**Search Studies an der HAW Hamburg**

Christiane Behnert*, Kim Plassmeier, Timo Borst und Dirk Lewandowski

# Evaluierung von Rankingverfahren für bibliothekarische Informationssysteme

https://doi.org/10.1515/iwp-2019-0004

**Zusammenfassung:** Dieser Beitrag beschreibt eine Studie zur Entwicklung und Evaluierung von Rankingverfahren für bibliothekarische Informationssysteme. Dazu wurden mögliche Faktoren für das Relevanzranking ausgehend von den Verfahren in Websuchmaschinen identifiziert, auf den Bibliothekskontext übertragen und systematisch evaluiert. Mithilfe eines Testsystems, das auf dem ZBW-Informationsportal EconBiz und einer web-basierten Software zur Evaluierung von Suchsystemen aufsetzt, wurden verschiedene Relevanzfaktoren (z. B. Popularität in Verbindung mit Aktualität) getestet. Obwohl die getesteten Rankingverfahren auf einer theoretischen Ebene divers sind, konnten keine einheitlichen Verbesserungen gegenüber den Baseline-Rankings gemessen werden. Die Ergebnisse deuten darauf hin, dass eine Adaptierung des Rankings auf individuelle Nutzer bzw. Nutzungskontexte notwendig sein könnte, um eine höhere Performance zu erzielen.

**Deskriptoren:** Bibliothek, Nutzer, Informationssystem, Ranking, Relevanz

### Evaluating ranking procedures for library information systems

**Abstract:** This article describes a study on the development and evaluation of ranking methods for library information systems. For this purpose, possible factors for relevance ranking were identified based on the procedures in web search engines, transferred to the library context and systematically evaluated. Using a test system based on the ZBW information portal EconBiz and web-based software for the evaluation of search systems, various relevance factors (e.g., popularity in connection with topicality) were tested. Although the tested ranking procedures are diverse on a theoretical level, no uniform improvements compared to the baseline rankings could be measured. The results indicate that an adaptation of the ranking to individual users or usage contexts is necessary in order to achieve a higher performance.

**Descriptors:** Library, User, Information system, Ranking, Relevance

### Évaluation des procédures de classement pour les systèmes d'information dans le domaine bibliothéconomique

**Résumé:** Cet article décrit une étude sur le développement et l'évaluation de procédures de classement pour les systèmes d'information dans le domaine bibliothéconomique. À cet effet, des facteurs pour le classement par pertinence ont été identifiés à partir de méthodes utilisés par les moteurs de recherche en ligne, puis ils ont été transférés dans le contexte de la bibliothèque et ils ont fait l'objet d'une évaluation systématique. À l'aide d'un système de test basé sur le portail d'informations ZBW EconBiz et d'un logiciel web d'évaluation des systèmes de recherche, différents facteurs de pertinence (tels que la popularité en combinaison avec l'actualité) ont été testés. Bien que les méthodes de classement testées soient différentes sur le plan théorique, aucune amélioration univoque n'a pu être mesurée comparée au classement de base. Les résultats indiquent qu'une adaptation du classement à des utilisateurs ou des contextes d'utilisation particuliers est nécessaire pour atteindre de meilleures performances.

**Descripteurs:** Bibliothèque, Utilisateur, Système d'information, Classement, Pertinence

***Kontaktperson: Christiane Behnert,** Hochschule für Angewandte Wissenschaften Hamburg, Fakultät DMI, Department Information, Finkenau 35, 22081 Hamburg, E-Mail: christiane.behnert@haw-hamburg.de
**Kim Plassmeier,** Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW), Abteilung Innovative Informationssysteme und Publikationstechnologien, Neuer Jungfernstieg 21, 20354 Hamburg, E-Mail: k.plassmeier@zbw.eu
**Dr. Timo Borst,** Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW), Abteilung Innovative Informationssysteme und Publikationstechnologien, Düsternbrooker Weg 120, 24105 Kiel, E-Mail: t.borst@zbw.eu
**Prof. Dr. Dirk Lewandowski,** Hochschule für Angewandte Wissenschaften Hamburg, Fakultät DMI, Department Information, Finkenau 35, 22081 Hamburg,
E-Mail: dirk.lewandowski@haw-hamburg.de

# 1 Einleitung

Das menschliche Such- und Informationsverhalten ist heutzutage stark durch die Websuche und insbesondere durch Websuchmaschinen wie Google beeinflusst. So werden bei der Nutzung von Websuchmaschinen die obersten Positionen der ersten Trefferliste, die als Ergebnis auf eine Suchanfrage angezeigt wird (z. B. Barry & Lardner, 2011; Jansen & Spink, 2006; Pan u. a., 2007; Schultheiß, Sünkler, & Lewandowski, 2018), präferiert und in der Regel recht kurze Suchanfragen eingegeben (Höchstötter & Koch, 2008). Auch bei bibliothekarischen Informationssystemen zeigt sich das gleiche Such- und Browsing-Verhalten (Hennies & Dressler, 2006). Um die Erwartungen an das Informationssuchverhalten zu erfüllen, sollten Bibliothekskataloge sich eher wie Suchmaschinen „anfühlen" (Connaway & Dickey, 2010), insbesondere durch eine Sortierung der Suchergebnisse nach Qualität bzw. Relevanz (Lewandowski, 2010). Moderne bibliothekarische Informationssysteme, sogenannte *Discovery Systeme*, verwenden zwar Rankingalgorithmen, beschränken sich aber oft auf Text-Matching und die Gewichtung einzelner Felder aus den Metadaten (so wird z. B. das Vorkommen eines Suchworts im Titel höher gewichtet als das Vorkommen des Suchwortes im Abstract). Die Möglichkeiten, die Rankingfunktionen für bibliothekarische Informationssysteme theoretisch bieten, wurden lange nicht ausgenutzt. Einige Discovery Systeme begannen, zusätzliche Parameter in das Ranking zu integrieren (siehe z. B. Blenkle, Ellis, Haake, & Zillmann, 2015; Langenstein & Maylein, 2009). Jedoch fehlten bisher ein gezielter Überblick über potenzielle Rankingfaktoren sowie deren systematische Evaluierung.

Vor diesem Hintergrund war das Ziel des hier vorgestellten DFG-Forschungsprojekts LibRank[1] die Entwicklung und systematische Evaluierung von neuartigen Rankingverfahren für bibliothekarische Informationssysteme aufbauend auf Erkenntnissen aus dem Bereich der Websuche. Dafür wurden zunächst auf den Bibliothekskontext übertragbare Relevanzfaktoren für Rankingalgorithmen identifiziert (Abschnitt 2) und in ein Relevanzmodell eingebettet (Abschnitt 3). Zur Erforschung der Rankingverfahren wurden diese in eine häufig genutzte Suchanwendung, das Fachportal EconBiz[2] des Leibniz-Informationszentrums Wirtschaft (ZBW), in einer Testumgebung integriert und mithilfe des *Relevance Assessment Tools* und menschlicher Jurorinnen und Juroren evaluiert (Abschnitt 4.1). Die Analyse der erhobenen Daten deutet darauf hin, dass nicht ein einzelner Rankingalgorithmus für alle Suchanfragen Ergebnisse in der gleichen Qualität bieten kann (Abschnitt 4.2). Der Beitrag diskutiert die Ergebnisse (Abschnitt 5) und schließt mit einem Fazit ab (Abschnitt 6).

# 2 Rankingfaktoren für bibliothekarische Informationssysteme

Die Sortierung der Suchergebnisse in einem Informationssystem, also das Ranking, will für eine informationssuchende Person die bestmögliche Rangfolge der Treffer generieren. Diese werden aus Sicht des Systems als relevant für die Suchanfrage bewertet (nach dem *best-match*-Ansatz) und auf Basis eines Rankingalgorithmus in absteigender Reihenfolge angezeigt. Dies bedeutet, je höher der Relevanzgrad (*score*), desto relevanter das Ergebnis und umso höher wird es gerankt. Für das Ranking in Websuchmaschinen wird eine Vielzahl unterschiedlicher Faktoren und Signale herangezogen. Diese lassen sich in Hinblick auf die dahinterliegenden Konzepte wenigen Gruppen zuordnen, die sich prinzipiell auch auf den Kontext (wissenschaftlicher) bibliothekarischer Informationssysteme übertragen lassen (Behnert & Lewandowski, 2015). Abbildung 1 zeigt fünf Gruppen, die neben den textstatistischen Verfahren als Basis in das Ranking mit einfließen: Popularität, Aktualität, Standort und Verfügbarkeit, Dokumenteigenschaften und Nutzerhintergrund (für eine tiefer gehende Darstellung der einzelnen Relevanzfaktoren verweisen wir auf Behnert & Lewandowski, 2015).

Insbesondere Faktoren der Gruppe Popularität, die auf dem Prinzip der „Weisheit der Vielen" (Surowiecki, 2005) beruhen, spielen im Ranking von Websuchmaschinen eine wichtige Rolle, da sie beispielsweise Klick- und Nutzungshäufigkeiten berücksichtigen (Lewandowski, 2018, S. 102ff.). Kommerzielle Anbieter wissenschaftlicher Informationssysteme wie Google Scholar integrieren ebenfalls Popularitätsfaktoren in ihr Ranking, z. B. die Zitationszahl eines Artikels (Beel & Gipp, 2009). Im Kontext von Biblio-

**1** Das Forschungsprojekt *LibRank: Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen* wurde kooperativ durch die Hochschule für Angewandte Wissenschaften Hamburg (HAW Hamburg) und die Deutsche Zentralbibliothek für Wirtschaftswissenschaften / Leibniz-Informationszentrum Wirtschaft (ZBW) durchgeführt. Informationen sowie Publikationen und Links sind auf der Projektwebseite www.librank.info/de [30.11.2018] zu finden.

**2** EconBiz ist das Fachportal für betriebs- und volkswirtschaftliche Literaturquellen, entwickelt und betrieben von der ZBW; https://www.econbiz.de/ [30.11.2018].

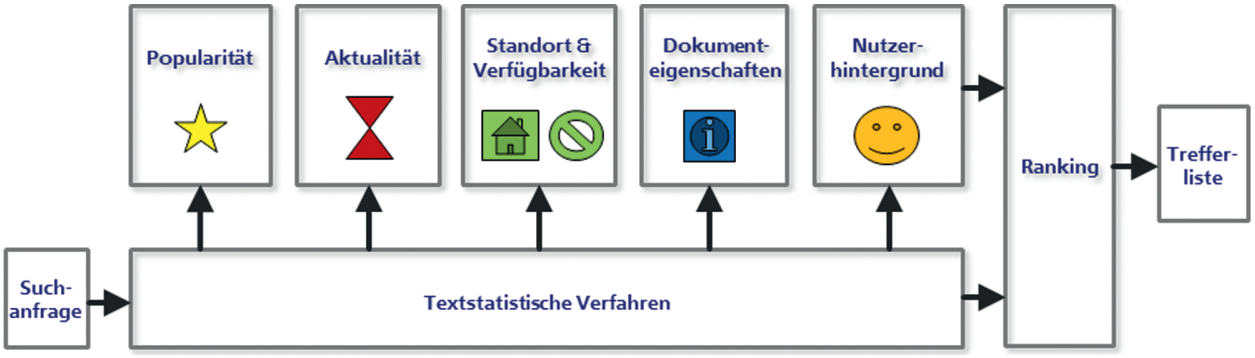


**Abbildung 1:** Gruppen von Rankingfaktoren. Quelle: Behnert & Lewandowski (2015, S. 389).

theken lassen sich unter anderem Ausleih- und Downloadzahlen dieser Gruppe zuordnen. Aktualität ist in traditionellen Bibliothekskatalogen der Standard, nach dem Ergebnisse in einer Trefferliste sortiert werden (absteigend nach Erscheinungsjahr). Die Kombination von Aktualität insbesondere mit Popularitätsfaktoren bietet großes Potenzial, denn so können Eigenschaften des Werkes mit Nutzungscharakteristika zusammen betrachtet werden.

Die Berücksichtigung des Standorts und der Verfügbarkeit liegt vor allem vor dem Hintergrund zunehmend elektronischer, lizenzierter Bestände nahe. Denn es ist anzunehmen, dass Personen, die sich physisch nicht in der Bibliothek befinden, sofort zugängliche, also freie oder von der Institution lizenzierte, Dokumente gegenüber gedruckten Materialien bevorzugen. Eigenschaften des Dokumentes wie Sprache oder Format können ebenfalls als Rankingfaktoren betrachtet werden, da selbst ein höchst relevantes Dokument nicht als solches beurteilt wird, wenn Suchende die Sprache nicht verstehen. In diesem Zusammenhang sind Informationen zum Hintergrund der informationssuchenden Person von großer Bedeutung.

## 3 Relevanzmodell

Zur Bewertung der Relevanz eines Dokumentes müssen die einzelnen Faktoren in einem Relevanzmodell zusammengeführt werden. Im einfachsten Fall könnten diese direkt in ein lineares Modell integriert werden. Dabei ergeben sich jedoch u. a. folgende Probleme:

- Es gibt keine gemeinsame Skala, um die verschiedenen Faktoren miteinander vergleichen zu können. Beispielsweise ist eine direkte Abwägung von Ausleihdaten mit Downloadzahlen nicht ohne weiteres möglich.
- Die Rohdaten enthalten systematische Verzerrungen. So weist beispielsweise die Häufigkeitsverteilung der Zitationen eine Abhängigkeit vom Alter der Werke auf (Zitationsobsoleszenz).

Unser Ansatz zur Abmilderung dieser Probleme basiert auf einer Methode aus der Szientometrie – der Methode der *Characteristic Scores and Scales* (CSS) von Glänzel & Schubert (1988). Die CSS-Methode wurde ursprünglich entwickelt, um in Zitationsverteilungen Klassen von Artikeln zu finden, die als „wenig zitiert“, „einigermaßen zitiert“, „bemerkenswert zitiert“ oder „hervorragend zitiert“ interpretiert werden können. Die Bestimmung der Klassen basiert dabei im Wesentlichen auf der Häufigkeitsverteilung der Zitationen.

Wir haben diese Methode auf die einzelnen Popularitätsfaktoren angewendet. Die Faktoren können dann innerhalb der einzelnen Klassen miteinander verglichen werden. Dabei ergeben sich plausible Äquivalenzen zwischen den Faktoren. Beispielsweise lässt sich daraus entnehmen, welche Anzahl von Downloads als gleichwertig mit der Anzahl an Zitationen angesehen werden kann. Diese Äquivalenzen sind aufgrund der Nichtlinearität der CSS-Methode ebenfalls nicht-linear.

Anschließend wurde definiert, wie die verschiedenen Klassen gegeneinander gewichtet werden sollen. Diese Klassengewichtung hat gegenüber eher technischen Parametern anderer Transformationen (z. B. Robertson, 2010) den Vorteil, dass sie für Fachexpertinnen und Fachexperten intuitiver anpassbar sein sollte. Im Folgenden haben wir ein lineares Schema verwendet (Gewicht der $i$-ten Klasse proportional zu $i$). Denkbar wäre aber auch ein exponentielles Schema (Gewicht der $i$-ten Klasse proportional zu $a^i$), damit wäre es schwieriger, Faktoren aus höheren Klassen durch Faktoren niedrigerer Klassen auszugleichen.

Dieses Vorgehen wurde ebenfalls benutzt, um systematische Effekte wie Zitationsobsoleszenz in den Daten zu reduzieren, und um verschiedene Datenbestände für einen einzelnen Faktor zusammenzuführen, wie beispielsweise Nutzungsdaten aus unterschiedlichen Systemen (Plassmeier, Borst, Behnert, & Lewandowski, 2015).

Da die thematische Relevanz eine zentrale Rolle bei der Relevanzbewertung spielt, wurde die Textstatistik-Komponente nicht in das lineare Modell integriert, sondern über

den sogenannten priorisierten Aggregationsoperator mit den anderen Faktoren gekoppelt (da Costa Pereira u. a., 2012). Die Idee hinter dem priorisierten Aggregationsoperator ist, eine Kompensation geringer thematischer Relevanz durch andere Rankingfaktoren, wie sie im linearen Modell möglich wäre, zu vermeiden[3]. Eine Abwandlung dieses Modells wurde ebenfalls im OpenSearch Track im Rahmen von TREC getestet (Schaer & Tavakolpoursaleh, 2016; Tavakolpoursaleh, Neumann, & Schaer, 2017).

# 4 Evaluierung der Rankingverfahren im EconBiz-Testsystem mit dem Relevance Assessment Tool

## 4.1 Methodik

Für das Design und die Durchführung der Testläufe wurde die von der HAW Hamburg entwickelte, web-basierte Software *Relevance Assessment Tool (RAT)* [4] verwendet. Mithilfe des RAT wurden bereits zahlreiche Studien[5] zur teilautomatisierten Evaluierung von Suchsystemen mithilfe menschlicher Juroren durchgeführt, auf deren Basis ein Framework zur Evaluierung entstand.

### 4.1.1 Anreicherung des Testsystems

Für den überwiegenden Anteil der Faktoren musste das Testsystem mit Daten aus externen Quellen angereichert werden, wodurch der Datenbestand teilweise eine starke Heterogenität aufwies (Behnert & Borst, 2015). Die Auswahl der Rankingfaktoren richtete sich vordergründig nach der Verfügbarkeit der notwendigen Datenquellen. Folgende Quellen wurden genutzt:

- Daten für nutzungsbasierte Faktoren, z. B. Anzahl der Views und Downloads, stammten aus der Web-Tracking-Komponente[6] von EconBiz und Daten aus RePEc[7] von LogEc[8] (Nutzungsverhalten; Popularität);
- Ausleihdaten, Informationen zu lokalen Exemplaren und den Beständen von anderen Bibliotheken wurden aus dem Bibliothekssystem des GBV abgerufen (Nutzungsverhalten, Erwerbungsverhalten; Popularität);
- für Daten zu Impact-Faktoren wurden die Daten von SCImago Journal Rank[9] integriert (Popularität);
- Zitationsdaten stammten von CitEc[10] und CrossRef[11] (Popularität);
- der Journal Ranking Guide (JRG)[12] offerierte Daten über den Begutachtungsprozess von Zeitschriften (Peer Review, Autorität; Popularität);
- anhand des Datensets von OCLC LOD Works[13] wurde die Anzahl an Auflagen eines Werkes bestimmt (Erwerbungsverhalten bzw. Nachfrage; Popularität).

Das Testsystem wurde mit diesen Daten angereichert und die Daten entsprechend den Ausführungen in Abschnitt 3 zum Relevanzmodell aufbereitet. Ein Großteil wurde für Faktoren der Gruppe Popularität genutzt. Informationen zum Nutzerhintergrund wurden nicht verwendet.

Tabelle 1 gibt einen Überblick über die evaluierten Rankingfaktoren im Rahmen von insgesamt drei Testläufen. Ein Textstatistik-Ranking, basierend auf der tf-idf-Implementierung von Solr mit heuristisch bestimmten Feldgewichtungen, und das EconBiz-Ranking (das Textstatistik-Ranking mit heuristischen, multiplikativen Boosts für Aktualität und frei verfügbare Dokumente) dienten jeweils als Baselines. Für ein Ranking (R23) wurden die Gewichtungsparameter für die einzelnen Faktoren über einen Learning-to-Rank Ansatz bestimmt. Dabei wurden die Parameter anhand der Relevanzbewertungen aus dem zweiten Testlauf durch direkte Optimierung der Kennzahl nDCG[14] über ein einfaches gradienten-freies Verfahren gefunden (*Coordinate Ascent*; Busa-Fekete, Szarvas, Élteto, & Kégl, 2012). Dieses Modell wurde im dritten Testlauf evaluiert. Für die übrigen Rankings wurden die Gewichtungs-

**3** Es wurde eine weniger strikte Variante des priorisierten Aggregationsoperators implementiert, die einen begrenzten Ausgleich durch andere Rankingfaktoren erlaubt. Dies trägt insbesondere den Unsicherheiten in den textstatistischen Verfahren bzgl. der thematischen Relevanz Rechnung (Plassmeier, 2016).

**4** Für eine detaillierte Beschreibung des Tools verweisen wir auf Lewandowski & Sünkler (2013).

**5** Eine Publikationsübersicht zu Studien, die das RAT verwendeten, ist zu finden unter http://searchstudies.org/relevance-assessment-tool/[30.11.2018].

**6** eTracker: https://www.etracker.com/ [30.11.2018].

**7** Research Papers in Economics (RePEc) ist eine Datenbank für internationale wirtschaftswissenschaftliche Publikationen und stellt verschiedene Services zur Verfügung; http://repec.org/[30.11.2018].

**8** Ein RePEc-Service ist LogEc, der Zugriffsstatistiken zu RePEc-Werken speichert; http://logec.repec.org/[30.11.2018].

**9** http://www.scimagojr.com/[30.11.2018].

**10** http://citec.repec.org/[30.11.2018].

**11** https://github.com/CrossRef/rest-api-doc [30.11.2018].

**12** Der Journal Ranking Guide war ein Service der ZBW und wird inzwischen nicht mehr angeboten.

**13** http://www.oclc.org/data.en.html [30.11.2018].

**14** Normalised Discounted Cumulated Gain.

**Tabelle 1:** Übersicht der Rankingfaktoren in den drei Evaluierungsläufen.

| R# | Evaluierungslauf 1 | R# | Evaluierungslauf 2 | R# | Evaluierungslauf 3 |
|---|---|---|---|---|---|
| R1 | Textstatistik (Baseline) | R11 | Textstatistik (Baseline) | R21 | Textstatistik (Baseline) |
| R2 | EconBiz-Ranking (Baseline) | R12 | EconBiz-Ranking (Baseline) | R22 | EconBiz-Ranking (Baseline) |
| R3 | Alle Faktoren | R13 | Alle Faktoren | R23 | Gelernte Variante |
| R4 | Nur Popularitätsfaktoren | R14 | Nur Popularitätsfaktoren | R24 | Nur Popularitätsfaktoren |
| R5 | Nur Aktualität | R15 | Nur Aktualität | R25 | Nur Aktualität |
| R6 | Nur Verfügbarkeit | R16 | Nur Verfügbarkeit | R26 | Nur Verfügbarkeit |
| R7 | Nur Dokumenteigenschaften | R17 | Nur Dokumenteigenschaften | R27 | Nur Dokumenteigenschaften |
| R8 | *Nutzungshäufigkeit + Erscheinungsdatum* | R18 | *Nutzungshäufigkeit + Erscheinungsdatum* | R28 | *Verfügbarkeit + Erscheinungsdatum + Nutzungshäufigkeit + Zitationen* |
| R9 | *Nutzungshäufigkeit + Zitationen* | R19 | *Nutzungshäufigkeit + Zitationen* | R29 | *Verfügbarkeit + Erscheinungsdatum + Nutzungshäufigkeit + Zitationen* |
| R10 | *Zitationen + Erscheinungsdatum* | R20 | *Zitationen + Erscheinungsdatum* | R30 | *Verfügbarkeit + Erscheinungsdatum + Nutzungshäufigkeit + Zitationen* |

Anmerkung: Die unterstrichenen Faktoren in den Rankings R28, R29, R30 wurden jeweils am höchsten bzw. höher gewichtet. In allen drei Evaluierungsläufen bestand ein Ranking aus rein textstatistischen Verfahren (R1, R11, R21) und aus dem EconBiz-Ranking (R2, R12, R22), das als Baseline diente. Rankings in kursiver Schrift bestanden aus einzelnen Faktoren aus unterschiedlichen Kategorien. So setzte sich R8 aus dem Popularitätsfaktor Nutzungshäufigkeit und dem Aktualitätsfaktor Erscheinungsdatum zusammen (ein weiterer Aktualitätsfaktor Zugangsdatum wurde nicht berücksichtigt).

parameter für die einzelnen Faktoren heuristisch festgelegt.

### 4.1.2 Design der Evaluierung

Für die ausgewählten Rankingverfahren wurden für jede Suchanfrage die jeweils ersten 20 Treffer extrahiert und aus diesen ein Pool an Dokumenten gebildet, der den Jurorinnen und Juroren zur Relevanzbewertung vorgelegt wurde (s. Abb. 2a). Wenn ein Suchergebnis von mehr als einem Ranking unter den ersten 20 Treffern erzeugt wurde, wurden die Dubletten entfernt (s. Abb. 2b). Das Design der Evaluierung beruht auf den Methoden zur Evaluierung von Websuchmaschinen bzw. Information-Retrieval-Systemen. Diese wurden auf den Kontext bibliothekarischer Informationssysteme übertragen (Behnert & Lewandowski, 2017). Die folgenden vier Kernbereiche bilden das Evaluierungsframework ab[15]:

- Suchanfragen (Darstellung, Arten, Quellen, Anzahl der Anfragen)
- Juroren (Auswahl und Anzahl der Juroren, auch im Hinblick auf Nutzermodelle)
- Suchergebnisse (Auswahl der Trefferelemente, Bewertungsablauf mit dem RAT)
- Testdurchführung (systematisches Vorgehen bei der Faktorenauswahl)

Den Jurorinnen und Juroren wurden die Suchergebnisse zu realen informationsorientierten Suchanfragen[16] zur subjektiven Relevanzbewertung vorgelegt. Abbildung 3 zeigt eine mit dem RAT zu bearbeitende Aufgabe, die aus einer Suchanfrage und einer kurzen Beschreibung eines dahinter liegenden Informationsbedürfnisses bestand. Das jeweilige Suchergebnis wurde sowohl mithilfe eines Schiebereglers auf einer visuellen Analogskala (zur Bewertung der graduellen Relevanz) als auch mittels binärer Entscheidung (relevant oder nicht relevant) bewertet, um analysieren zu können, ab welchem Relevanzgrad für Nutzerinnen und Nutzer ein Dokument eine gewisse Relevanz besitzt.

15 Für eine detaillierte Darstellung des Evaluierungsdesigns siehe Behnert (2016).

16 Im Gegensatz zu navigationsorientierten Suchanfragen erfordern informationsorientierte Anfragen in der Regel mehr als einen Treffer, um das Informationsbedürfnis Suchender zu befriedigen. Die Suchanfragen wurden aus den EconBiz-Logfiles generiert und manuell klassifiziert.

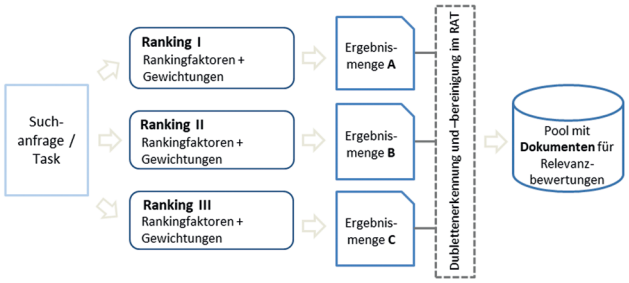


**Abbildung 2a:** Schematische Darstellung der Dokumentenpoolbildung für die Relevanzbewertung im RAT durch Juroren.

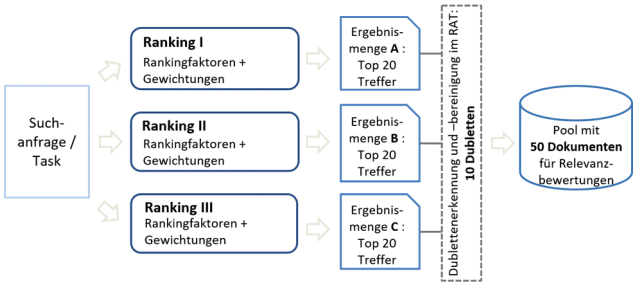


**Abbildung 2b:** Exemplarische Dublettenbereinigung im RAT.

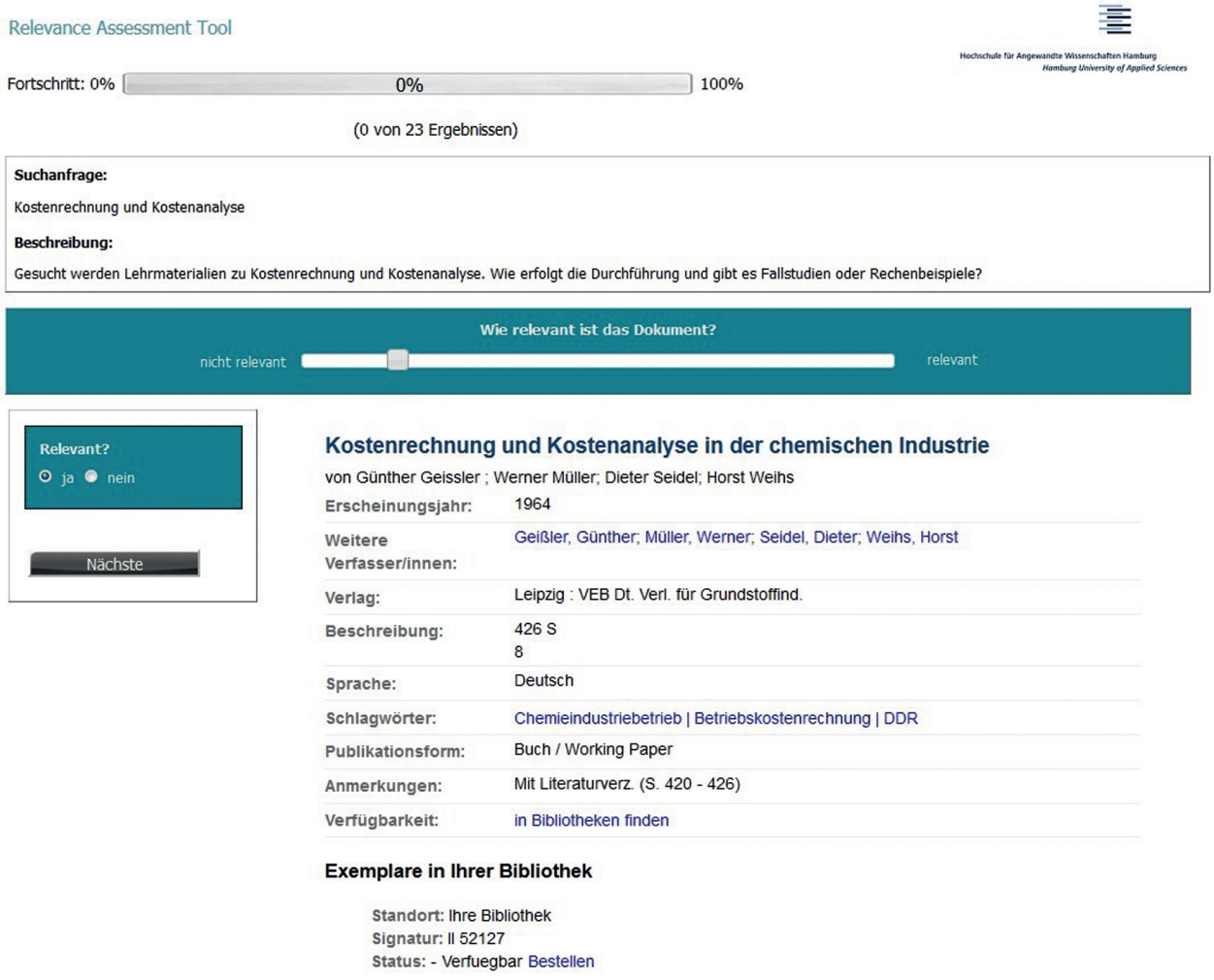


**Abbildung 3:** Beispiel einer Aufgabenansicht im Relevance Assessment Tool im Rahmen der drei Evaluierungsläufe.

## 4.2 Ergebnisse

Insgesamt wurden mithilfe des RAT in drei Testläufen 555 Suchaufgaben (Tasks) bearbeitet. So lagen schließlich 35.158 bewertete Dokumente vor. Die Daten wurden bereinigt und einer Plausibilitätsprüfung unterzogen, somit konnten Relevanzbewertungen aus 311 Tasks in die Auswertung einbezogen werden. Tabelle 2 gibt einen Überblick über die Eckdaten der Evaluierungsläufe.

In Abbildung 4 werden die mittlere Performance der getesteten Rankings mittels nDCG@10 und Precision@10 für die drei Testläufe sowie die simultanen 95 Prozent

**Tabelle 2:** Übersicht über die drei Evaluierungsläufe.

| Testlauf | Anzahl Juroren | Jurorengruppe | Anzahl Tasks gesamt | Anzahl Tasks pro Juror | Anzahl bearb. Tasks | Anzahl Tasks in Analyse | Anzahl bew. Dokumente |
|---|---|---|---|---|---|---|---|
| #1 | 4 | Fachreferenten (ZBW) | 120 | 30 | 83 | 41 | 3.470 |
| #2 | 8 | | 120 | 15 | 109 | 108 | 8.114 |
| #3 | 45 | Studierende | 450 | (var.) | 363 | 162 | 23.574 |
| Insg. | 57 | n.n. | 690 | n.n. | 555 | 311 | 35.158 |

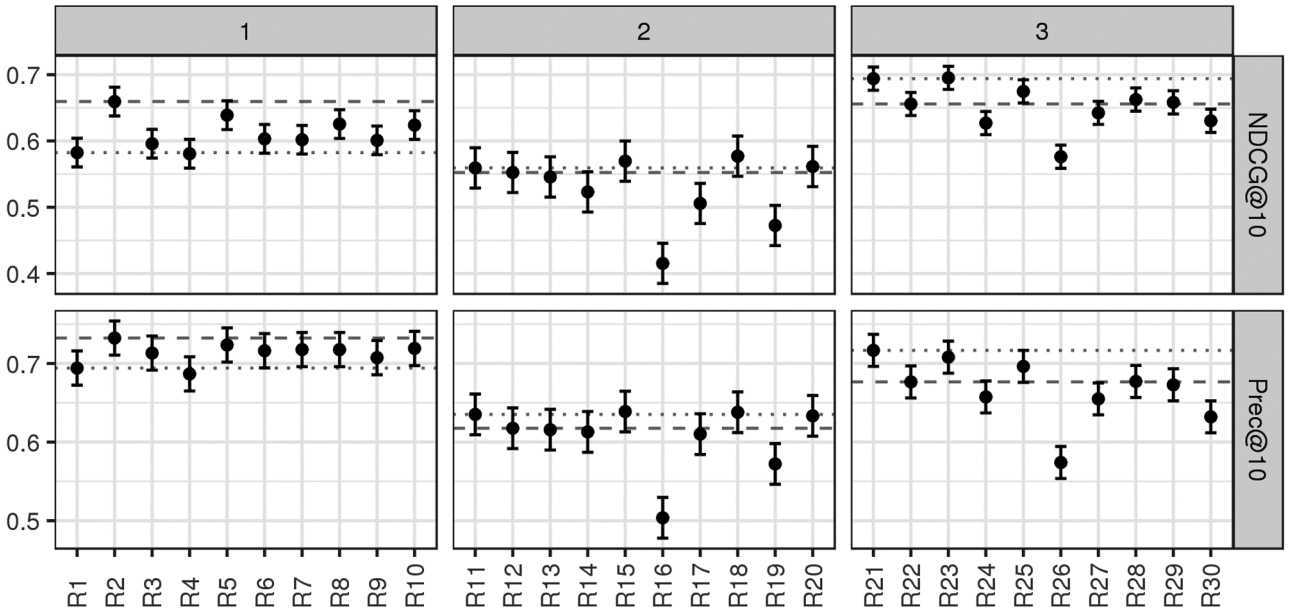


**Abbildung 4:** Durchschnittliche Performance der Rankingverfahren in den drei Testläufen. Die Fehlerbalken repräsentieren die simultanen 95 Prozent Konfidenzintervalle. Performance der Baselines: Textstatistik (gepunktet), EconBiz (gestrichelt).

Konfidenzintervalle dargestellt. Die Unterschiede wurden Sakai (2014) folgend mittels Tukeys HSD-Test auf statistische Signifikanz getestet.

Insgesamt konnte für keines der getesteten Rankings eine einheitliche Verbesserung gegenüber den Baseline-Rankings gemessen werden. Im ersten Lauf erzielte stattdessen die EconBiz-Baseline die beste Performance, im dritten Lauf die Textstatistik-Baseline. Statistisch signifikante Verbesserungen gegenüber einer Baseline konnte nur Aktualität (R5) im Vergleich zur Textstatistik-Baseline (R1) in Lauf 1 erzielen. Das gelernte Ranking (R23) erzielte in Lauf 3 die zweitbeste Performance und weist statistisch signifikante Verbesserungen gegenüber den Rankings R24, R26, R27 und R30 auf.

Im Vergleich zu der Textstatistik-Baseline enthalten die Top-10 Listen der getesteten Rankings im Mittel zwei bis fünf gemeinsame Dokumente. Die Varianz über die Suchaufgaben ist dabei groß. Somit gibt es einen deutlichen Effekt der Rankings auf die jeweils zehn besten Suchergebnisse, wobei dieser für einige Rankingverfahren sowie bei einigen Suchaufgaben weniger prägnant ausfällt. Bei den Testläufen zwei und drei ist die Tendenz beobachtbar, dass die mittlere Performance geringer ausfällt, wenn die Anzahl der gemeinsamen Dokumente abnimmt.

## 5 Diskussion

Die Ergebnisse folgen dem typischen Muster von Evaluierungsstudien (W. E. Webber, 2010): Es kann eine hohe Varianz über die Tasks (Informationsbedürfnisse) beobachtet werden[17]. Für einige Informationsbedürfnisse erreichen demnach alle Rankings eine gute Performance,

17 Detailreiche Darstellung in Behnert & Plassmeier (2016).

während für andere keines der Rankings eine gute Performance erreicht. Die hohe Varianz über die Performance der Rankings pro Task unterliegt ebenfalls großen Schwankungen zwischen den Aufgaben, wohingegen die Varianz der mittleren Performance der Modelle relativ klein ist. Da die Performance-Werte pro Task mit einer Reihe von Fehlern behaftet sind (z. B. durch inkonsistente Bewertungen oder Reihungseffekte), kann nicht verlässlich daraus geschlossen werden, dass ein bestimmtes Ranking für bestimmte Aufgaben bessere Ergebnisse liefert.

Das in absoluten Zahlen hohe Performanceniveau ist auffällig, aber kein verlässlicher Indikator für die Qualität eines Systems (Webber, Moffat, & Zobel, 2008). Es ist vielmehr ein Artefakt der Testkollektion und vor allem auf einen hohen Anteil als relevant bewerteter Dokumente zurückzuführen. Dies deutet entweder auf inhärent einfache Tasks oder eher wohlwollende Bewertungen hin. Carterette & Soboroff (2010) zeigten, dass wohlwollende Bewertungen zu einer reduzierten Varianz zwischen den Systemen führen. Strengere Relevanzbewertungen könnten demzufolge zu stärker ausgeprägten Unterschieden zwischen den Rankings führen.

Bei dem hohen Performanceniveau stellt sich außerdem die Frage, in welcher Größenordnung die Performance gesteigert werden kann. Metrikov et al. (2012) zeigen, dass allein aufgrund der Varianz der Bewertungen zwischen verschiedenen Jurorinnen und Juroren die maximal erreichbare Performance bereits deutlich reduziert wird. Diese obere Schranke wird durch Unsicherheiten in den einzelnen Relevanzbewertungen sowie sogenannte Multi-Intent-Suchanfragen[18] weiter abgesenkt.

Für Multi-Intent-Suchanfragen mag ein Ranking nach Popularität implizit die Generierung von diversifizierten Ergebnislisten, die relevante Dokumente für verschiedene Informationsbedürfnisse enthalten, begünstigen. Dies kann jedoch auf der Grundlage von Relevanzbewertungen für spezifische Informationsbedürfnisse, wie sie den Jurorinnen und Juroren in den Testläufen vorgelegt wurden, nicht detektiert werden.

Da die Ergebnislisten teilweise dieselben Dokumente wie die Top-10 Treffer der Textstatistik-Baseline enthalten, könnte der Effekt der Textstatistik-Komponente im priorisierten Aggregationsoperator zu stark ausfallen. Dies mag an einer zu geringen „Kompatibilität" der Textstatistik-Score-Verteilung mit der Verteilung der anderen Faktoren liegen. Für die Kopplung dieser Verteilungen ist es daher möglicherweise hilfreich, auch die Textstatistik-Scores zu normalisieren (Nottelmann & Fuhr, 2003). Die Verwendung eines exponentiellen Gewichtungsschemas für die CSS-Transformation könnte den Effekt der anderen Faktoren ebenfalls verstärken (s. Abschnitt 3).

Letztlich hat auch die Qualität der Daten für die Faktoren einen Einfluss auf die Effektivität der Rankings. Die Daten stammen überwiegend aus externen Systemen, die in der Regel nur Teilkollektionen von EconBiz abdecken, wodurch zwangsweise Lücken und Fehler auftreten. Eine Verbesserung lässt sich durch den weiteren Ausbau offener Datenbanken durch Projekte wie die „Initiative for Open Citation (I4OC)" oder „Linked Open Citation Database (LOC-DB)" erwarten.

## 6 Fazit

Faktoren für das Relevanzranking in bibliothekarischen Informationssystemen können in Anlehnung an das Ranking in Websuchmaschinen theoretisch weit über textstatistische Verfahren hinausgehen. So können insbesondere Popularitätsindikatoren wie Zitations- und Nutzungshäufigkeiten in Verbindung mit Aktualität in das Ranking einfließen. Ausgewählte Rankingfaktoren wurden in eine EconBiz-basierte Testumgebung implementiert und mithilfe des Relevance Assessment Tools und Relevanzbewertungen durch menschliche Jurorinnen und Juroren evaluiert. Insgesamt konnte für keines der getesteten Rankings eine einheitliche Verbesserung gegenüber den Baseline-Rankings gemessen werden. Das bedeutet, gegenüber den klassischen textstatistischen Verfahren wurde kein „erfolgreicheres" Rankingmodell gefunden, das für alle Suchanfragen bzw. Informationsbedürfnisse die relevantesten Ergebnisse auf den oberen Trefferpositionen anzeigen kann. Insbesondere die hohe Varianz über die Suchaufgaben deutet darauf hin, dass eine Adaptierung des Rankings auf individuelle Nutzer bzw. Nutzungskontexte, notwendig sein könnte, um erfolgreichere Rankingmodelle zu erstellen und eine höhere Performance zu erzielen.

Obwohl sich die rein textstatistischen Verfahren als kompetitiv erwiesen haben, ist eine Offenlegung der Rankingverfahren weiterhin wünschenswert. Insbesondere im Kontext wissenschaftlicher Informationssysteme besteht ein explizites Interesse an der Transparenz und „Wissenschaftlichkeit" der Rankingmechanismen sowie an Alternativen zu denen kommerzieller Anbieter. Vor diesem Hintergrund regen wir an, dass interessierte Einrichtungen mit ihrem eigenen Datenbestand und anpassbaren Relevanzfaktoren Tests durchführen sowie die Bearbeitung weiterer Forschungsfragen verfolgen. Ein im Rahmen des

**18** Multi-Intent-Suchanfragen lassen unterschiedliche Interpretationen in Hinblick auf das entsprechende Informationsbedürfnis zu.

LibRank-Projektes entwickelter Demonstrator[19] wurde bereitgestellt, um die Auswirkungen von beliebigen Kombinationen der Faktoren interaktiv evaluieren zu können. Die Testkollektion[20] (inkl. 690 realer Suchanfragen und 35.158 Relevanzbewertungen) steht zudem im GESIS-Datorium für die Nachnutzung zur Verfügung.

## Förderhinweis

Die hier vorgestellten Erkenntnisse sind das Ergebnis des Forschungsprojektes *LibRank: Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen*, das von März 2014 bis Februar 2016 von der Deutschen Forschungsgemeinschaft (Projektnummer 246011126) gefördert wurde.

## Literatur

Barry, C., & Lardner, M. (2011). A study of first click behaviour and user interaction on the Google SERP. In J. Pokorny, V. Repa, K. Richta, W. Wojtkowski, H. Linger, C. Barry, & M. Lang (Hrsg.), Information Systems Development (S. 89–99). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4419-9790-6_7.

Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: The impact of citation counts (an empirical study). In 2009 Third International Conference on Research Challenges in Information Science (S. 439–446). IEEE. http://doi.org/10.1109/RCIS.2009.5089308.

Behnert, C. (2016). Evaluation methods within the LibRank project. Working Paper. http://www.librank.info/wp-content/uploads/2016/07/Working_paper_LibRank2016.pdf [1.12.2018]

Behnert, C., & Borst, T. (2015). Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen: Das DFG-Projekt LibRank. Bibliothek Forschung und Praxis, 39(3), 384–393. http://doi.org/10.1515/bfp-2015-0052.

Behnert, C., & Lewandowski, D. (2015). Ranking search results in library information systems — Considering ranking approaches adapted from web search engines. The Journal of Academic Librarianship, 41(6), 725–735. http://doi.org/10.1016/j.acalib.2015.07.010.

Behnert, C., & Lewandowski, D. (2017). A framework for designing retrieval effectiveness studies of library information systems using human relevance assessments. Journal of Documentation, 73(3), 509–527. http://doi.org/10.1108/JD-08-2016-0099.

Behnert, C., & Plassmeier, K. (2016). Results of evaluation runs and data analysis in the LibRank project. Working Paper. http://www.librank.info/wp-content/uploads/2016/10/AP4_Evaluierungsbericht.pdf [1.12.2018].

Blenkle, M., Ellis, R., Haake, E., & Zillmann, H. (2015). Nur die ersten Drei zählen! Optimierung der Rankingverfahren über Popularitätsfaktoren bei der Elektronischen Bibliothek Bremen (E-LIB). O-Bib, 2, 33–42. http://doi.org/10.5282/o-bib/2015H2S33-42.

Busa-Fekete, R., Szarvas, G., Élteto, T., & Kégl, B. (2012). An apple-to-apple comparison of Learning-to-rank algorithms in terms of Normalized Discounted Cumulative Gain. 20th European Conference on Artificial Intelligence (ECAI 2012): Preference Learning: Problems and Applications in AI Workshop. Ios Press.

Carterette, B., & Soboroff, I. (2010). The Effect of Assessor Errors on IR System Evaluation. Information Sciences, 539–546. http://doi.org/10.1145/1835449.1835540.

Connaway, L. S., & Dickey, T. J. (2010). The digital information seeker: Report of findings from selected OCLC, RIN and JISC user behaviour projects. Higher Education Funding Council for England (HEFCE). http://www.jisc.ac.uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf [1.12.2018].

da Costa Pereira, C., Dragoni, M., Pasi, G., Pereira, C. da C., Dragoni, M., & Pasi, G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized Information Retrieval setting. Information Processing & Management, 48(2), 340–357. http://doi.org/10.1016/j.ipm.2011.07.001.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. Journal of Information Science, 14(2), 123–127. http://doi.org/10.1177/016555158801400208.

Hennies, M., & Dressler, J. (2006). Clients information seeking behaviour: An OPAC transaction log analysis. In click 06, ALIA 2006 Biennial Conference, 19-22 September 2006. Perth, AU.

Höchstötter, N., & Koch, M. (2008). Standard parameters for searching behaviour in search engines and their empirical evaluation. Journal of Information Science, 35(1), 45–65. http://doi.org/10.1177/0165551508091311.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 42(1), 248–263.

Langenstein, A., & Maylein, L. (2009). Relevanz-Ranking im OPAC der Universitätsbibliothek Heidelberg. B.I.T. Online, 12(4), 408–413.

Lewandowski, D. (2010). Using search engine technology to improve library catalogs. Advances in Librarianship, 32, 35–54. http://doi.org/10.1108/S0065-2830(2010)0000032005.

Lewandowski, D. (2018). Suchmaschinen verstehen (2. Aufl.). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-662-56411-0.

Lewandowski, D., & Sünkler, S. (2013). Designing search engine retrieval effectiveness tests with RAT. Information Services and Use, 33(1), 53–59. http://doi.org/10.3233/ISU-130691.

Metrikov, P., Pavlu, V., & Aslam, J. A. (2012). Impact of assessor disagreement on ranking performance. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval – SIGIR '12 (S. 1091). New York, New York, USA, New York, USA: ACM Press. http://doi.org/10.1145/2348283.2348484.

Nottelmann, H., & Fuhr, N. (2003). From retrieval status values to probabilities of relevance for advanced IR applications. Informa-

**19** http://librank-demo.zbw.eu/ [30.11.2018]; der Quellcode für den Demonstrator steht ebenfalls zur Verfügung unter https://github.com/LibRank-Project/LibRank-Demonstrator-UI[30.11.2018].

**20** Die Testkollektion bestehend aus A) Suchanfragen und Beschreibungen; B) URL der Dokumente mit den dazugehörigen Relevanzbewertungen; C) Rankingfaktoren und Gewichtungen) sind abrufbar unter: https://datorium.gesis.org/xmlui/handle/10.7802/1253 [30.11.2018].

tion Retrieval, 6(3/4), 363–388. http://doi.org/10.1023/A:1026080230789.
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. http://doi.org/10.1111/j.1083-6101.2007.00351.x.
Plassmeier, K. (2016). Relevance model. Working Paper. http://www.librank.info/wp-content/uploads/2016/10/AP3_Relevanzmodell.pdf [1.12.2018].
Plassmeier, K., Borst, T., Behnert, C., & Lewandowski, D. (2015). Evaluating popularity data for relevance ranking in library information systems. In Proceedings of the 78th ASIS&T Annual Meeting (Bd. 51). https://www.asist.org/files/meetings/am15/proceedings/submissions/posters/270poster.pdf [1.12.2018].
Robertson, S. E. (2010). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval, 3(4), 333–389. http://doi.org/10.1561/1500000019.
Sakai, T. (2014). Statistical reform in information retrieval? ACM SIGIR Forum, 48(1), 3–12. http://doi.org/10.1145/2641383.2641385.
Schaer, P., & Tavakolpoursaleh, N. (2016). Popularity ranking for scientific literature using the Characteristic Scores and Scale method. Trec.Nist.Gov. http://trec.nist.gov/pubs/trec25/papers/THKoeln-GESIS-O.pdf [1.12.2018].
Schultheiß, S., Sünkler, S., & Lewandowski, D. (2018). We still trust in Google, but less than 10 years ago: An eye-tracking study. Information Research, 23(3), paper 799. http://www.informationr.net/ir/23-3/paper799.html [1.12.2018].
Surowiecki, J. (2005). Die Weisheit der Vielen: warum Gruppen klüger sind als Einzelne und wie wir das kollektive Wissen für unser wirtschaftliches, soziales und politisches Handeln nützen können (1. Aufl.). München: Bertelsmann.
Tavakolpoursaleh, N., Neumann, M., & Schaer, P. (2017). IR-Cologne at TREC 2017 OpenSearch Track: Rerunning popularity ranking experiments in a living lab. https://trec.nist.gov/pubs/trec26/papers/IR-Cologne-O.pdf [1.12.2018].
Webber, W. E. (2010). Measurement in information retrieval evaluation. University of Melbourne. http://www.williamwebber.com/research/wew-thesis-PhD.pdf [1.12.2018].
Webber, W., Moffat, A., & Zobel, J. (2008). Score standardization for inter-collection comparison of retrieval systems. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '08 (S. 51). New York, New York, USA: ACM Press. http://doi.org/10.1145/1390334.1390346.

**Christiane Behnert**
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät DMI, Department Information
Finkenau 35
22081 Hamburg
**christiane.behnert@haw-hamburg.de**

Christiane Behnert ist seit 2014 Wissenschaftliche Mitarbeiterin am Department Information der HAW Hamburg und beschäftigt sich seitdem mit dem Thema Relevanz. Seit 2016 arbeitet sie im Rahmen einer dreijährigen Promotionsförderung in Kooperation mit der Universität Hildesheim an ihrer Dissertation zu subjektiven Kriterien bei der Relevanzbewertung in akademischen Informationssystemen. Sie absolvierte an der Fachhochschule Potsdam ein Studium als Diplom-Bibliothekarin und kehrte nach zwei Jahren an der Universitätsbibliothek Greifswald für ein Masterstudium an die FH Potsdam zurück, das sie 2013 erfolgreich beendete.

**Kim Plassmeier**
Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW)
Abteilung Innovative Informationssysteme und Publikationstechnologien
Neuer Jungfernstieg 21
20354 Hamburg
**k.plassmeier@zbw.eu**

Diplomphysiker Kim Plassmeier arbeitet seit 2011 als Softwareentwickler und Datenanalyst bei der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW). Seine Forschungsinteressen sind Data Science und Information Retrieval.

**Dr. Timo Borst**
Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW)
Abteilung Innovative Informationssysteme und Publikationstechnologien
Düsternbrooker Weg 120
24105 Kiel
**t.borst@zbw.eu**

Timo Borst ist seit 2007 Leiter der Abteilung Innovative Informationssysteme und Publikationstechnologien bei der ZBW. Seine Forschungsinteressen sind Bibliotheksanwendungen, Information Retrieval und Repositorien für wissenschaftliche Fachinformationen. Er studierte Informatik an der Technischen Universität Berlin und wurde 1999 zum Dr. phil. an der Universität Marburg promoviert.

**Prof. Dr. Dirk Lewandowski**
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät DMI, Department Information
Finkenau 35
22081 Hamburg
**dirk.lewandowski@haw-hamburg.de**

Dirk Lewandowski ist Professor für Information Research & Information Retrieval an der HAW Hamburg. Seine Forschungsinteressen sind Web Information Retrieval, Qualitätsfaktoren von Suchmaschinen, das Rechercheverhalten der Suchmaschinen-Nutzer sowie die gesellschaftlichen Auswirkungen des Umgangs mit den Web-Suchmaschinen.