschaften.
- Günther, M. (2016). Welches Weltbild vermitteln Suchmaschinen? Untersuchung der Gewichtung inhaltlicher Aspekte von Google- und Bing-Ergebnissen in Deutschland und den USA zu aktuellen internationalen Themen. [Masterarbeit] Hamburg: Hochschule für Angewandte Wissenschaften.
- Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring Search Engine Quality. *Information Retrieval*, 4(1), 33–59.
- Höchstötter, N., & Lewandowski, D. (2009). What users see – Structures in search engine results pages. *Information Sciences*, 179(12), 1796–1812. <https://doi.org/10.1016/j.ins.2009.01.028>
- Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Infor-*

- mation Retrieval, 3(1–2), 1–224. <https://doi.org/10.1561/1500000012>.
- Koch, W. & Frees, B. (2016). Ergebnisse der ARD/ZDF-Online-Studie: Dynamische Entwicklung bei mobiler Internetnutzung sowie Audios und Videos. In *Media Perspektiven* 9(2016), S. 418–437. http://www.ard-zdf-onlinestudie.de/files/2016/0916_Koch_Frees.pdf [1.12.2018].
- Krewinkel, A., Sünkler, S., Lewandowski, D., Finck, N., Tolg, B., Kroh, L. W., ... Fritsche, J. (2016). Concept for automated computer-aided identification and evaluation of potentially non-compliant food products traded via electronic commerce. *Food Control*, 61, 204–212. <https://doi.org/10.1016/j.foodcont.2015.09.039>.
- Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, 64(6), 915–937. <https://doi.org/10.1108/00220410810912451>.
- Lewandowski, D. (2012). A framework for evaluating the retrieval effectiveness of search engines. In C. Jouis (Ed.), *Next Generation Search Engines* (pp. 456–479). Hershey, PA: IGI Global. <http://doi.org/10.4018/978-1-4666-0330-1.ch020>.
- Lewandowski, D. (2015). Evaluating the retrieval effectiveness of Web search engines using a representative query sample. *Journal of the Association for Information Science & Technology*, 66(9), 1763–1775. <https://doi.org/DOI: 10.1002/asi.23304>.
- Lewandowski, D., & Sünkler, S. (2013). Designing search engine retrieval effectiveness tests with RAT. *Information Services & Use*, 33(1), 53–59.
- Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2003). *Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen*. In M. Machill & C. Welp (Eds.), *Wegweiser im Netz* (pp. 13–490). Gütersloh: Bertelsmann Stiftung.
- Otterbacher, J., Bates, J., & Clough, P. (2017). Competent Men and Warm Women. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems – CHI '17* (pp. 6620–6631). New York, New York, USA: ACM Press. <https://doi.org/10.1145/3025453.3025727>.
- Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3), 801–823.
- Podgajnik, L. (2013). Relevanzanalyse niedrig gerankter Google-Suchergebnisse auf informationsorientierte Suchanfragen mit dem Relevance Assessment Tool. [Bachelorarbeit] Hamburg: Hochschule für Angewandte Wissenschaften.
- Purcell, K., Brenner, J., & Raine, L. (2012). *Search Engine Use 2012*. Search. Washington, DC. http://pewinternet.org/~media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf [1.12.2018].
- Schaer, P., Mayr, P., Sünkler, S., & Lewandowski, D. (2016). How Relevant is the Long Tail? In N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, ... N. Ferro (Eds.), *CLEF 2016* (Vol. 9822, pp. 227–233). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-445649_20.
- Singer, G., Norbisrath, U., Vainikko, E., Kikkas, H., & Lewandowski, D. (2011). Search-Logger: Analysing Exploratory Search Tasks. In *SAC2011. Proceedings of the 2011 ACM Symposium on Applied Computing* (pp. 751–756). New York: ACM.
- Spink, A., Jansen, B. J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major Web search engines. *Information Processing & Management*, 42(5), 1379–1391. <https://doi.org/10.1016/j.ipm.2005.11.001>
- Sünkler, S., & Lewandowski, D. (2017). Does it matter which search engine is used? A user study using post-task relevance judgments. In *Proceedings of the 80th Annual Meeting of the Association of Information Science and Technology*, Crystal City, VA, USA (pp. 405–414). <https://doi.org/10.1002/pr.a2.2017.14505401044>
- Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4), 467–490.
- Tosques, F., & Mayr, P. (2009). Programmierschnittstellen der kommerziellen Suchmaschinen. *Handbuch Internet-Suchmaschinen*, 116–147.
- Voorhees, E. M. (2009). I Come Not to Bury Cranfield, but to Praise It. In *Proceedings of the workshop on human-computer interaction and retrieval* (pp. 13–16). <https://doi.org/10.1080/2158379X.2013.849369>
- White, R. W. & Horvitz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in Web search. *ACM Trans. Inf. Syst.* 27, 4, Article 23 (November 2009), 37 pages. <http://dx.doi.org/10.1145/1629096.1629101> [1.12.2018].
- White, R. (2016). *Interactions with Search Systems*. New York: Cambridge University Press.

Prof. Dr. Dirk Lewandowski
Hochschule für Angewandte
Wissenschaften Hamburg (HAW Hamburg)
Fakultät Design, Medien und Information,
Department Information
Finkenau 35
22081 Hamburg
dirk.lewandowski@haw-hamburg.de

Dirk Lewandowski ist Professor für Information Research & Information Retrieval an der Hochschule für Angewandte Wissenschaften Hamburg. Weitere Informationen: www.searchstudies.org/dirk

Sebastian Sünkler
Hochschule für Angewandte
Wissenschaften Hamburg (HAW Hamburg)
Fakultät Design, Medien und Information,
Department Information
Finkenau 35
22081 Hamburg
sebastian.suenkler@haw-hamburg.de

Sebastian Sünkler ist wissenschaftlicher Mitarbeiter und Promovierender an der Hochschule für Angewandte Wissenschaften Hamburg. Weitere Informationen: www.searchstudies.org/sebastian