Analyse der Übertragbarkeit allgemeiner Rankingfaktoren von Web-Suchmaschinen auf Discovery-Systeme

Bachelorarbeit

Bibliothekswissenschaft

Fakultät für Informations- und Kommunikationswissenschaften

Technische Hochschule Köln

vorgelegt von:

Julia Walz

am 16.02.2018 bei Prof. Dr. Gernot Heisenberg

Technology
Arts Sciences
TH Köln

Abstrakt

Ziel: Ziel dieser Bachelorarbeit war es, die Übertragbarkeit der allgemeinen Rankingfaktoren, wie sie von Web-Suchmaschinen verwendet werden, auf Discovery-Systeme zu analysieren. Dadurch könnte das bisher hauptsächlich auf dem textuellen Abgleich zwischen Suchanfrage und Dokumenten basierende bibliothekarische Ranking verbessert werden.

Methode: Hierfür wurden Faktoren aus den Gruppen Popularität, Aktualität, Lokalität, Technische Faktoren, sowie dem personalisierten Ranking diskutiert. Die entsprechenden Rankingfaktoren wurden nach ihrer Vorkommenshäufigkeit in der analysierten Literatur und der daraus abgeleiteten Wichtigkeit, ausgewählt.

Ergebnis: Von den 23 untersuchten Rankingfaktoren sind 14 (61 %) direkt vom Ranking der Web-Suchmaschinen auf das Ranking der Discovery-Systeme übertragbar. Zu diesen zählen unter anderem das Klickverhalten, das Erstellungsdatum, der Nutzerstandort, sowie die Sprache. Sechs (26%) der untersuchten Faktoren sind dagegen nicht übertragbar (z.B. Aktualisierungsfrequenz und Ladegeschwindigkeit). Die Linktopologie, die Nutzungshäufigkeit, sowie die Aktualisierungsfrequenz sind mit entsprechenden Modifikationen übertragbar.

Schlagwörter: Ranking, Web-Suchmaschine, Discovery-Systeme, Rankingfaktoren, Faktorengruppen

Abstract

Purpose: The purpose of this bachelor thesis was to analyze the transferability of the general ranking factors, as used by web search engines, to Discovery-Systems. As a result of this investigation, the library ranking, which used to be based primarily on textual matching between search query and documents, could be improved.

Method: For this purpose ranking factors from the groups popularity, freshness, locality, technical factors as well as the personalized ranking factors were discussed. The corresponding ranking factors were selected according to their frequency of occurrence in the literature analyzed and the importance derived therefrom.

Results: 23 ranking factors were examined, 14 (61%) are directly transferable from the web search engines to the discovery systems. These include for example click popularity, publication date, user location, and language. However, six (26%) of the investigated factors are not transferable (e.g., update frequency and page loading rate). The Link-based ranking, the frequency of usage, as well as the update frequency are transferable with appropriate modifications.

Keywords: ranking, web search engines, Discovery-Systems, ranking factors, factor groups

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

CSS	Characteristics Scores and Scales
d.h.	Das heist
evtl.	Eventuell
FRBR	Functional Requirements for Bibliographic Records
h-Index	Hirsch Index
IDF	Inverse Document Frequency
i.d.R.	In der Regel
JIF	Journal Impact Factor
o.ä.	Oder ähnliches
SaaS	Software as a Service
SEO	Search Engine Optimization
SERP	Search Engine Result Page
SuUB	Staats- und Universtitätsbibliothek Bremen
TF	Term Frequency
TLA	Transaction Log Analysis
u.a.	Unter anderem
URL	Uniform Resource Locator
WDF	Within-Document Frequency
WWW	World Wide Web
z.B.	Zum Beispiel

Einleitung

Die vorliegende Arbeit analysiert die Übertragbarkeit der allgemeinen Rankingfaktoren von Web-Suchmaschinen auf Bibliothekskataloge, die auf Suchmaschinentechnologie basieren, so genannte Discovery-Systeme. Konkret sollen die durch die verwendete Literatur ermittelten wichtigsten Rankingfaktoren der Gruppen Textstatistik, Popularität, Aktualität, Lokalität, technische Faktoren und Personalisierung hinsichtlich ihrer Anwendungsmöglichkeiten im bibliothekarischen Ranking diskutiert werden. Der Fokus dieser Analyse liegt auf der Anwendung von Discovery-Systemen in wissenschaftlichen Bibliotheken.

Die Rankingfaktoren der Web-Suchmaschinen ermöglichen die Sortierung der Ergebnistreffer einer Suchanfrage nach Relevanz. Dies ist im Angesicht der Größe des Web und der hohen Anzahl an Ergebnistreffern, die auf jede Suchanfrage gelistet werden, ein essentieller Beitrag zur Bewältigung der Informationsflut. Da die heutigen Bibliothekskataloge ihre Treffermengen ebenfalls aus sehr großen Datenpools generieren, sind auch deren Trefferlisten nicht mehr komplett manuell durchsehbar. Somit wird auch in den Discovery-Systemen das Relevanzranking angewandt. Dieses beruht jedoch hauptsächlich auf dem textuellen Abgleich zwischen den Begriffen der Suchanfrage und den Dokumenten der Datenbanken. Folglich können diese nur eine textstatistische Relevanzbeurteilung, jedoch keine Aussagen zur inhaltlichen Qualität der Dokumente leisten. Viele der Web-Rankingfaktoren zielen dagegen auf eine solche Inhaltsbewertung ab. Da die Relevanz zudem eine subjektive Bewertung ist, könnten Rankingfaktoren für z.B. Popularität, Lokalität oder Personalisierung die oberen Positionen der Trefferliste stärker an die Informationsbedürfnisse des Einzelnen anpassen. Folglich könnte das verstärkte Hinzufügen von Rankingfaktoren, zur qualitativen Inhaltsbewertung der Ergebnistreffer, durch die Übertragung der Web-Rankingfaktoren auf das bibliothekarische Ranking, einen deutlichen Vorteil erbringen.

Ein weiterer Vorteil, der sich durch eine solche Übertragung ergeben könnte, wäre eine Anpassung des Suchsystems an die Recherchefähigkeiten der heutigen Nutzer[1]. Diesen bereitet offensichtlich die Formulierung ihres Informationsbedürfnisses Schwierigkeiten, was zur Folge hat, dass überwiegend Suchanfragen gestellt werden, die nur aus ein bis zwei Wörtern bestehen. Zudem fehlen die Kenntnisse des Suchvokabulars, zur Verknüpfung unterschiedlicher Anfragen, wie beispielsweise die Boolschen Operatoren[2]. Somit wird lediglich die Einfache Suche genutzt, wohingegen die Erweiterte Suche kaum Beachtung findet. Gängigen Studien

[1] Aus Gründen der besseren Lesbarkeit wird in dieser Arbeit zur Bezeichnung männlicher wie weiblicher Personen das generische Maskulinum verwendet.

[2] Boolsche Operatoren dienen der Verknüpfung von zwei oder mehr Suchtermini, um komplexe Suchanfragen zu formulieren. Diese sind AND, OR und NOT.

zufolge, wird des Weiteren bei der Auswahl der relevanten Ergebnistreffer darauf vertraut, dass die tatsächlich besten Treffer als erste gelistet werden[3]. Durch das schlagkräftige Ranking der Web-Suchmaschinen, wird es möglich, tatsächlich relevante Dokumente auch ohne komplexe Formulierung der Suchanfragen auf den obersten Rängen zu finden, welche meist das Informationsbedürfnis ausreichend befriedigen. Ein erweitertes Suchformular oder die Kenntnis des Suchvokabulars wird so vermeintlich nicht mehr benötigt.

In dieser Arbeit werden zunächst in Kapitel eins die Begrifflichkeiten der Web-Suchmaschine und des Discovery-Systems geklärt und weiterführend die theoretischen Grundlagen und Ideen des Rankings und der Relevanzbewertung dargelegt. Im zweiten Kapitel werden anschließend die technischen Grundlagen in Bezug auf die Funktionsweise der Suchmaschinentechnologie in beiden Systemen, sowie des Index als Basis für das Ranking thematisiert. Darauf aufbauend folgt in Kapitel drei die Analyse der Übertragbarkeit der Web-Rankingfaktoren auf das bibliothekarische Ranking. Dieses ist in die oben genannten sechs Hauptgruppen der Rankingfaktoren untergliedert. Für jede Faktorengruppe werden die entsprechenden Rankingfaktoren vom Webkontext ausgehend beschrieben und hinsichtlich ihres Anwendungspotentials auf Discovery-Systeme untersucht. Wann immer möglich wird der bisherige Einsatz des Faktors anhand eines Bibliotheksbeispiels angegeben. Als Basis hierfür dienten vornehmlich die Angaben der vier kommerziellen Hauptanbieter für Discovery-Systeme (siehe Kapitel 1.2. Discovery-Systeme), die raren Berichte über das Ranking von durch einzelne Bibliotheken selbstentwickelten Discovery-Systemen, sowie den Projektberichten des nach zwei Jahren Laufzeit 2016 abgeschlossenen Projektes „LibRank“. Dieses beschäftigte sich mit der Integration von Rankingfaktoren aus dem Webkontext in das Fachportal der Wirtschaftswissenschaften „EconBiz“[4].

1. Theoretische Grundlagen

In diesem Kapitel sollen zunächst die theoretischen Grundlagen zum Verständnis des später diskutierten Ranking der Suchergebnisse geschaffen werden. Einführend werden die beiden in dieser Arbeit miteinander verglichenen Systeme, Web-Suchmaschinen und Discovery-Systeme, kurz vorgestellt. Ein Schwerpunkt des Kapitels liegt jedoch in den Theorien, die der Sortierung nach Relevanz und der Relevanzbewertung selbst zu Grunde liegen. Im Zusammenhang letzterer wird abschließend die Anfrageinterpretation thematisiert.

---

[3] Vgl. hierzu Bues, Johannes (2015): Klickwahrscheinlichkeiten in den Google SERPs. In: SISTRIX Blog.; Nutzerverhalten auf Google-Suchergebnisseiten. Eine Eyetracking-Studie im Auftrag des Arbeitskreises Suchmaschinen-Marketing des Bundesverbandes Digitale Wirtschaft (BVDW) e.V..

[4] Leibniz-Informationszentrum Wirtschaft. Homepage/Forschung/LibRank.

1.1. Web-Suchmaschinen

Web-Suchmaschinen, umgangssprachlich auch nur als Suchmaschinen bezeichnet, sollen das World Wide Web (WWW) systematisch durchsuchbar machen[5]. Die Herausforderung liegt dabei in dessen enormer Anzahl an enthaltenen Dokumenten und ständigen Veränderung, da kontinuierlich Dokumente hinzugefügt, gelöscht oder verändert werden. Die vielen verschiedenen Dokumenttypen und Web-Seiten innerhalb des WWW sind untereinander durch Hyperlinks verbunden. Diese Linkstrukturen ermöglichen den Suchmaschinen eine automatische Erfassung dieser miteinander vernetzten Dokumente (siehe Kapitel 2.1. Funktion von Suchmaschinen). Suchmaschinen sind demnach umfangreiche Computerprogramme und die Webseite eines Anbieters, wie z.B. www.google.de, die umgangssprachlich als „Suchmaschine" bezeichnet werden, ist lediglich die Schnittstelle zwischen Nutzer und System[6]. Die wichtigste Komponente einer Suchmaschine ist jedoch, laut Erlhofer, „das automatische Sammeln und Auswerten von Webseiten." [7]

Die Suchmaschine Google-Suche der Firma Alphabet Inc. ist das Web-Suchinstrument, das sich in den westlichen Ländern am deutlichsten durchgesetzt hat. Neben Google gibt es jedoch auch die Suchmaschine Bing von Microsoft. Auf dieser baut die drittwichtigste Suchmaschine Yahoo! auf[8]. Alle drei Genannten haben sich zur Aufgabe gemacht, das weltweite Internet zu verzeichnen, dennoch ist bei diesen eine eindeutige Bevorzugung der Sprachen aus den so genannten westlichen Ländern zu erkennen. In Sprachräumen in denen nicht das lateinische Alphabet gebräuchlich ist, sind dagegen andere Suchmaschinen populär. So gibt es beispielsweise für den chinesischen Markt die Suchmaschine Baidu.

1.2. Discovery-Systeme

Auch Discovery-Systeme sind Suchmaschinen. Sie wurden entwickelt, um die Quellen der Bibliotheken zentral durchsuchbar zu machen. Diese Quellen sind dabei ebenso heterogen, wie die des World Wilde Web. Breeding führt diese exemplarisch auf: „a comprehensive view of library collections today consists of many components: physical print and media collections, locally created digital collections, subscriptions to ejournals and databases, ebook collections,

---

5 Vgl. Erlhofer, Sebastian (2016): Suchmaschinen-Optimierung, Kpt. 5.

6 Ebd.

7 Ebd., hier: S. 198.

8 Bidder, Benjamin; Schultz, Stefan (2009): Microsoft und Yahoo verbünden sich gegen Google. In: Spiegel Online.

and selected free materials on the web such as open access scholarly journals or digital collections."[9]

Da Discovery-Systeme somit ebenfalls auf sehr großen Datenmengen basieren, ist die Notwendigkeit einer hohen Präzision des Rankings, einer damit einhergehenden guten Aufbereitung der Daten im Voraus, sowie eine leichte Bedienbarkeit für die Nutzer, wie bei Web-Suchmaschinen, gegeben[10]. Aus diesem Grund basieren Discovery-Systeme auf Suchmaschinentechnologie (vgl. Kapitel 2.2. Funktion von Discovery-Systemen).

Sie werden als Software as a Service (SaaS), d.h. nicht durch die Bibliotheken selbst, betrieben[11]. SaaS ist eine Form des cloud computering, bei dem alle Daten auf einem extern betriebenen Server gespeichert werden. Die Software und die IT-Infrastruktur werden dabei von einem Anbieter der Bibliothek als Dienstleistung zur Verfügung gestellt. Aktuell gibt es auf dem Markt vier solcher Anbieter: ProQuest mit ihrem Produkt Summon, Ex Libris Group mit ihrem Produkt Primo Central, EBSCO Industries mit EBSCO Discovery Service und OCLC mit WorldCat Local.

Es gibt jedoch auch Bibliotheken, die ihre eigenen Discovery-Systeme entwickelt haben, so zum Beispiel die UB Heidelberg, oder die UB Bremen.

1.3. Ranking

In diesem Unterkapitel sollen zunächst die Ziele des Ranking und anschließend dessen Zustandekommen und Vorteile dargelegt werden. Abschließend werden die sich aus dem Ranking ergebenden Probleme diskutiert.

Das Ranking ist eine Sortierung von Dokumenten nach einer Rangordnung, in diesem Falle nach angenommener Relevanz. Es ist ein auf komplexen Algorithmen beruhender Rechenvorgang einer Suchmaschine, der die Position der einzelnen Dokumente auf der Ergebnisseite, der so genannten SERP (Search Engine Result Page), ermittelt[12].

Ziel ist es, die Dokumente, die in einem ersten Schritt für die Suchanfrage als passend ermittelt wurden, in eine für den Suchenden sinnvolle Rangordnung zu bringen. Die „Rangordnung […] ist eine Reihenfolge mehrerer vergleichbarer Objekte, deren Sortierung eine Bewertung festlegt."[13] Die Bewertung findet Ausdruck in der Positionierung bzw. Platzierung der einzelnen

---

9 Breeding, Marshall (2012): Library Web-Scale. In: Computers in Libraries, hier: S.21.

10 Neubauer, Karl Wilhelm (2010): Die Zukunft hat schon begonnen. In: B.I.T.online, Heft 1.

11 Ebd.

12 Suchmaschinenranking. In: Wikipedia – Die freie Enzyklopädie.

13 Rangordnung. In: Wikipedia – Die freie Enzyklopädie.

Treffer im Ranking. Dem ersten Dokument in der Reihung wird dabei die höchste, dem Letzten die niedrigste Relevanz in Bezug auf die Suchanfrage beigemessen. Diese Relevanzbewertung beruht jedoch lediglich auf einer, durch die jeweilige Suchmaschine individuell, vermuteten Relevanz[14]. Aufgrund der Individualität der Algorithmen und vor allem deren Gewichtung, die von den einzelnen Suchmaschinen für das Ranking eingesetzt werden, erhält ein Suchender für dieselbe Anfrage in unterschiedlichen Suchmaschinen ebenso unterschiedliche Trefferlisten.

Die Relevanz eines Dokuments wird mithilfe von Rankingfaktoren beurteilt. In der Suchmaschinenoptimierung wird ein solcher Rankingfaktor wie folgt definiert: „Ein 'Rankingfaktor' ist ein (Multiplikations-)Faktor innerhalb des Algorithmus einer Suchmaschine, bei dessen positiver bzw. negativer Erfüllung eine Webseite in den Ergebnissen einer Suchmaschine steigt bzw. fällt.“[15] Von diesen Faktoren gibt es sehr viele, Google gibt etwa an, mehr als 200 Faktoren für sein Ranking zu verwenden[16], Microsoft verwendet in seiner Suchmaschine Bing, nach eigenen Angaben sogar mehr als 1000[17]. Trotz der hohen Anzahl der verwendeten Rankingfaktoren können diese sechs Hauptgruppen zugeordnet werden:

- **Textstatistik:** Gleicht die Suchbegriffe hinsichtlich des Vorkommens in den Dokumentenbeständen ab, und bildet die Grundlage für die weiteren fünf Rankingfaktorengruppen
- **Popularität:** Funktioniert nach dem Prinzip „was für andere bei dieser Suche hilfreich war, ist es auch jetzt“ indem z.B. Klickzahlen und Verweildauer in Bezug gesetzt werden
- **Aktualität:** Reihung u.a. nach Veröffentlichungsdatum eines Dokuments
- **Lokalität:** berücksichtigt z.B. den aktuellen Standort des Suchenden in der Ergebnisreihung
- **Technische Faktoren:** Berücksichtigt z.B. das Dateiformat oder dessen Sprache
- **Personalisierung:** Es wird durch Auswertung der Logfileanalysen u.a. das persönliche Interesse, Lese- und Suchverhalten im Ranking berücksichtigt

Diese Faktorengruppen sollen dieser Arbeit als Grundlage dienen.

Zudem können die Rankingfaktoren in anfrageabhängige Faktoren, sowie anfrageunabhängige Faktoren unterteilt werden[18]. Diejenigen, die von der Suchanfrage abhängig sind, beziehen sich auf den textuellen Vergleich der Suchbegriffe (siehe Kapitel 3.1. Textstatistik) und der Doku-

[14] Vgl. Erlhofer, Sebastian (2016), Kpt. 4.
[15] Was ist Rankingfaktor. In: SEO-united.de – Glossar.
[16] Google – Alles über die Suche. Algorithmen.
[17] Nadella, Satya (2010): New Signals in Search. In: Bing blogs
[18] Vgl. Lewandowski, Dirk (2005): Web Information Retrieval, Kpt. 6.1.

mente. Sie waren historisch gesehen, die ersten verwendeten Rankingfaktoren. Da diese jedoch für die SEO (Search Engine Optimization, engl. für Suchmaschinenoptimierung[19]) relativ leicht zu beeinflussen sind, wurden zusätzlich Faktoren entwickelt, die zur Qualitätsermittlung der zu rankenden Dokumente dienen.

Aus dieser Entwicklung wird ersichtlich, dass das Ranking zum einen als Reaktion der Suchmaschinenanbieter auf die Manipulationsversuche der SEO entstanden ist. Zum anderen ist die Weiterentwicklung des Ranking auch der Anpassung an das Nutzerverhalten geschuldet. Diese haben, wie bereits erwähnt, u.a. Schwierigkeiten sowohl ihr eigentliches Informationsbedürfnis, als auch ihre Suchanfragen gezielt an das System zu formulieren[20]. Die Suchmaschinen reagierten darauf, indem sie u.a. den Nutzern an der Spitze der SERP eine möglichst große Vielfalt an Informationen (neben den „normalen“ Dokumenten, z.B. auch Bilder, News, etc.) anbieten, die untereinander getrennt gerankt werden.

Das Ranking der Suchergebnisse ist somit notwendig, um das Informationsbedürfnis der Nutzer, trotz deren Schwierigkeiten der Definition desselben, zuverlässig zufriedenstellen zu können. Ein zweiter Grund für die Notwendigkeit einer Relevanzsortierung, liegt in der großen Menge der vorhandenen Dokumente. Da eine manuelle Durchsicht aller vorhandenen Dokumente unmöglich ist, ist eine Beschränkung auf die Relevantesten essentiell. Ziel des Ranking ist somit auch, die Bewältigung des so genannten „information-overload“ maßgeblich zu unterstützen.

Dennoch gibt es Gründe der Ergebnisreihung einer Suchmaschine zu misstrauen. Ein Hauptproblempunkt, sowohl in Websuchmaschinen, als auch in Discovery-Systemen, ist hierbei die Intransparenz der Rechenalgorithmen[21]. Zum einen sind die Rankingfaktoren, die im Einzelnen Verwendung finden, nicht bekannt. Es lässt sich folglich nicht ermitteln, welche Faktoren, in welchem Umfang und in welcher Gewichtung für ein Ranking durch die Suchmaschine berücksichtigt werden. Zum anderen werden durch die Suchmaschinenanbieter keine Angaben über die Herkunft der für das Ranking verwendeten Daten (z.B. aus Metadaten/-tags, Volltexten, etc.) gemacht. Im Web-Kontext haben diese Tatsachen nur geringen Einfluss auf die Zufriedenheit und das Vertrauen der Nutzer in die Suchmaschine, da hier der Informationsbedarf dennoch vermeintlich immer schnell gedeckt ist. Im Bibliotheks- bzw. wissenschaftlichen Kontext

[19] Ziel der SEO ist, die eigenen Webseiten so zu gestalten, dass sie vom Rankingalgorithmus als sehr relevant eingestuft und damit auf die höchsten Positionen der Ergebnisliste (SERP) platziert werden.

[20] Vgl. Lewandowski, Dirk (2016): Suchmaschinenkompetenz als Baustein der Informationskompetenz. In: Handbuch Informationskompetenz, S. 115-126.

[21] Roscher, Mieke (2014): Fachdisziplinäre Bedürfnisse in der Gestaltung von Discovery-Lösungen. Masterarbeit, Humboldt-Universität zu Berlin.

jedoch, ist es essentiell, jede Veröffentlichung zu einem Forschungsobjekt auffinden und damit den aktuellen Forschungsstand ermitteln zu können. Die Wissenschaftler müssen sich folglich auf die Relevanzbewertung der Suchmaschine absolut verlassen können. Dies kann durch die aktuell verwendeten Algorithmen zur Relevanzbewertung nicht gewährleistet werden. Dementsprechend groß ist auch das Misstrauen der Forscher gegenüber den Ergebnisseiten der Discovery-Systeme. Dies belegt eine Studie[22], die 2014 im Zuge einer Masterarbeit durchgeführt wurde.

Misstrauensgründe gegenüber den Rankingalgorithmen gibt es auch im Hinblick auf die einzelnen Faktorengruppen. Hierbei sind insbesondere die Faktoren der Gruppen Popularität (Unterdrückung weniger verbreiteter Meinungen) und der Personalisierung (Filterblasen) zu nennen. Diese werden in den Kapiteln 3.2 Popularität und 3.6. Personalisierung ausführlich diskutiert.

1.4. Relevanz und Pertinenz

Im vorangegangenen Kapitel wurde bereits festgestellt, dass die Sortierung der Ergebnisdokumente für eine Suchanfrage, d.h. das Ranking, einer Suchmaschine nach einer Bewertung der Relevanz erfolgt. Was die einzelnen Suchmaschinen jedoch unter Relevanz verstehen, ist einerseits abhängig von der Gewichtung der Rankingfaktoren im Algorithmus, andererseits vom subjektiven Empfinden der Nutzer. Im Folgenden soll dies detaillierter dargelegt werden.

So stellt Lewandowski fest, dass „es einen großen Unterschied aus [macht], ob mit Relevanz das Matching von Dokument und Suchanfrage gemeint ist oder ob man Relevanz anhand der Übereinstimmung von Information und Informationsbedürfnis messen möchte.“[23] In diesem Sinne wird zwischen Relevanz und Pertinenz unterschieden.

Die Relevanz wird aus den Informationen errechnet, die aus der Interaktion der tatsächlich eingegebenen Suchanfrage und dem System anfallen[24]. Sie zielt demnach auf die nutzerunabhängige, technisch objektiv messbare Seite.

Die Pertinenz bezeichnet das subjektive Informationsbedürfnis des Nutzers, d.h. die empfundene Nützlichkeit und Verstehbarkeit der gefundenen Information[25]. Die Bewertung der einzelnen Dokumente erfolgt durch die Nutzer i.d.R. unsystematisch und intuitiv[26]. Folglich kann

[22] Vgl. Roscher, Mieke (2014), S. 54.
[23] Lewandowski, Dirk (2005), hier: S. 96.
[24] Vgl. Lancaster, Frederick W.; Gale, V. (2003): Pertinence and Relevance. In: Encyclopedia of Library and Information Science, S. 2311.
[25] Vgl. Ebd.
[26] Vgl. Stock, Wolfgang G. (2007): Information-Retrieval, Kpt. 6.

Pertinenz vom System nicht direkt gemessen werden, ist aber für die Zufriedenheit der Systemnutzer entscheidend.

Mizzaro hat das Zusammenspiel dieser beiden Gruppen in seinem Modell 1997 dargestellt (Abbildung 1):

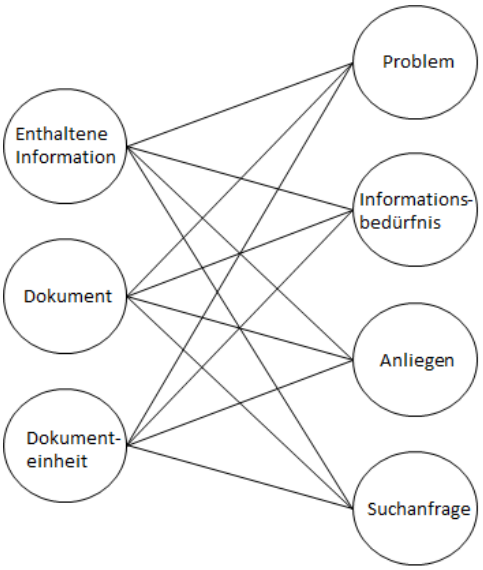


Abbildung 1 Formen der Relevanz nach Mizzaro. Die Grafik zeigt das Zusammenspiel der systemseitigen Relevanz (linke Seite) und der Nutzerseitigen (rechte Seite). Quelle: Lewandowski, Dirk (2005), S. 96, modifiziert J.Walz.

Auf der linken Seite sind die Faktoren des Systems dargestellt: die in einem Dokument enthaltenen Informationen, das Dokument, sowie die tatsächliche Repräsentation des Dokumentes, die Dokumenteinheit (Dateiformat, gedrucktes Buch o.ä.). Rechts ist die Nutzerseite dargestellt: Zunächst das Problem, das der Nutzer durch das Finden einer Information lösen will. Dieses drückt sich in einem Informationsbedürfnis aus, das der Nutzer in ein Anliegen, d.h. eine umgangssprachliche Formulierung seines Informationsbedürfnisses, übersetzt. Die Suchanfrage stellt schließlich die Übersetzung des Anliegens in die tatsächlich an das System gestellte Anfrage dar.

Jede Verbindung stellt nach Mizzaro eine Art von Relevanz dar. Die Relevanz im oben beschriebenen Sinne, d.h. die streng nutzerunabhängige Relevanz, werden durch die drei von der Suchanfrage (rechts unten) ausgehenden Verbindungen dargestellt. Wobei die Verbindung zwischen der Suchanfrage und der Dokumenteinheit die „algorithmische Relevanz“ ausdrückt. Aus dieser erfolgt das Relevanzranking basierend auf den Faktoren der Textstatistik (siehe Kapitel 3.1. Textstatistik)[27]. Die Pertinenz wird durch die Verbindungen symbolisiert, die von den nutzerseitigen Faktoren Informationsbedürfnis und Anliegen ausgehen (Mitte, rechts)[28].

Ein evidentes Problem liegt in der Möglichkeit, dass Relevanz und Pertinenz nicht übereinstimmen. So können durch die Suchmaschine Dokumente bezüglich der Suchanfrage als relevant eingestuft werden, die das Informationsbedürfnis des Nutzers dennoch nicht befriedigen. Gründe liegen in der Schwierigkeit das Informationsbedürfnis präzise zu definieren, sowie der

[27] Stock, Wolfgang G. (2007), Kpt. 19.
[28] Ebd.

Herausforderung dieses anschließend in eine systemgerechte Anfrage zu übersetzen[29] (vgl. Einleitung).

Der Faktor Zeit stellt ein weiteres Problem der Relevanz dar[30]. Eine Information kann aktuell sehr wertvoll sein, später dagegen gänzlich irrelevant. Ebenso können bereits veraltete Informationen zu einem späteren Zeitpunkt wieder relevant werden. Aus den Suchanfragen, die oft aus nur ein bis zwei Wörtern bestehen[31], werden die Nutzungskontexte und Änderungen der Informationsbedürfnisse nicht ersichtlich. Da das Ranking jedoch auf der wahrscheinlichen Relevanz beruht, ist die Ermittlung der Pertinenz, d.h. das Verständnis der Suchanfragen, für eine Suchmaschine essentiell.

1.5. Anfrageinterpretation

Um eine Reihung nach Relevanz erstellen zu können, ist es, wie im vorherigen Kapitel bereits erörtert, für die Suchmaschine notwendig, herauszufinden, was der Nutzer sich von seiner Suchanfrage erwartet, d.h. was für ihn relevant, was irrelevant sein könnte. Dies wird als Anfrageinterpretation bezeichnet[32].

Im Kontext der Web-Suchmaschinen können drei Anfragetypen bestimmt werden[33]: navigationsorientierte, informationsorientierte und transaktionsorientierte Anfragen.

Unter navigationsorientierten Anfragen wird das gezielte (Wieder-) Finden einer bereits bekannten Webseite, wie z.B. der Homepage eines Unternehmens, verstanden. Dies bedeutet, dass i.d.R. nur ein richtiges bzw. relevantes Dokument existiert und bei dessen Auffinden das Informationsbedürfnis sofort befriedigt ist.

Im Gegensatz hierzu, kann bei informationsorientierten Anfragen das Informationsbedürfnis nicht durch ein einziges Dokument befriedigt werden. Diese Anfragen zielen auf das Verständnis eines ganzen Themas, wie z.B. „Windenergie“, ab, wodurch viele verschiedene Informationen/Dokumente als relevant eingestuft werden müssen.

Transaktionsorientierte Anfragen dienen dem Auffinden einer Webseite, auf der dann eine Transaktion, wie z.B. der Kauf eines Produkts, oder der Download einer Datei, stattfinden soll. Ziel dieser Anfragen ist demnach die Interaktion mit der gesuchten Webseite.

---

29 Lewandowski, Dirk (2005).

30 Stock, Wolfgang G. (2007).

31 Vgl. Lewandowski, Dirk (2015): Suchmaschinen verstehen, Kpt. 4.5.2.

32 Ebd., Kpt. 3.5.

33 Broder, Andrei (2002): A taxonomy of web search. In: SIGIR Forum.

Dokumente die durch navigations- und transaktionsorientierte Anfragen aufgefunden werden sollen, sind, aufgrund des dem Nutzer bekannten genauen Ziels seiner Suche, für Suchmaschinen leicht zu ermitteln und dementsprechend hoch zu ranken[34]. Die Dokumente für die informationsorientierten Anfragen unterliegen jedoch der subjektiven Relevanzbewertung der Nutzer (siehe Kapitel 1.4. Relevanz und Pertinenz).

Diese Einteilung der Anfragetypen lässt sich auf den Bibliothekskontext übertragen[35]:

Den navigationsorientierten Anfragen entsprechen die so genannten Known-Item-Searches, d.h. die Suche nach einem bekannten Titel bzw. Werk. Die informationsorientierten Anfragen entsprechen der thematischen Suche. Hierbei wird weitergehend unterschieden zwischen konkretem und problemorientiertem Informationsbedarf. Ersterer bedeutet die Suche nach Fakteninformationen, d.h. mit dem Finden des Faktums, wie z.B. die Einwohnerzahl einer Stadt, ist das Informationsbedürfnis befriedigt. Letzt genannter bezeichnet die Suche nach Literatur/Informationen zu einem bestimmten Thema, was mehrere Dokumente zur Befriedigung des Informationsbedürfnisses benötigt. Diese Unterscheidung trifft ebenfalls auf die informationsorientierten Anfragen an Web-Suchmaschinen zu. Den transaktionsorientierten Anfragen steht im Bibliothekskontext die Suche nach einer Datenbank o.ä., in der die Recherche mit weiteren Anfragen fortgesetzt werden kann, gegenüber.

Für die Suchmaschine ist, aufgrund der oft nur aus wenigen Worten bestehenden Suchanfragen (siehe Kapitel 1), nicht ersichtlich, welcher Anfragetyp vorliegt. So kann beispielsweise die Suchanfrage „1. FC Köln“ navigationsorientiert sein, wenn dessen Homepage gefunden werden will, informationsorientiert, wenn sich der Suchende über den Club informieren möchte, oder transaktionsorientiert, wenn die Intention besteht, Tickets für das nächste Spiel zu kaufen. Um dennoch eine befriedigende Relevanzbewertung treffen zu können, ist die Suchmaschine gezwungen, den Anfragetyp implizit über Logfile-Analysen zu ermitteln[36]. Hierbei werden die durch das System automatisch erstellten Protokolle einer Suche (=Logfile) analysiert, wodurch die Anfragen mit Kontextinformationen angereichert werden. So wird beispielsweise auf die Suchhistorie der Vergangenheit, die Suchanfragen der aktuellen Session[37], Informationen darüber, welche Dokumente angeklickt wurden und wie lange diese besucht wurden, etc. zuge-

[34] Vgl. Lewandowski, Dirk (2016): Suchmaschinenkompetenz als Baustein der Informationskompetenz. In: Handbuch Informationskompetenz, S. 115-126.

[35] Lewandowski, Dirk (2010): Der OPAC als Suchmaschine. In: Handbuch Bibliothek 2.0.

[36] Lewandowski, Dirk (2010).

[37] Als Session wird Zeitraum bezeichnet, der entweder durch eine, durch die Suchmaschine, festgelegte Zeit oder längere Inaktivität von Nutzerseite beendet wird. Siehe Wikipedia Sitzung (Informatik).

griffen[38]. Diese Daten werden sowohl für den aktuell Suchenden erhoben, wie auch für alle Nutzer (mit derselben) Suchanfrage insgesamt. Letztere Datenverwertung erfolgt nach dem Prinzip der „Weisheit der Vielen“[39]. Dabei wird davon ausgegangen, was andere Nutzer in dieser Situation für relevant erachteten, ist auch für den aktuell Suchenden von Interesse.

Zusammenfassung:

Dieses Kapitel hat gezeigt, dass das Relevanzranking bei den großen Datenmengen, die die Suchsysteme und deren Nutzer zu bewältigen haben, essentiell ist. Ein weiterer Vorteil ist die nicht mehr benötigte Fähigkeit der komplexen Anfrageformulierung durch die Erhebung von Kontextinformationen in Form von Rankingfaktoren. Dem gegenüber stehen die Nachteile des Ranking, allen voran die Intransparenz der Algorithmen, die besonders im wissenschaftlichen Kontext einer Bibliothek von Bedeutung sind.

Ebenso wurden die Aspekte der Relevanz thematisiert, von denen die Pertinenz, d.h. die subjektive Bewertung der Suchergebnisse durch die Nutzer, der wesentlichste Aspekt darstellt. Zur Bestimmung dieser ist das Verständnis der Suchanfragen durch die Suchmaschine essentiell, was hauptsächlich durch implizite Kontextdatenerhebung realisiert wird.

2. Technische Grundlagen

Im vorangegangenen Kapitel wurden die Theorien die der Sortierung nach Relevanz und der Relevanzbewertung selbst zu Grunde liegen thematisiert. In diesem Kapitel sollen nun die technischen Prozesse, die dem Ranking zu Grunde liegen, dargelegt werden. So soll zunächst die Funktion der Web-Suchmaschinen und im Anschluss die der Discovery-Systeme umrissen werden. Abschließend wird die grundlegende Bedeutung des Index für das Relevanzranking aufgezeigt.

2.1. Funktion von Suchmaschinen

In diesem Unterkapitel soll die Funktion der Web-Suchmaschinen dargelegt werden. Die Funktion der Discovery-Systeme kann später hieran abgeleitet werden, da diese auf derselben Technologie beruhen.

[38] Vgl. Lewandowski, Dirk (2015), Kpt. 3.5.
[39] Ebd.

Abbildung 2 zeigt die schematische Funktion einer Web-Suchmaschine. Das World Wide Web das der Suchmaschine als Datengrundlage dient ist, in Anlehnung an die Struktur, wie Broder sie 2000 darstellte[40], in einen Kern-, IN- und OUT-Bereich untergliedert dargestellt.

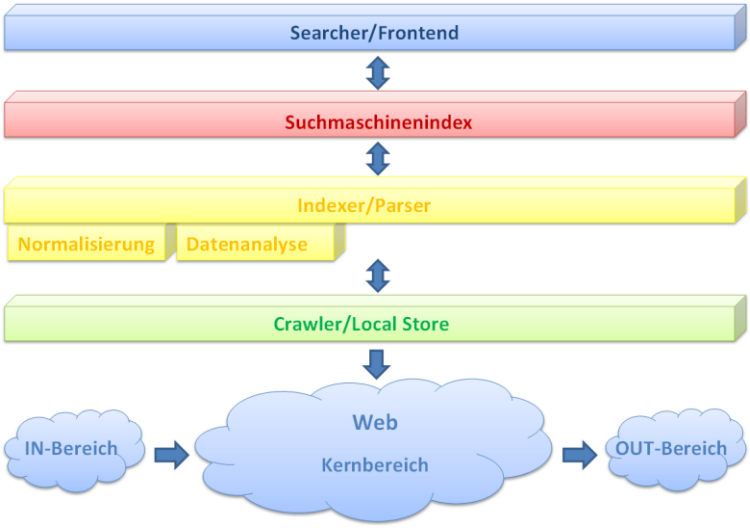


Abbildung 2 Schematische Funktion einer Web-Suchmaschine. Das Web, als Wolke dargestellt, ist in den Kernbereich, den IN-Bereich und den OUT-Bereich untergliedert. Der Crawler erstellt durch die Verfolgung der Linkstrukturen eine Kopie des Web, den Local Sore. Der Indexer erstellt daraus den Suchmaschinenindex auf den der Searcher beim Ranking zugreift.

Aufgrund der Dynamik des Webs ist dieses als Wolke dargestellt. Zum Kernbereich des Web zählen die Dokumente, die untereinander sehr stark verlinkt sind. In diesen Bereich führen die Links von Dokumenten aus dem IN-Bereich (links vom Kernbereich). Diese Dokumente verlinken Dokumente aus den anderen Bereichen, werden allerdings selbst von diesen nicht verlinkt. Die Dokumente, die vom Kernbereich und z.T. vom IN-Bereich verlinkt werden, bilden den OUT-Bereich (rechts vom Kernbereich).

Diese Linkstrukturen des Web macht sich der Crawler (in Abb. 2 grün) der Suchmaschine zu Nutze. Der Crawler ist ein Computerprogramm, mit der Aufgabe neue und aktualisierte Dokumente durch die Verfolgung der Linkstrukturen aufzufinden. Er ist, neben dem Indexer und dem Searcher, eine der Hauptkomponenten, aus denen sich die Suchmaschinen zusammensetzen. Der Crawler erstellt für die Suchmaschine eine Kopie des Webs (Local Store)[41], d.h. wird eine Suchmaschine wie etwa Google zur Recherche verwendet, wird nicht das tatsächliche Internet durchsucht, sondern die kopierte Version in den Datenbanken der Suchmaschine. Dies ist notwendig, da ansonsten, im Falle einer Suchanfrage, erst alle Dokumente des Webs gefunden werden müssten. In der Kopie der Suchmaschine können die Dokumente dagegen bereits im Vorfeld hinsichtlich der Relevanzbewertung analysiert werden (vgl. anfrageunabhängige Rankingfaktoren Kapitel 1.3. Ranking).

[40] Vgl. Lewandowski, Dirk (2015), Kpt. 6.5.
[41] Vgl. Ebd., Kpt. 3.3.

Der Crawler besteht wiederum aus vier Komponenten: Gatherer, Loader, URL-Datenbank[42] und Checker[43].

Der Gatherer ist das Element des Crawler-Systems, das tatsächlich den Links folgt und eine (neue) Kopie der gefundenen Seite erstellt, d.h. die Dokumente im Web sammelt und aktualisiert[44]. Welche Links besucht werden, wird durch die URL-Datenbank gemanagt. Findet der Gatherer ein neues Dokument, wird es sofort gedownloadet, oder die URL wird zum späteren Download an die URL-Datenbank weitergegeben. Zudem werden die weiterführenden Links des Dokuments für einen späteren Besuch des Gatherers extrahiert. Führt der Link zu keinem Dokument, wird diese Information mit einer Löschaufforderung an die URL-Datenbank weitergeleitet[45]. Theoretisch kann so, wenn alle Links besucht wurden, das gesamte Web in die Datenbanken der Suchmaschine kopiert werden. Dies ist jedoch nicht möglich, da Dokumente existieren (siehe IN-Bereich) auf die nicht durch andere Dokumente verwiesen werden[46]. Die Dokumente des Kernbereiches und damit auch die Dokumente des OUT-Bereiches können dagegen, aufgrund ihrer guten Verlinkung untereinander, leicht gefunden werden. Aus Gründen der Effizienz verfügt eine Suchmaschine über mehrere Gatherer.

Der Loader organisiert und vergibt die von den Gatherern auszuführenden Aufträge[47]. Ebenso wird durch den Loader die Ausführung der Aktualisierungs- und Erstcrawl-Aufträge überwacht und entsprechend der Auslastung der einzelnen Gatherer optimiert.

Die URL-Datenbank verwaltet die gespeicherten Links[48]. Durch sie wird die Liste der zu besuchenden Links erstellt, die über den Loader an die Gatherer weitergeleitet wird. Die URL-Datenbank ist eine relationale Datenbank, die sowohl dem System bekannte URLs mit ihren Crawlperioden, als auch neue Webadressen speichert. Die Crawlperiode, d.h. wie oft ein Dokument durch einen Gatherer besucht wird, ist u.a. abhängig von der Popularität und der Änderungsfrequenz (vgl. Nachrichtenseiten) eines Dokuments, aber auch z.B. durch die Wichtigkeit für das System selbst (wenn das Dokument viele Links enthält)[49]. Die Crawlperiode hat ebenfalls Einfluss auf das Ranking, da das Ziel einer Suchmaschine sein kann, immer sehr Aktu-

[42] URL steht für ‚Uniform Resource Locator' und stellt die Adresse einer Webseite dar.
[43] Glöggler, Michael (2003): Suchmaschinen im Internet, Kpt. 3.
[44] Ebd.
[45] Ebd.
[46] Lewandowski, Dirk (2015), Kpt. 3.1.
[47] Glöggler, Michael (2003), Kpt. 3.
[48] Ebd.
[49] Lewandowski, Dirk (2015), Kpt. 3.

elles auf die hohen Positionen zu ranken. Somit würden Dokumente mit älterem Datum eine Abwertung im Ranking erhalten[50].

Der Checker als vierte Komponente des Crawler-Systems, entscheidet was indexiert wird[51]. Er überprüft alle vom Gatherer gelieferten Dokumente und überwacht die Einhaltung der Systemvorgaben zur Erstellung des Local Store (in Abb. 2 ebenfalls grün) als Endprodukt des Crawlers. Diese Vorgaben sind beispielsweise das Dokumentformat, die technische Verfügbarkeit des Servers, Dublettenerkennung oder das Entfernen von Spam.

Der Local Store, indem alle gefundenen Dokumente lediglich nach Eingangszeitpunkt sortiert sind, liefert dem Indexer (in Abb. 2 gelb) die Rohdaten. Der Indexer ist die Komponente einer Suchmaschine, deren Aufgabe die Aufbereitung der Dokumente für das Ranking ist[52]. Dies dient zur Ermittlung der anfrageunabhängigen Rankingfaktoren (vgl. Kapitel 1.3. Ranking).

Die Datenaufbereitung erfolgt in zwei Schritten[53]: der Datennormalisierung und der Datenanalyse. Bei der Datennormalisierung werden, zur Ermöglichung einer gleichartigen Verarbeitung während des Ranking, alle Dokumente in ein einheitliches Datenformat gebracht. Ebenfalls werden durch die Entfernung des Programmiercodes, Informationen z.B. aus den Metatags extrahiert (vgl. Kapitel 3.1. Textstatistik). Bei der Datenanalyse werden die Volltexte der Dokumente aufbereitet[54]. So werden die Texte zunächst in ihre einzelnen Wörter zerlegt, es folgt u.a. die Extraktion von inhaltrepräsentierenden Wörtern (z.B. aus Überschriften), oder die Phrasenerkennung. Bei Dokumenten wie Bildern oder Videos werden diese Informationen aus den Umgebungstexten, den Metadaten, sowie direkt aus dem Dokument (Farben etc.) gewonnen.

Seine Ergebnisse speichert der Indexer in einer durchsuchbaren Datenstruktur, dem Suchmaschinenindex (in Abb. 2 rot).

Der Index ist das Herz der Suchmaschine und soll somit möglichst vollständig sein bzw. auf einer möglichst vollständigen Kopie des Web basieren[55]. Dies ist jedoch aus den in Kapitel 2.3. Bedeutung des Suchmaschinenindexes aufgeführten Gründen nicht erreichbar.

---

50 Glöggler, Michael (2003), Kpt. 3.
51 Ebd.
52 Lewandowski, Dirk (2015), Kpt. 3.
53 Erlhofer, Sebastian (2016), Kpt. 5.4.
54 Ebd.
55 Lewandowski, Dirk (2013): Suchmaschinenindices. In: Handbuch Internet-Suchmaschinen, Bd. 3, S. 143-162.

Der Searcher (in Abb. 2 blau) ist die Komponente einer Suchmaschine mit der der Nutzer interagiert[56]. Seine Aufgabe ist es die eingegebene Suchanfrage zu interpretieren und in diesem Zuge diese auch mit Kontextinformationen anzureichern. Ebenso wird durch den Searcher auf Grundlage des Indexes das Ranking erstellt[57].

2.2. Funktion von Discovery-Systemen

Dyscovery-Systeme sollten gegenüber den alten Bibliothekskatalogen im Wesentlichen eine Verbesserung in der Bedienung durch die Nutzer sein[58]. Hierfür sollten sie die verschiedenen Datenquellen der Bibliotheken, die zuvor alle separat angesteuert werden mussten, unter einer Suchmaske vereinen. Ebenso sollte die Suche intuitiver und einfacher, d.h. google-like mit Einschlitzsuche und ohne Notwendigkeit einer komplexen Suchanfrageformulierung, gestaltet werden[59].

Zur Erreichung dieses Zweckes zeigt die Abbildung 3 die Anwendung der Suchmaschinentechnologie in Discovery-Systemen in ihren Grundfunktionen.

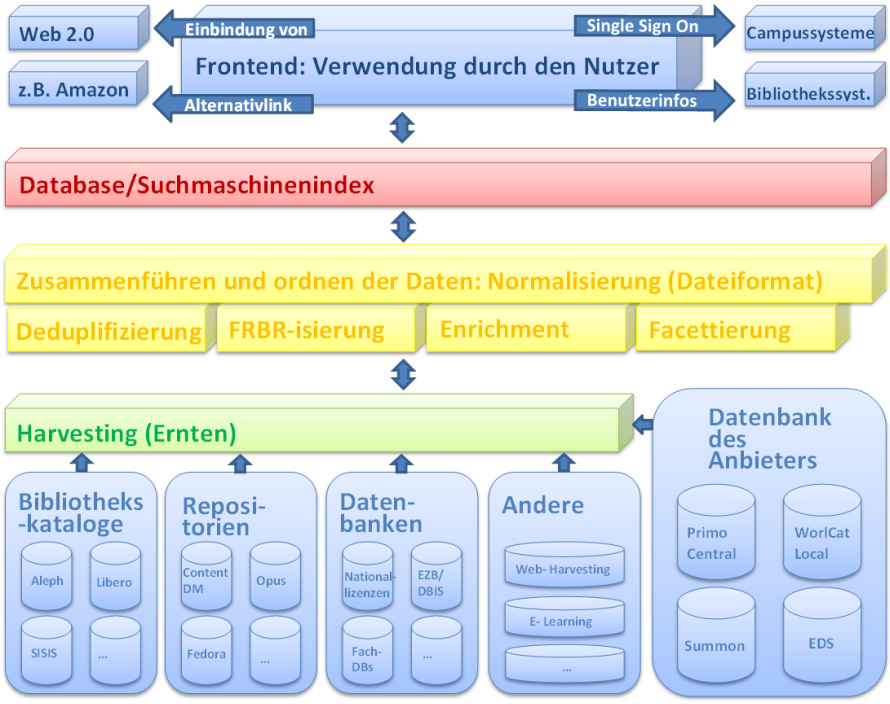


Abbildung 3 Suchmaschinentechnologie in Discovery-Systemen. Der Harvester sammelt die Dokumente in den Datenbanken, die in einem weiteren Schritt zusammengeführt und normalisiert werden, woraufhin der Suchmaschinenindex erstellt wird. Dieser bildet, wie bei den Web-Suchmaschinen, die Basis für den Searcher, der die Schnittstelle zum Frontend darstellt. Quelle: Neubauer, Karl Wilhelm (2010), modifiziert J. Walz.

---

56 Lewandowski, Dirk (2015), Kpt. 3.5.

57 Ebd.

58 Neubauer, Karl Wilhelm (2010): Die Zukunft hat schon begonnen. In: B.I.T.online, Heft 1.

59 Kaluza, Harald (2013): Google Scholar versus EBSCO Discovery Service. In: B.I.T.online – Innovativ Bd. 44, S. 59-80.

In der Grafik sind, synonym zum WWW bei den Web- Suchmaschinen, die heterogenen Datenquellen blau dargestellt, die das Discovery-System vereinen soll. Dies sind zum einen, die lokal von der Bibliothek lizensierten bzw. erstellten Daten, d.h. das lokale Bibliothekssystem, verschiedene Repositorien, kommerzielle und freie Fachdatenbanken, sowie unter „Andere", ergänzende Datenquellen aus dem Web oder digitale E-Learning Tutorials. Zum anderen ist dies die in der Regel sehr viel größere Datenbank des SaaS-Anbieters. Dabei stellen die Tonnensymbole exemplarisch jeweils eine Datenbank dar. Ein Programm, der so genannte Harvester (in Abb. 3 grün), sammelt aus allen Datenbanken in regelmäßigen Abständen neue bzw. aktualisierte Datensätze ein und speichert diese für die Normalisierung (Abb. 3 gelb) auf dem zentralen Server des Discovery-Systems. Dieser Vorgang entspricht dem Crawling-Prozess der Web-Suchmaschinen (vgl. Kapitel 2.1. Funktion von Suchmaschinen); die Harvester haben eine ähnliche Funktionalität. Die Datensätze werden anschließend in einem ersten Schritt der Normalisierung in ein einheitliches Dateiformat umgewandelt[60]. Im Falle von Primo ist dies beispielsweise das PNX-Format (PNX = Primo Normalized XML)[61]. Als ein weiterer Schritt erfolgt die Deduplifizierung, d.h. ein Abgleich mit den bereits vorhandenen Datensätzen und ggf. die Löschung der Dublette. Die bibliothekarische Datensatzbeschreibung, die Katalogisierung, basiert grundlegend auf den Functional Requirements for Bibliographic Records (FRBR). Dieses Modell unterteilt eine Publikation u.a. in die Hauptebenen der geistigen Idee eines Werkes, der Manifestationen im Sinne von Chargen und der individuellen Exemplare[62]. Gemäß dieser Unterteilung erhalten die Datensätze bei der FRBR-isierung, als weiteren Schritt der Normalisierung, eine entsprechende Markierung von beispielsweise verschiedener Auflagen oder Erscheinungsformen. Auf diese Weise können bei der späteren Trefferanzeige die anderen FRBR-Ebenen mit angezeigt bzw. verlinkt werden. Ebenso erfolgt eine Zuordnung der Datensätze zu den entsprechenden Facetten (Rubriken), sowie die Anreicherung mit Inhaltsverzeichnissen, Covern, Klappentexten, etc.

Nach der Normalisierung erfolgt die Indexierung der Datensätze. Indexiert werden dabei u.a. die Metadaten, Abstracts, Inhaltsverzeichnisse und ggf. Volltexte. Auf diese Weise entsteht, wie bei den Web-Suchmaschinen, durch den Indexer der Gesamtindex (in Abb. 3 rot). In diesem sind alle Datensätze aus den Datenbanken der Bibliothek, inklusive der des Anbieters vertreten.

---

60 Neubauer, Karl Wilhelm (2010).

61 Zumstein, Philipp (2011): Suchmöglichkeiten in Primo auf dem Prüfstand.

62 Für einen detaillierteren Einblick siehe z.B. Tillett, Barbara (2003): What is FRBR? In: Technicalities 25/Heft 5.

Durchgeführt wird die Suche durch den Nutzer ausgehend vom Frontend, in Abb. 3 oben in blau dargestellt. Auch hier erstellt der Searcher, unter Zuhilfenahme des Indexes, das Ranking der Suchergebnisse.

Bei der Trefferanzeige ist es zudem möglich Daten aus dem lokalen Bibliothekssystem, wie Verfügbarkeitsstatus oder Nutzerdaten anzuzeigen, oder eine alternative Verlinkung zu z.B. Amazon oder Google-Books oder zu Web 2.0 Angeboten zu implementieren(siehe Abb. 3 obere Ecken). Ebenso sind, falls vorhanden, die Volltexte der Dokumente über einen Link verfügbar[63]. Von der Bibliothek lizensierte, kommerzielle Datenbanken können somit durch die Implementierung eines Single-Sign-On-Mechanismus, d.h. mit einmaliger Authentifizierung, genutzt werden.

2.3. Bedeutung des Suchmaschinenindexes

Der Suchmaschinenindex spielt, wie schon erwähnt, für das Ranking eine zentrale Rolle. In den vorangegangenen Unterkapiteln (Funktion von Web-Suchmaschinen und Dyscovery-Systemen) wurde aufgezeigt, wie der Index zustande kommt. In diesem Unterkapitel soll nun der Index, als Basis für die Relevanzbewertung betrachtet werden.

Der Index ist die Repräsentation der Dokumente in Form einer invertierten Liste, bestehend aus allen der Suchmaschine bekannten Wörtern. Unter invertiert wird dabei die zu den einzelnen Wörtern zusätzlich gegebenen Positionsangaben, d.h. in welchem Dokument sie vorkommen, verstanden[64]. Somit wird es möglich von den Wörtern auf die Dokumente zu schließen, ein Durchsuchen aller Dokumente im Moment der Suche ist folglich nicht nötig. Dies ermöglicht der Suchmaschine, eine effiziente Eingrenzung der Dokumentenmenge. Für das darauffolgende Ranking hält der Index zudem alle im Voraus der Suche ermittelbaren Relevanzindikatoren bereit, um auch hier die Bearbeitungszeit zu minimieren[65].

Diese Relevanzindikatoren sind beispielsweise das Gesamtvorkommen des Wortes im Dokument, der Rang eines Begriffes, d.h. vorkommen in einer Überschrift o.ä., oder die Formatierung um eventuelle Hervorhebungen zu ermitteln[66]. Aus den Eigenschaften der Begriffe wird die Relevanz eines Dokumentes in Bezug auf die Suchanfrage errechnet. Der Relevanzbewertung bzw. dem Ranking liegt folglich die Gewichtung einzelner Begriffe zu Grunde.

---

63 Roscher, Mieke (2014).

64 Lewandowski, Dirk (2015), Kpt. 3.4.

65 Erlhofer, Sebastian (2016), Kpt. 5.5.3.

66 Ebd.

Der Index hat somit einerseits die Aufgabe das Vorkommen der (Such-) Begriffe in den jeweiligen Dokumenten zu vermerken[67]. Dadurch ist es dem Searcher, durch einen einfachen Textabgleich zwischen Suchanfrage und den im Index geführten Begriffen, möglich, die Treffermenge zunächst einzuschränken. Andererseits ist die Aufgabe des Indexes anfrageunabhängige Indikatoren zur Relevanz einzelner Begriffe zu speichern[68]. Dies ermöglicht dem Searcher eine schnelle Relevanzbewertung der infrage kommenden Dokumente, die dann lediglich noch durch die abfrageabhängigen Faktoren ergänzt werden muss.

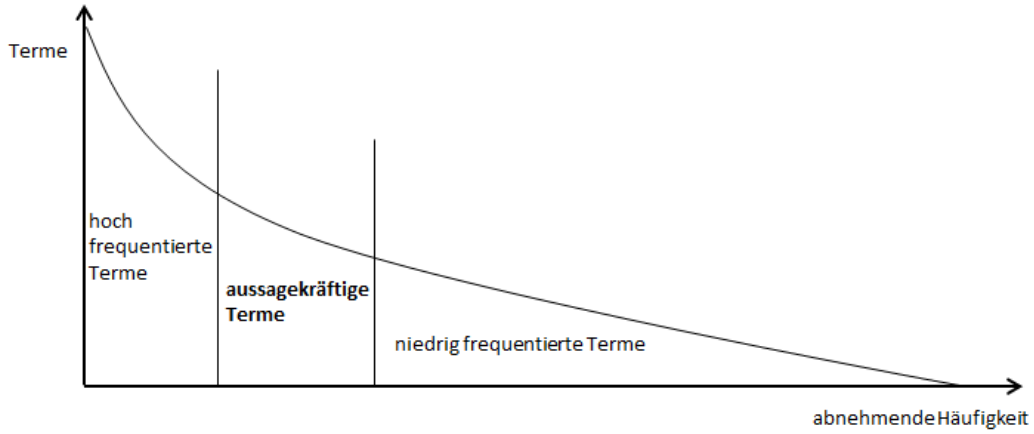
Wie bereits in Kapitel 2.1. Funktion von Suchmaschinen angesprochen, sind die Suchmaschinenanbieter im Web stets um Vollständigkeit ihres Indexes bemüht. Diese wird allerdings schon durch die Struktur des Webs torpediert. So können auf Grund der Verlinkungsstrukturen Dokumente aus dem Kern- und OUT-Bereich durch die Crawler leicht gefunden werden, Dokumente aus dem IN-Bereich jedoch nicht, da diese von keinem anderen Dokument verlinkt sind (siehe Kapitel 2.1. Funktion von Suchmaschinen). Ebenso ist es den Suchmaschinen nicht möglich Webseiten zu indexieren, deren Inhalte entweder durch deren Betreiber vor dem Zugriff der Crawler geschützt sind, wie beispielsweise Social-Media- und E-Mail-Accounts, oder deren Inhalte erst beim Aufrufen erzeugt werden, wie Zugverbindungen o.ä.[69]. Zudem erfolgt ein willentlicher Ausschluss von Dokumenten durch die Anbieter, wie z.B. Spam oder auf Grund von Urheberrechten[70]. Unter Spam werden dabei Dokumente verstanden, die mit dem Ziel erstellt werden, eine hohe Platzierung im Ranking zu erhalten und Schadsoftware zu verbreiten oder die Ergebnisqualität der Suchergebnisliste zu minimieren. Folglich beinhaltet der Index einer Web-Suchmaschine niemals das gesamte Internet.

Im Bibliothekskontext existieren solche Probleme nicht, da hier die Daten bzw. Dokumente ausnahmslos aus geschlossenen Datenbanken und aus exakt definierten, personell verifizierten Quellen stammen[71]. Hier stellt sich bezüglich der Vollständigkeit des Indexes jedoch die Frage der Führung von Stoppwortlisten. Dies sind Listen der Wörter, die von der Indexierung ausgeschlossen werden, i.d.R. die, die in den Dokumenten in hoher Frequenz auftreten, wie z.B. Artikel[72]. Dies wird von einigen Discovery-Systemen, darunter der Katalog der TH-Köln, praktiziert, da Luhn bereits in den 1950er feststellte, dass ein Zusammenhang zwischen dem Inhalt

---

67 Lewandowski, Dirk (2015), Kpt. 3.4.
68 Ebd.
69 Ebd., Kpt. 3.3.
70 Ebd.
71 Roscher, Mieke (2014).
72 Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012): Informationserschließung und automatisches Indexieren, Kpt. 5.1.5.

eines Dokuments und der Häufigkeit der Wörter in dessen Text besteht[73]. Luhn verdeutlicht diese These mittels eines Schaubilds (siehe Abbildung 4), das die Terme eines Dokuments und ihre Häufigkeit abbildet.



<u>Abbildung 4</u> Häufigkeitsverteilung von Begriffen innerhalb eines Textes. Hochfrequente Begriffe (links) sind z.B. Artikel und Präpositionen. Niedrig frequentierte Begriffe sind dagegen zu speziell, als dass ein Suchender sie kennen könnte. Im Schaubild mittig befinden sich die aussagekräftigen Terme, die den Text repräsentieren können. Quelle: Gödert, Winfried [u.a.] (2012), S. 257, modifiziert J. Walz.

Die hochfrequentierten Begriffe, in der Grafik links, identifiziert Luhn als Artikel, Pronomen, Präpositionen, Adverbien, etc.. Die niederfrequentierten Wörter (in Abb. 4 rechts) dagegen, benennt er als für den Text sehr spezifische Begriffe, die seinen Inhalt sehr gut repräsentieren. Jedoch sei es aufgrund jener hohen Spezifität unwahrscheinlich, dass der Suchende diese kennen und sie als Suchbegriffe verwendet würde[74]. Folglich sind sowohl die hoch- als auch die niederfrequenten Terme für eine inhaltliche Repräsentation des Dokuments nicht geeignet. Die entscheidungsstarken Terme dagegen, ermittelt Luhn in der Mitte der Grafik. Die mittlere Auftrittshäufigkeit dieser Wörter innerhalb des Dokuments würde einen Rückschluss auf den Textinhalt zulassen[75].

Somit können anhand der Frequenz, in der ein durch den Nutzer eingegebener Suchbegriff, in den Dokumenten einer Datenbankkollektion vorkommt, Rückschlüsse auf die Relevanz eines Dokuments gezogen werden.

Der Ausschluss von Termen aus dem Index hat jedoch einige Nachteile: Gödert [u.a.][76] führen hier an, dass Artikel auch Akronyme sein könnten, so ist beispielsweise der Artikel „DIE“ die Abkürzung des „Deutschen Instituts für Entwicklungspolitik“. Aber auch spezifische Eigennamen von Firmen, wie z.B. DER Touristik, können hochfrequente Terme enthalten, die in sol-

[73] Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012).
[74] Vgl. Ebd., S. 258.
[75] Ebd., Kpt. 5.1.5.
[76] Ebd., S. 260.

chen Fällen jedoch aussagekräftige Begriffe darstellen. Gödert [u.a.] verweisen zudem auf ein Problem der Stoppwortlisten das sich bei multilingualen Datenbankkollektionen ergibt. Wörter die im Deutschen ein hochfrequentes Vorkommen haben, könnten sich, bei gleicher Orthographie, in anderen Sprachen unter den entscheidungsstarken Termen befinden (vergl. z.B. den deutschen Artikel „die“ mit dem englischen Verb „die“). So würde eine „sprachunabhängige Stoppwortliste […] den Indexeintrag potentiell entscheidungsstarker Terme verhindern“[77]. Die Bedingungen für sprachspezifische Stoppwortlisten wäre allerdings, dass die Sprache für jedes zu indizierende Dokument jeder Zeit bekannt sein müsste und sich in der Kollektion keine bilingualen Dokumente befinden. Dies lässt sich jedoch kaum erfüllen. Gödert [u.a.] raten aus diesen Gründen von einer Führung von Stoppwortlisten ab und argumentieren unterstützend, dass die Websuchmaschinen dies ebenfalls nicht tun, was durch eine Suche des Artikels „die“ in Google, Ask und Yahoo, sowie vergleichend der Deutschen Nationalbibliothek und des Gemeinsamen Bibliothekverbundes (GBV) belegt wird[78].

Zusammenfassung:

Das Kapitel hat gezeigt, dass der Suchmaschinenindex durch die zusätzlich hinterlegten Relevanzindikatoren für das Relevanzranking eine zentrale Rolle spielt. Er beschleunigt auf diese Weise, in Anbetracht der riesigen Datenmengen, die Bearbeitungsdauer einer Suchanfrage durch den Searcher wesentlich. Der Index ist das Resultat eines komplizierten Prozesses aus Crawling bzw. Harvesting, Normalisierung, sowie Datenanalysen. Im Webkontext ist die angestrebte Vollständigkeit des Indexes nicht machbar, wohingegen sich im Bibliothekskontext die Frage nach dem Führen von Stoppwortlisten stellt, die die Bibliotheken bzw. Anbieter der Discovery-Systeme jeweils für sich beantworten müssen.

3. Analyse der Übertragbarkeit der Rankingfaktoren

In den vorangegangenen Kapiteln wurden die theoretischen und technischen Grundlagen des Relevanzranking ausführlich dargelegt. In diesem Kapitel sollen nun die Rankingfaktoren der Web-Suchmaschinen hinsichtlich ihrer Übertragbarkeit auf Discovery-Systeme erörtert werden.

Wie in Kapitel 1.3. Rankingbereits thematisiert, ist das Ranking von der Gewichtung der einzelnen Faktoren in den Algorithmen abhängig. Diese sind jedoch ein gut gehütetes Betriebsgeheimnis der Suchmaschinenanbieter. Es ist daher nicht bekannt, welche tatsächlichen Rankingfaktoren und in welcher Gewichtung diese Anwendung finden. Dennoch konnten über die Jah-

[77] Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012).
[78] Ebd.

re ausführliche Retrievaltests der SEO und anderer Einflussgruppen ermittelt werden, die für die Relevanzgewichtung definitiv eine Rolle spielen[79]. Diese sind, wie bereits in Kapitel 1.3 genannt, Textstatistik, Popularität, Personalisierung, Aktualität, Lokalität und technische Faktoren. In diese lassen sich einzelne, ebenso durch diese Tests bestimmte, Faktoren einordnen[80].

Die Textstatistik ist dabei die Faktorengruppe, die allen anderen zu Grunde liegt (siehe Abbildung 5). Durch sie wird mittels eines einfachen Textvergleichs zwischen Suchanfrage und Dokumenten zunächst die Treffermenge eingeschränkt. Alle Faktoren der weiteren Gruppen dienen weiterführend lediglich dazu, diese Treffermenge nach Qualität zu sortieren[81].

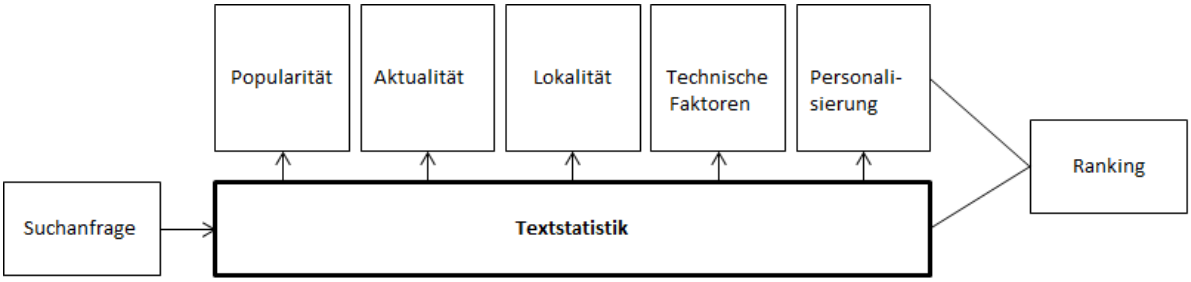


Abbildung 5 Textstatistik als Basis der anderen Rankingfaktorengruppen. Durch den textuellen Abgleich der Suchanfrage und der Dokumente, bildet die Textstatistik eine Teilergebnismenge. Diese bildet die Grundlage für die weitere Relevanzbewertung. Quelle: Lewandowski, Dirk (2015), S. 94, modifiziert J. Walz.

Dieses Kapitel wurde nach diesen Einflussgruppen gegliedert. Für jede einzelne Gruppe werden die, durch die Literatur ermittelten, wichtigsten Rankingfaktoren in ihrer Funktion und Anwendung, stets ausgehend von den Web-Suchmaschinen, mit der in den Discovery-Systemen verglichen und anschließend ihr Übertragbarkeitspotential diskutiert.

3.1. Textstatistik

Wie bereits mehrfach erwähnt, wird durch die Textstatistik die Grundlage für die weitere Relevanzbewertung gelegt. Es erfolgt zunächst ein textueller Abgleich zwischen der Suchanfrage und allen Dokumenten bzw. deren Repräsentation, dem Index (Kapitel 2.3. Bedeutung des Suchmaschinenindexes). Die Menge der Dokumente, in denen die Suchbegriffe vorkommen, bilden die Treffermenge. Alle weiteren Relevanzbestimmungen erfolgen lediglich auf dieser kleineren Menge[82]. In diesem Kapitel werden nun folgende Rankingfaktoren diskutiert, die auf der Textstatistik basieren:

[79] Lewandowski, Dirk (2015), Kpt. 5.1.
[80] Ebd.
[81] Lewandowski, Dirk (2015), Kpt. 5.2.
[82] Lewandowski, Dirk (2015), Kpt. 5.2.

- Häufigkeiten: Termhäufigkeit und Dokumenthäufigkeit
- Reihenfolge und Entfernung
- Position
- Längeninformationen
- Hervorhebungen & Ankertext

3.1.1. Häufigkeiten: Termhäufigkeit und Dokumenthäufigkeit

Die Rankingfaktoren Termhäufigkeit und Dokumenthäufigkeit befolgen das Prinzip, nach welchem das Dokument, das den Suchbegriff am meisten enthält, am relevantesten ist. Sie basieren auf der von Luhn gestellten These, dass von der Häufigkeit in der ein (Such-) Begriff in einem Dokument vorkommt, auf die Relevanz des Dokuments geschlossen werden kann (vgl. Kapitel 2.2. Funktion von Discovery-Systemen).

Termhäufigkeit:

Die einfache Termhäufigkeit TF (term frequency), d.h. das Vorkommen eines Begriffs in einem Dokument, wird gemäß Luhn gezählt[83]. Dies verschafft jedoch längeren Dokumenten, kürzeren gegenüber einen Vorteil, da ein Begriff in langen Dokumenten zwangsläufig öfters vorkommt als in kurzen. Daher sollte diese Zahl ins Verhältnis zu der Gesamtzahl aller Wörter eines Dokuments gesetzt werden[84]. Diese relativierte Worthäufigkeit (within-document frequency = WDF) errechnet sich:

WDF = Term je Dokument / Gesamtzahl der Terme je Dokument

Für die Relevanz Sortierung bedeutet dies, dass die Dokumente, die den höchsten Wert für TF bzw. WDF aufweisen, an die Spitze der Trefferliste gesetzt werden. Ein Problem hierbei besteht jedoch darin, dass die Hochfrequenzterme, sofern sie nicht durch Stoppwortlisten ausgeschlossen wurden, ebenso hohe TF- bzw. WDF-Werte generieren[85]. Das System ist dabei nicht in der Lage zu erkennen, dass diese Begriffe nicht nur im aktuell zu bewertenden Dokument häufig vorkommen, sondern in allen Dokumenten der Kollektion.

[83] Stock, Wolfgang G. (2007), Kpt. 19.
[84] Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012), Kpt. 5.3.5.
[85] Ebd., Kpt. 5.3.5.

Dokumenthäufigkeit:

Aus diesem Grund, wird zusätzlich der Wert der Dokumenthäufigkeit (Inverse document frequency = IDF) errechnet. Dieser gibt die Häufigkeit des Terms bezogen auf dessen Frequenz im gesamten Dokumentbestand an[86], welche direkt aus den Fundstellenangaben des Index entnommen werden kann (siehe Kapitel 2.3. Bedeutung des Suchmaschinenindexes). Der Wert ist am größten, wenn der Term in vielen Dokumenten des Gesamtbestandes vorkommt, wie beispielsweise hochfrequente Terme, wie Artikel und Präpositionen. Somit stellt die Dokumenthäufigkeit einen senkenden Gewichtungswert für das Ranking dar.

Den Rankingfaktoren Term- und Dokumenthäufigkeit liegen folglich nachstehende Annahmen zu Grunde[87]:

- „Je häufiger ein Term in einem Dokument vorkommt, desto wichtiger ist er für dessen Inhalt“
- „Je häufiger ein Term in der gesamten Dokumentkollektion vorkommt, desto weniger ist er als Indexterm geeignet“

Beide Faktoren werden im TF*IDF Ansatz kombiniert. Dieser garantiert die höhere Gewichtung von seltenen Begriffen, gegenüber den Stoppwörtern[88]. Die Formel des TF*IDF Ansatzes

$$TF * \mathrm{IDF} = \frac{TF\ bzw. WDF}{Dokumenth\ddot{a}ufigkeit}$$

findet in komplexerer Version sowohl im Ranking der Web-Suchmaschinen, als auch in Discovery-Systemen Anwendung[89].

Die Termhäufigkeit wird im Webkontext auch dazu verwendet, die Authentizität und damit in gewissem Maße, die Qualität der Webdokumente zu bestimmen. Mit der Messung der maximalen Keyword-Dichte (keyword density) werden Texte im Ranking stark abgewertet, die einen Begriff zu oft enthalten und damit als unnatürlich eingestuft wurden[90]. Dies resultiert aus der Erfahrung, dass Webseitenersteller in der Vergangenheit bestimmte Schlagworte, für den Nutzer unsichtbar, in großer Zahl aneinanderreihten, um ihren Seiten so eine höhere Position

[86] Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012).
[87] Ebd., hier: S. 291.
[88] Behnert, Christiane; Borst, Timo (2015): Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen. In: Bibliothek – Forschung und Praxis, 39/Heft 3, S. 384-393.
[89] Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012); Stock, Wolfgang G. (2007); Erlhofer, Sebastian (2016).
[90] Lewandowski, Dirk (2015), Kpt. 5.2.2.

zu verschaffen[91]. Im Bibliothekskontext entfällt diese Funktion der Termhäufigkeitsberechnung, da hier die Dokumente, die in den Datenbankbestand übernommen werden, nicht durch Jedermann erstellt werden können und somit i.d.R. peer reviewed, d.h. durch eine Redaktion von Fachleuten verifiziert sind.

3.1.2. Reihenfolge und Entfernung

Die Reihenfolge in der die Suchbegriffe durch den Nutzer eingegeben wurden und deren Distanz bei ihrem Auftreten in einem Dokument stellen zwei weitere Rankingfaktoren der Textstatistik dar. Beide Faktoren sind anfrageabhängig, d.h. sie können erst nach der Eingabe der Suchanfrage durch einen Abgleich der eingegebenen Begriffe mit der Dokumentenkollektion bestimmt werden.

Reihenfolge der Begriffe:

Der Rankingfaktor „Reihenfolge“ findet im Falle einer Eingabe von mehreren Suchbegriffen durch den Nutzer Anwendung. Hierbei werden einerseits die zuerst stehenden Suchbegriffe als signifikanter angenommen und erhalten dementsprechend eine größere Bedeutung in der weiteren Relevanzbewertung[92] die z.B. durch Position oder Hervorhebungen ermittelt wird (siehe Kapitel 3.1.3. Position3.1.5. Hervorhebungen und Ankertext). Andererseits werden Dokumente höher gewichtet, die die Suchbegriffe in exakt der Reihenfolge enthalten, in der sie durch den Nutzer eingegeben wurden. ExLibris gibt zu diesem Rankingfaktor an: „Primo accords greater weight to an item […] if the order of the query terms or phrases is the same in the query and the record“[93]. Lewandowski nennt hier zur Verdeutlichung das Suchbeispiel von „Paris Hilton“ (Person) bzw. „Hilton Paris“ (Hilton Hotel in Paris)[94].

Entfernung der Suchbegriffe:

Auch die Ermittlung der Entfernung der Begriffe steht in unmittelbarem Zusammenhang mit der Eingabe mehrerer Suchbegriffe. Dem zu Grunde liegt die Annahme, dass je näher die eingegebenen Suchbegriffe in einem Dokument beieinander stehen, desto höher die Wahrscheinlichkeit ist, dass das gesuchte Thema in diesem Dokument behandelt wird[95]. Lewandowski[96] nennt hier als Extrembeispiel die zwingende Nähe von Vor- und Nachnamen einer gesuchten Person. Im Ranking werden folglich Dokumente, in denen die Suchbegriffe weit auseinander

---

91 Erlhofer, Sebastian (2016), Kpt. 3.
92 Lewandowski, Dirk (2005), Kpt. 6.1.
93 Ex Libris Ltd. (2015): Primo Discovery. Search, Ranking, and Beyond.
94 Lewandowski, Dirk (2015), S. 97.
95 Vgl. Lewandowski, Dirk (2005), Kpt. 6.1.
96 Lewandowski, Dirk (2015), S. 97.

liegen abgewertet, wohingegen Dokumente, die die gesuchten Begriffe nah beieinander beinhalten, aufgewertet werden.

Sowohl die Reihenfolge, als auch die Entfernung der Begriffe voneinander werden in diesem Sinne in Web-Suchmaschinen und in Discovery-Systemen in den Rankingalgorithmen berücksichtigt[97].

3.1.3. Position

Die Position der Begriffe innerhalb eines Dokumentes ist ein weiterer Rankingparameter der Faktorengruppe Textstatistik. Hierbei erhalten Dokumente, die die gesuchten Begriffe an markanten Stellen enthalten, eine Aufwertung im Ranking[98]. Diese besonderen Dokumentstellen können beispielsweise der Titel bzw. eine Überschrift, ein Abstract, die URL, etc. sein. Steht der Suchbegriff z.B. in einer Überschrift oder gar im Titel des Dokuments, wird angenommen, dass der Begriff im nachfolgenden Text thematisiert wird, was die Relevanz des Dokuments steigert[99]. Hierbei werden allerdings die unterschiedlichen Positionen auch verschiede Gewichtungen beigemessen: „So wird beispielsweise das Titelfeld höher gewichtet als das Abstract und dieses wiederum höher als der eigentliche Text.“[100]

Die Metatags im header[101] eines HTML-Dokuments können auch als eine solche eminente Stelle angesehen werden. Diese Meta-Elemente ermöglichen dem Ersteller seinem Dokument eine inhaltliche Beschreibung, einen Titel, Schlagwörter o.ä. zuzuteilen. Suchmaschinen können auf diese hinterlegten Informationen zugreifen und diese nutzen. Wie Lewandowski jedoch schon 2005 schrieb, werden auf Grund von zahlreichen Manipulationsversuchen, diese Metaangaben durch keine Web-Suchmaschine mehr abgegriffen[102]. Im Bibliothekskontext entsprechen diesen Metaangaben die Metadaten der Katalogisierung, wie beispielsweise Hauptsachtitel, Autor, Verlag, etc. Diese können im Sinne des Rankingfaktors „Position“ für das Ranking eingesetzt werden. Ein Problem stellt hierbei jedoch die Heterogenität der in einem Discovery-System vorliegenden Metadaten dar. Dieses wird in Kapitel 3.1.6. Übertragbarkeit näher diskutiert.

---

97 Behnert, Christiane; Lewandowski, Dirk (2015): Ranking search results in library information systems. In: The Journal of Academic Librarianship, 41/Heft 6, S. 731.

98 Lewandowski, Dirk (2005), Kpt. 6.1.

99 Lewandowski, Dirk (2015), Kpt. 5.2.3.

100 Behnert, Christiane; Borst, Timo (2015).

101 Der header ist ein Teil der HTML-Struktur von Webdokumenten, der durch den Ersteller zur Beschreibung und Strukturierung des Dokuments verwendet wird.

102 Lewandowski, Dirk (2005), S. 92.

Ebenfalls wird durch diesen Rankingfaktor die Position des Begriffs im Dokument allgemein ermittelt. Es werden folglich Dokumente im Ranking höher bewertet, die den Suchbegriff am Dokumentbeginn aufweisen, im Gegensatz zu Dokumenten die diesen erst am Ende enthalten[103]. Dem liegt die Annahme zu Grunde, dass die Nutzer durch ein Anlesen der Dokumente entscheiden, ob dieses für ihren Informationsbedarf von Relevanz ist. Findet sich der gesuchte Begriff nicht zu Beginn des Dokuments, wird dieses folglich als irrelevant bewertet[104].

Auch der Rankingfaktor der Position der Begriffe innerhalb eines Dokuments wird sowohl in Web-Suchmaschinen, als auch im traditionellen Bibliothekskatalogranking verwendet[105].

3.1.4. Längenangaben

Zwei weitere Rankingfaktoren der Textstatistik sind die Längenangaben. Diese umschließen zwei Aspekte: zum einen die Größe der Site auf der das Dokument abgelegt ist und zum anderen die Dokumentlänge[106].

Größe der Site:

Eine Site ist der gesamte Webauftritt einer privaten oder unternehmerischen Inhalteanbieters im Web. Zu ihr gehören neben der Homepage, alle Unterseiten und eventuelle herunterladbare Dateien o.ä. Die Größe der Site entspricht daher der Anzahl der zu ihr gehörenden Dokumente und dient in dieser Form den Web-Suchmaschinen als Rankingfaktor. Ihm zu Grunde liegt die Annahme, dass Sites mit einer hohen Zahl an Dokumenten, das entsprechende Thema besonders umfangreich bzw. ausführlich behandeln und somit ihre Dokumente eine höhere Relevanz besitzen[107]. Folglich werden Dokumente von großen Sites im Ranking aufgewertet, während Dokumente von kleinen Sites abgewertet werden[108].

Im Bibliothekskontext entfällt dieser Rankingfaktor in dieser Form, da die Datenbestände von Discovery-Systemen nicht untereinander vernetzt bzw. verlinkt sind, sodass sich keine zusammengehörenden „Auftritte“ erkennen lassen. Vernetzungen liegen hier lediglich in Form von gemeinsamen Verlagen, Herausgebern oder Autoren vor. Folglich wäre lediglich die Messung der „Publikationsgröße“ einer Person bzw. Institution möglich, die der Site eines Webinhalteanbieters entsprechen könnte. Dokumente von Personen bzw. Institutionen mit vielen Publikationen würden so im Ranking bevorzugt. Dies wäre jedoch im Hinblick auf die nicht zwangsläu-

[103] Lewandowski, Dirk (2015), Kpt. 5.2.3.
[104] Lewandowski, Dirk (2015).
[105] Yang, Sharon Q.; Hofmann, Melissa A. (2010): The next generation library catalog. In: Information Technology & Libraries, 29/Heft 3, S. 143.
[106] Behnert, Christiane; Lewandowski, Dirk (2015).
[107] Lewandowski, Dirk (2005).
[108] Ebd.

fig geringere inhaltliche Qualität von weniger oft publizierenden Wissenschaftlern oder Institutionen nicht gerechtfertigt.

Dokumentlänge:

Die Dokumentlänge stellt den zweiten Aspekt der Längenangaben dar. Dieser Rankingfaktor ist im Kontext des Webs, als Umgebung in die Jedermann Inhalte erstellen kann, ein weiterer Indikator für die Authentizität der Dokumente. Die Dokumente müssen lang genug sein, dass sie als aussagekräftig angesehen werden können, dürfen allerdings auch nicht zu lang sein[109]. Es werden folglich „Dokumente ab und bis zu einer gewissen Länge“[110], die als sinntragend definiert wurde, im Ranking bevorzugt.

Die Dokumentlänge ist im Bibliothekskontext komplementär zu der Dokumentlänge der Volltexte, oder entspricht im Falle von gedruckt vorliegenden Exemplaren, der Anzahl der Seiten[111]. Dieser Rankingfaktor kann demnach von den Web-Suchmaschinen übernommen werden. Hierbei ist jedoch zu berücksichtigen, dass die Länge der Publikationen von der entsprechenden Forschungsdisziplin abhängig ist[112]. So werden in den Naturwissenschaften beispielsweise deutlich mehr Kurzbeiträge publiziert, als z.B. in den Geisteswissenschaften.

3.1.5. Hervorhebungen und Ankertext

Hervorhebungen von einzelnen Begriffen in einem Dokument, von denen Ankertexte u.a. eine Variation darstellen, agieren ebenfalls als ein Rankingfaktor der Gruppe Textstatistik. Sie beziehen sich hauptsächlich auf die Textformatierung im Dokument.

Hervorhebungen können beispielsweise durch eine bestimmte Formatierung, z.B. Fett- oder Kursivdruck, durch das Schreiben eines Begriffes in Großlettern, oder durch die Wahl einer zur Standardschrift des Dokuments differenten Schriftgröße oder -art, Ausdruck finden[113]. Enthält ein Dokument den Suchbegriff in Form einer Hervorhebung, wird dieses im Ranking begünstigt. Dies erfolgt auf der Annahme, dass der Autor den Term im inhaltlichen Kontext als besonders wichtig empfindet und der Inhalt somit tatsächlich das gesuchte Thema beinhaltet.

Ankertexte:

Ankertexte sind Begriffe hinter denen ein Hyperlink zu einem weiteren Dokument verankert ist. Ein anschauliches Beispiel hierfür sind die verlinkten Wörter innerhalb eines Wikipediaarti-

[109] Lewandowski, Dirk (2005).
[110] Ebd., hier: S. 94.
[111] Behnert, Christiane; Lewandowski, Dirk (2015).
[112] Ebd.
[113] Erlhofer, Sebastian (2016), Kpt. 5.5.

kels, die zu einem weiteren Artikel führen. Alle Ankertexte die auf ein bestimmtes Dokument verweisen, werden als dessen alternative inhaltliche Beschreibung angesehen[114] und können so ebenfalls als eine Variante der Hervorhebung interpretiert werden[115]. Da der Aufbau des Index im Webkontext auf eben diesen Verlinkungsstrukturen basiert, stellt das Zusammentragen und Analysieren dieser dokumentexternen Linktexte für die Suchmaschinen kein Problem dar. Im Bibliothekskatalog existieren derartige Verlinkungen jedoch nicht, weshalb die Ankertexte als Variation des Rankingfaktors „Hervorhebungen" in Discovery-Systemen keine Anwendung finden.

3.1.6. Übertragbarkeit der Faktorgruppe Textstatistik

Wie in den vorangegangenen Unterkapiteln der Faktorengruppe Textstatistik dargelegt, findet eine Mehrzahl der diskutierten Rankingfaktoren dieser Gruppe in Discovery-Systemen bereits Anwendung. Hierzu zählen Term- und Dokumenthäufigkeit, Reihenfolge, Entfernung und Position der Begriffe, Dokumentlänge, sowie Hervorhebungen durch Textformatierung. Keine Anwendung finden lediglich die Größe der Site, sowie Ankertexte als Form der Hervorhebung.

Das Relevanz Ranking der kommerziell vertriebenen Discovery-Systeme basiert hauptsächlich auf textstatistischen Verfahren. Der Anbieter EBSCO gibt hierzu beispielsweise auf seiner Webseite an: „The major contributing factor in relevance scoring is the frequency of the user's search terms in matching database metadata and full-text records"[116].

Auch ExLibris nennt für sein Produkt Primo Discovery als erstes Rankingkriterium: „The degree to which an item matches the query. For example, Primo accords greater weight to an item if the query terms occur in specific metadata fields […] and if the order of the query terms or phrases is the same in the query and the record."[117]

Ähnlich äußern sich auch OCLC und ProQuest[118] zu ihren Produkten. Auch in den durch Bibliotheken selbst designten Discovery-Systemen spielt die Textstatistik eine zentrale Rolle. Oberhauser nennt hier für die UB Heidelberg u.a. insbesondere die Übereinstimmung zwischen Suchanfrage und den Metadatenfeldern Autor, Titel, Schlagwort usw., ebenso wie den Match mit der exakten Reihenfolge der eingegebenen Begriffe.

[114] Behnert, Christiane; Lewandowski, Dirk (2015).
[115] Lewandowski, Dirk (2005), Kpt. 6.1.
[116] EBSCO Inc.: Relevance Ranking.
[117] Ex Libris Ltd. (2015): Primo Discovery. Search, Ranking, and Beyond, hier: S. 3.
[118] Siehe hierzu: OCLC Inc.: How does relevance ranking work in WorldCat Local?; Ex Libris: How does the Summon service determine the order of search results?.

Probleme bei der Übertragbarkeit:

Um die Faktoren der Textstatistik vergleichbar anwenden zu können, muss eine konstante Menge an Text für jeden Datensatz vorhanden sein[119]. Diese Menge schwankt im bibliothekarischen Kontext jedoch von einfachen bibliographischen Daten bis hin zu mehreren hundert Seiten starken Volltextdokumenten. Somit werden bei der Textstatistik nicht nur Volltexte, Abstracts und Zusammenfassungen etc. mit einbezogen[120], sondern auch die Inhalte der Metadatenfelder.

Die durch Fachpersonal erstellten bzw. kontrollierten Metadaten der Discovery-Systeme stellen gegenüber den unsystematisch, durch Jedermann erstellten Metatags der Webumgebung einen großen Vorteil für das Ranking der Textstatistik dar. Dellit und Boston[121] nennen hier insbesondere die Eignung der Metadatenfelder für Schlagworte, Titel, Deskriptoren, Dokumenttyp o.ä. Aufgrund dieser hohen Qualität der Metadaten funktioniert die Klassifizierung und Facettierung, d.h. die Zuordnung der einzelnen Dokumente zu bestimmten Themengebieten, der Datenbankbestände sehr gut. Prekär ist hierbei jedoch, dass nur ein geringer Teil der Datensätze eines Discovery-Systems tatsächlich durch die Bibliothek erstellt wird, der überwiegende Teil der Daten wird aus externen Datenquellen importiert[122] (vgl. Kapitel 2.2. Funktion von Discovery-Systemen). So fehlt bei Fremddaten häufig die Verknüpfung von Personennamen mit deren Pseudonymen, was je nach Schreibweise unterschiedliche Treffermengen erzeugt[123]. Ebenso sind durch die fehlende Normierung in der Schlagwortvergabe die Klassifikationen oft auf einen Teil der Dokumente nicht anwendbar[124]. Auch im Metadatenfeld der Publikationsform fehlt die Normierung[125]. So werden verschiedene Begriffe für ein und dasselbe verwendet, wie z.B. Zeitschriftenartikel, Fachzeitschriftenartikel oder Magazinartikel. Diese Beispiele zeigen, dass eine durchgehend gute Klassifizierung nur durch eine ebenso durchgehende hohe Qualität der Metadaten aller Dokumente einer Kollektion zu erreichen ist[126], welche durch die Datenheterogenität der importierten Datensätze nicht gegeben ist[127]. Im Webkontext ist eine solche durch die Möglichkeit zur Inhalteerstellung durch Jedermann ebenfalls

[119] Lewandowski, Dirk (2009): Ranking library materials. In: Library Hi Tech, 27/Heft 4, S. 584-593.
[120] Oberhauser, Otto (2010): Relevance Ranking in den Online-Katalogen der „Nächsten Generation". In: Mitteilungen der VÖB, 63/Heft 1/2, S. 25-35.
[121] Dellit, Alison; Boston, Tony (2010): Relevance ranking of results from MARC-based catalogues.
[122] Pfeffer, Magnus; Wiesenmüller, Heidrun (2016): Resource Discovery Systeme. In: Handbuch Informationskompetenz, S. 105-114.
[123] Ebd.
[124] Pfeffer, Magnus; Wiesenmüller, Heidrun (2016).
[125] Ebd.
[126] Ebd.
[127] Ebd.

nicht gegeben. Web-Suchmaschinen werten daher Dokumente mit schlechter Datenqualität im Ranking ab, da dies als Indikator für eine niedrige Qualität des Inhalts angesehen wird[128]. Ein entsprechender Umgang im bibliothekarischen Kontext hätte jedoch fatale Auswirkungen, da eine schlechte Repräsentation der Datensätze nicht zwingend mit einer geringen inhaltlichen Qualität einhergeht.

Eine offensichtliche Lösung des Problems der Datenheterogenität wäre eine Angleichung der Erschließungssysteme der externen Datenquellen[129]. Zudem müsste eine Datenverbesserung der bestehenden Datensätze im bibliothekarischen Sinne erfolgen[130]. Dies hätte jedoch zur Voraussetzung, dass die differenten Anbieter den Zugriff auf ihre Daten erlauben müssten, um diese Optimierung des Metadatenmanagements durchzuführen[131]. Als besonders wichtig führen Pfeffer und Wiesenmüller hier die „Homogenisierung in zentralen Punkten“[132], wie der inhaltlichen Klassifizierung, oder des Metadatenfeldes der Publikationsform an. Letzteres könne beispielsweise durch eine Nutzung von hierarchischen Strukturen ermöglicht werden, sodass z.B. alle Artikel derselben Zeitschrift automatisch den gleichen Inhalt im Feld Publikationsform erhalten.

Kann eine Angleichung der Metadaten nicht erfolgen, ist auch eine Anwendung der identischen Rankingalgorithmen auf die unterschiedlichen Datensätze nicht möglich[133]. Laut Lewandowski[134] müssten stattdessen, je nach Menge der vorhandenen Daten, Gruppen gebildet werden, die separat gerankt und anschließend zu einer gesamten Trefferliste zusammengesetzt werden müssten. In rudimentärer Form wird dies in den Discovery-Systemen bereits angewandt. So werden für jede Suchanfrage jeweils zwei Trefferlisten, eine für den (gedruckten) Buchbestand und eine für Aufsätze, generiert[135].

Zudem stellen die geringen Textmengen bibliografischer Daten, im Vergleich zu denen von Volltexten, ein Problem dar, welches zur Folge hat, dass die Algorithmen der Textstatistik im bibliothekarischen Kontext nicht effektiv genug eingesetzt werden können[136]. Die Textstatistik bietet somit eine erste, auf Textmatching basierte Relevanzbewertung. Weitere Rankingfakto-

[128] Lewandowski, Dirk (2015).
[129] Roscher, Mieke (2014), Kpt. 6.2.1.
[130] Pfeffer, Magnus; Wiesenmüller, Heidrun (2016).
[131] Ebd.
[132] Vgl. Ebd.
[133] Lewandowski, Dirk (2009).
[134] Vgl. Ebd., S. 588f.
[135] Zumstein, Philipp (2011).
[136] Behnert, Christiane; Borst, Timo (2015).

ren, wie sie im Folgenden diskutiert werden, können diese Sortierung um stärker auf inhaltliche Qualitätsbewertung abzielende Aspekte ergänzen.

3.2 Popularität

Im vorigen Kapitel wurde die Übertragbarkeit der auf Textstatistik basierenden Rankingfaktoren diskutiert. Es wurde festgestellt, dass diese sehr gut geeignet sind, um die Treffermenge einzugrenzen und erste Relevanzbewertungen vorzunehmen. Da diese jedoch lediglich auf Textmatching zwischen Suchanfrage und Dokumenten basieren, werden weitere Faktoren benötigt, um auch die inhaltliche Qualität der Dokumente in das Ranking mit einbeziehen zu können.

Die Rankingfaktoren der Gruppe Popularität versuchen dies, aufgrund der Annahme, dass „Qualität durch Popularität zum Ausdruck kommt“[137].

Die Faktoren der Gruppe Popularität können in die zwei verschiedenen Datenerhebungsmethoden Linktopologie und Nutzungsstatistik unterschieden werden. Dementsprechend ist die Gliederung dieses Kapitel. Nachstehende Rankingfaktoren, die auf der Ermittlung der Popularität eines Dokuments beruhen, werden im Folgenden erörtert:

- Linktopologie: PageRank bzw. Zitations-Relevanz
- Nutzungsstatistik
  - Klickverhalten
  - Verweildauer
  - Explizite Bewertungen
  - Nutzungshäufigkeit und Erwerbungsverhalten

3.2.1 Linktopologie

Die Linktopologie orientiert sich an der, durch die Webautoren mittels Linksetzung ausgedrückten, Popularität von Dokumenten. Die grundlegende Annahme dabei ist, dass Autoren von Webseiten nur auf andere Seiten verlinken, deren Inhalte sie als qualitativ hochwertig und für ihr Thema als wichtig empfinden[138].

Die linktopologischen Verfahren machen sich folglich die Verlinkungsstrukturen im Web zu nutze. Die Informationen zu sowohl den OUT-Links eines Dokuments als auch den IN-Links, soweit ermittelbar (vgl. Kapitel 2.1. Funktion von Suchmaschinen), entstehen als Nebenpro-

[137] Lewandowski, Dirk (2015), hier: S. 99.
[138] Lewandowski, Dirk (2015).

dukt des Crawlingprozesses, wodurch die Errechnung des Rankingwerts zeitsparend, anfrageunabhängig erfolgen kann.

3.2.1.1 PageRank bzw. Zitations-Relevanz

Das bekannteste linktopologische Verfahren, ist der PageRank von Google, der als wichtiger Rankingfaktor der Gruppe Popularität Verwendung findet und hier exemplarisch für den Webbereich vorgestellt werden soll. Auch andere Web-Suchmaschinen verwenden, unter anderem Namen (z.B. Web Rank bei Yahoo!), einen solchen Wert, der jedoch ähnlich errechnet wird.

Der PageRank ordnet jedem Webdokument einen Wert zu, der die Popularität, d.h. die inhaltliche Qualität der Seite, auf Grund der durch die Inhalteersteller gesetzten Links, wieder gibt. Er ist hoch, wenn das Dokument viele IN-Links auf sich vereinen kann[139]. Dies führt zu einer Aufwertung im Ranking. Wenn Dokumente mit einem eigenen hohen PageRank auf ein anderes verlinken, steigt dessen Wert stärker an, als wenn der IN-Link von einem Dokument mit geringem eigenem PageRank käme[140]. Der PageRank kann somit „vererbt" werden. Dies liegt darin begründet, dass bestimmte Dokumente, wie z.B. redaktionell geprüfte Nachrichtenseiten oder Seiten eines wissenschaftlichen Instituts, mehr Vertrauen in ihre inhaltliche Qualität, d.h. einen höheren Trust-Wert, besitzen, als z.B. Seiten einer Privatperson[141].

Der PageRank als Rankingfaktor ermöglicht die Identifikation und eine entsprechend hohe Positionierung potentiell hochwertiger Dokumente. Ebenso wird durch ihn die Unterscheidung zwischen originalen Webseiten und deren Fälschungen ermöglicht, da das Original mehr IN-Links auf sich vereinen kann, als die Fälschung[142].

Die Errechnung des PageRank erfolgt durch die Errechnung der Wahrscheinlichkeit, dass das Dokument bei zufälligem Verfolgen der Links, gefunden wird. Dies entspricht dem Modell des Zufallssurfers, der einem hypothetischen Nutzer entspricht, der ein Dokument aufruft, auf diesem zufällig einem Link folgt und dies auf jedem Dokument, auf das er auf diese Weise gelangt, wiederholt[143]. Der Abbruch wird ebenfalls zufällig gewählt. Die Formel hierzu lautet[144]:

Formel: $PR(A) = (1-d) + d (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$

PR(A) = PangeRank von Seite A
PR(Ti) = PageRank der Seiten i-n, die auf Seite A verlinken
C(Ti) = Anzahl aller Links auf Seite Ti (Ti = eine Seite zwischen T1 und Tn)
d = Dämpfungsfaktor zwischen 0 und 1

[139] Erlhofer, Sebastian (2016), Kpt. 7.2.
[140] Lewandowski, Dirk (2015).
[141] Ebd.
[142] Lewandowski, Dirk (2015).
[143] Erlhofer, Sebastian (2016), Kpt. 7.2.
[144] Ebd.

Der PageRank von Seite A wird mit Hilfe des PageRank der Seiten errechnet, die auf Seite A verlinken. Durch die Addierung wird mit jedem weiteren IN-Link der PageRank von Seite A gesteigert[145]. Aus der Gesamtanzahl der OUT-Links auf den verweisenden Dokumenten, wird die Wahrscheinlichkeit, dass ein Klick auf den Link zu Seite A erfolgt, in die Berechnung miteinbezogen. Folglich sinkt der PageRank, je mehr OUT-Links ein Dokument hat, da somit die Wahrscheinlichkeit sinkt, dass der Zufallsserver auf der Seite bleibt[146].

Ergänzend wird ein Dämpfungsfaktor dazu gerechnet, der den Stopp des Zufallssurfers miteinkalkuliert. Dieser Wert ist eine manuell definierte Konstante zwischen 0 und 1. Je kleiner dieser Wert ist, desto höher liegt die Wahrscheinlichkeit, dass das Verfolgen der Links abgebrochen wird. Somit haben Seiten mit keinen IN-Links generell einen geringen PageRank[147].

Der PageRank wurde für ein System von Dokumenten entwickelt, die durch gegenseitige Verweise (Links) untereinander verbunden sind. Da das Wissen schrittweise durch die kritische Auseinandersetzung mit schon Vorhandenem entsteht, ist auch die wissenschaftliche Literatur, durch das gegenseitige Rezipieren, auf ähnliche Weise miteinander vernetzt.

Der Rankingfaktor Zitations-Relevanz funktioniert für die wissenschaftliche Literatur simultan zur Popularitätsmessung des RageRank im Webkontext: Je öfter ein Werk zitiert wird, desto größer ist die ihm, durch andere Autoren, zugemessene Wichtigkeit, d.h. desto höher ist sein bibliometrischer Wert. Einem IN-Link entspricht hier folglich einem Zitat.

Die Zitations-Relevanz beruht auf Methoden der statistischen Auswertung von Publikationen, der so genannten Bibliometrie[148]. Die diesen Verfahren zugrundeliegende Annahme, dass ein Werk mit wenigen Zitaten zwangsläufig ein qualitativ schlechteres ist, als eines mit vielen Zitaten, wird im wissenschaftlichen Kontext allerdings stärker kritisiert, als für den PageRank im Webkontext. Dennoch finden die mit Hilfe der Bibliometrie erstellten Bestenlisten regelmäßig große Resonanz[149].

Im Gegensatz zu den Web-Suchmaschinen, die die Datengrundlage für die Errechnung des PageRank in Form ihrer Kopie des Web[150] auf ihren eigenen Servern ohne Zusatzleistung bereit hält, muss für die bibliometrischen Methoden eine externe Datenbank hinzugezogen werden. In diesen Datenbanken sind möglichst alle wissenschaftlichen Publikationen mit ihrer manuel-

[145] Erlhofer, Sebastian (2016).
[146] Ebd.
[147] Ebd.
[148] Behnert, Christiane; Borst, Timo (2015).
[149] Ball, Rafael (2015): Bibliometrie im Zeitalter von Open und Big Data.
[150] Siehe Kapitel 2.1. Funktion von Suchmaschinen.

len und maschinellen statistischen Zitationsauswertung verzeichnet. Da es jedoch keine Datenbank geben kann, in der tatsächlich alle Publikationen enthalten sind, besteht hierin ein strukturelles Problem aller hier aufgeführten biblimethrischen Methoden. Durch diese Unvollständigkeit der Datenbasis entstehen Benachteiligungen für die Artikel, Autoren etc. für die keine Zitationsdaten vorliegen[151]. Die gängigste dieser Datenbanken ist der Science Citation Index, aber auch Scopus von Elsevier, oder Google Scholar bieten entsprechende Daten[152].

Die Qualität die der PageRank mittels der Linkstrukturen für Webseiten errechnet, kann im Bibliothekskontext für drei Ebenen bestimmt werden: Für die Autoren, die Zeitschriften, in denen veröffentlicht wird und die einzelnen Artikel[153].

Ebene des Autors:

Für die Leistungsbewertung der Urheber wurde von Hirsch 2005 der Hirsch-Index vorgestellt[154]. Er wird aus der Anzahl der Publikationen einer Person und der Zitierhäufigkeit in einem bestimmten Zeitraum ermittelt. Dies erfolgt indem zunächst alle Publikationen einer Person nach ihrer Zitierhäufigkeit abnehmend sortiert gelistet werden und anschließend mit einer laufenden Nummer versehen werden. Der h-Index kann aus dieser Tabelle direkt abgelesen werde. Er stellt den Wert dar, bei dem die laufende Nummer mindestens so groß ist, wie die Anzahl an Zitierungen[155].

Auf diese Weise eliminiert der h-Index positive, wie negative Extremwerte und bevorzugt die durchschnittliche Zitierrate. Folglich erhält eine Person mit wenigen Publikationen, von der eine sehr häufig zitiert wird, dennoch keinen hohen h-Index. Dieser gemittelte Wert drückt den Einfluss, d.h. die Relevanz, die ein Autor besitzt, gemessen an der Zitierrate seiner Publikationen, aus.

Zudem ist der h-Index, wie bereits erwähnt, vom Zeitraum für den er ermittelt wird abhängig. In der Regel wird der Wert daher für das gesamte Forscherleben einer Person erhoben[156]. Hieraus ergibt sich jedoch die Kritik, dass ältere Autoren, gegenüber Autoren mit kürzerer Karriere, automatisch über einen höheren Wert und damit einen Vorteil im Ranking verfügen. Um

[151] Behnert, Christiane; Borst, Timo (2015).
[152] Ball, Rafael (2014): Bibliometrie. Einfach, verständlich, nachvollziehbar.
[153] Yan, Erjia; Ding, Ying (2010): Weighted citation. An indicator of an article's prestige.
[154] Hirsch, Jorge E. (2005): An index to quantify an individual's scientific research output. In: PNAS, 102/Heft 46, S. 16569–16572.
[155] Ball, Rafael (2015).
[156] Ebd.

dies näherungsweise auszugleichen, existiert eine Variation des h-Indexes in der der ermittelte Wert zusätzlich durch das wissenschaftliche Alter dividiert wird[157].

Weitere Anpassungen erfolgten auf Grund der Nichtberücksichtigung der verschiedenen Publikationskulturen von einzelnen Disziplinen[158], sowie der Anzahl der Koautoren. So wurden zuvor beispielsweise Autoren, die ihre Publikation ohne Koautoren verfassen, in keiner Weise bevorzugt[159]. Ebenso ist es möglich, die Selbstzitate der Autoren aus der Auswertung herauszunehmen, um keine Wettbewerbsvorteile zu erzeugen[160].

Mit dem Wert den der h-Index liefert, können folglich populäre, und damit renommierte, Autoren bzw. deren Publikationen im Ranking aufgewertet und so leichter gefunden werden.

Ebene der Zeitschrift:

Der Journal Impact Factor (JIF) wurde zur Messung des Einflusses bzw. der Popularität einzelner Zeitschriften durch Garfield entwickelt[161]. Mit Hilfe des JIF-Wertes können demnach wichtige Zeitschriften bzw. deren Artikel im Ranking aufgewertet werden. Ermittelt wird er, indem jeweils für die vergangenen zwei Jahre, die Anzahl der Zitate durch die Anzahl der erschienen Artikel dividiert wird[162]. Auch hier gilt, je höher der JIF-Wert, d.h. je mehr die Zeitschrift zitiert wurde, desto höher ist die angenommene Qualität und das Interesse der Wissenschaftler an der Zeitschrift. Dabei kann keine Aussage zur Qualität spezifischer Artikel getroffen werden, sodass der hohe Wert einer Zeitschrift eventuell von nur wenigen, hochfrequenten Artikeln zustande kommt.

Weitere Kritikpunkte an dieser Methode sind die Manipulierbarkeit des JIF durch Zitierzirkel und der willkürlich festgesetzte Zeitraum von zwei Jahren, für den die Auswertung erfolgt[163]. Dieser ist nicht für jede Disziplin sinnvoll, da in einigen Fächern häufiger als in Anderen, neue

[157] Plassmeier, Kim; Borst, Timo; Behnert, Christiane [u.a.] (2015): Evaluating Popularity Data for Relevance Ranking in Library Information Systems. In: Proceedings of the ASIST.
[158] Ball, Rafael (2015).
[159] Hirsch, Jorge E. (2010): An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. In: Journal Scientometrics, 85/Heft 3, S. 741-754.
[160] Ball, Rafael (2014).
[161] Garfield, Eugene (1972): Citation analysis as a tool in journal evaluation. In: Science, 178/Heft 4060, S. 471–479.
[162] Ball, Rafael (2014).
[163] Ebd.

Publikationen erscheinen (vgl. z.B. Informatik und Germanistik)[164]. Zudem können in bestimmten Fächern Zeitschriften, die nur wenig publizieren, dennoch hohen Einfluss haben[165].

Ebene der Artikel:

Glänzel & Schubert stellten 1988 die Characteristic Scores and Scales (CSS) Methode vor, mit der Aufsätze bzw. Zeitschriftenartikel entsprechend ihrer Zitationshäufigkeit beurteilt werden können[166]. Hierzu werden die Artikel entsprechend in die Klassen „sehr wenig zitiert", „wenig zitiert", „häufig zitiert" und „sehr häufig zitiert" eingeteilt[167]. Dabei gilt auch hier, je stärker ein Artikel zitiert wird, desto höher ist sein Ansehen und damit seine Position im Ranking.

Durch die Gewichtung der Zitate, kann, zusätzlich zur Zitierungshäufigkeit, auch z.B. der Veröffentlichungszeitpunkt des Artikels in die Berechnung des Wertes einfließen. So könnten Artikel die sofort nach Veröffentlichung hoch zitiert werden, einen höheren Wert erhalten, als die, die erst zu einem späteren Zeitpunkt rezipiert werden[168]. Ebenso könnte der Wert der Zeitschrift, in der der Artikel erschienen ist, in die Berechnung mit eingebunden werden[169]. Dies kann gemäß dem PageRank erfolgen, der, wie oben ausgeführt, die ‚Vererbung' des Wertes ermöglicht, d.h. ein IN-Link von einer Seite mit hohem PageRank, erhöht den eigenen Wert der Seite stärker, als der IN-Link einer weniger wichtigen Seite. So würde der Wert des Artikels durch den JIF-Wert der Zeitschrift, in dem er veröffentlicht wurde, durch diesen positiv oder negativ beeinflusst. Diese Vererbbarkeit ist auch auf die bibliometrischen Werte von Autoren und Zeitschriften übertragbar[170]. Demnach wäre etwa das Zitat einer populären, viel zitierten Zeitschrift mehr wert, als das Zitat einer tendenziell Unbekannten[171].

Allerdings stellt sich auch auf der Ebene der Artikel die Frage, nach dem Maß der Vergleichbarkeit. Zum einen entstehen Verzerrungen der Werte durch positive oder negative Spitzenwerte. Für die Normalisierung dieser Zitationsverteilungen kann ebenfalls die CSS Methode, wie für

[164] Ball, Rafael (2014).
[165] Garfield, Eugene (2006). The history and meaning of the journal impact factor. In: JAMA, 295/Heft 1, S. 90-93.
[166] Glänzel, Wolfgang; Schubert, András (1988): Characteristic Scores and Scales in Assessing Citation Impact. In: Journal of Information Science, 14/Heft 2, S. 123–127.
[167] Zit. n. Plassmeier, Kim; Borst, Timo; Behnert, Christiane [u.a.] (2015).
[168] Yan, Erjia; Ding, Ying (2010).
[169] Ebd.
[170] Ebd.
[171] Ebd.

das Projekt LibRank geschehen[172], verwendet werden. Auf diese Weise können die Effekte von Extremwerten reduziert werden, was die Vergleichbarkeit der Zitationshäufigkeiten erhöht[173].

Zum anderen besteht auch hier eine Nichtberücksichtigung der Publikationskonventionen einzelner Forschungsdisziplinen. So könnten je nach Disziplin 20 Zitationen eines Artikels als viel oder wenig interpretiert werden[174].

3.2.1.2 Zusammenfassung Linktopologie

Zusammenfassend lässt sich sagen, dass die linktopologischen Verfahren der Web-Suchmaschinen sich in Form der genannten bibliometrischen Methoden auf das bibliothekarische Ranking übertragen lassen. Die untersuchten kommerziellen Anbieter der Discovery-Systeme (OCLC, ProQuest, Ex Libris Group und EBSCO Industries) geben allesamt an, die Zitations-Relevanz in ihrem Ranking zu berücksichtigen[175].

Eine wesentliche Kritik an der Verwendung der Linktopologie im Ranking liegt auf wissenschaftlicher Seite jedoch in der Nichtbeachtung der disziplinabhängenden Publikations- und Zitierkonventionen. Dies führt zu einer Erschwerung der Vergleichbarkeit und daher zu unstimmigen Rankingergebnissen[176]. Deshalb ist ein für jede Disziplin, wenn nötig auch für jede Teildisziplin, separates Ranking auf Grundlage der bibliometrischen Werte zu empfehlen, dessen Ergebnisse anschließend im weiteren Rankingprozess oder auf der SERP zusammengeführt werden.

Linktopologische Verfahren können, sowohl im Web, als auch im Discovery-System, nur ein Indikator für Qualität sein. Wie bereits erwähnt, können Mängel in der Datenbasis zu Fehlern in der Berechnung führen, ebenso muss ein wenig zitiertes Werk nicht zwangsläufig eine niedere Qualität besitzen. Folglich sollte der PageRank bzw. die Zitations-Relevanz nicht der einzige Rankingfaktor in der Gruppe Popularität sein.

3.2.2 Nutzungsstatistik

Die zweite Kategorie der Popularitätsfaktoren stellen die nutzungsstatistischen Faktoren dar. Diese orientieren sich, im Gegensatz zu den linktopologischen Faktoren, an den Nutzern einer Suchmaschine, indem deren Interaktion mit dem System analysiert wird.

---

[172] Plassmeier, Kim (2016): Relevance Model. Working Paper.
[173] Plassmeier, Kim; Borst, Timo; Behnert, Christiane [u.a.] (2015).
[174] Behnert, Christiane; Borst, Timo (2015).
[175] OCLC Inc.; EBSCO Inc.; Ex Libris; Ex Libris Ltd. (2015).
[176] Ball, Rafael (2015).

Das nutzungsstatistische Popularitätsranking funktioniert nach dem Prinzip ‚was früher für diese Suchanfrage relevant war, ist es auch jetzt'. Dies muss nicht immer der Wahrheit entsprechen, erzielt im Webkontext jedoch effektive Rankingergebnisse.

Die Daten können sowohl in Bezug auf alle Nutzer der Suchmaschine, für einzelne Nutzergruppen, als auch auf einzelne Personen erhoben werden. Dieses Kapitel beschreibt lediglich die Erstgenannten. Das personalisierte Ranking wird dagegen in einem eigenen Kapitel diskutiert (siehe Kapitel 3.6. Personalisierung).

3.2.2.1 Klickverhalten

Das Klickverhalten der Nutzer stellt den ersten Rankingfaktor der auf Nutzungsstatistik basierenden Popularitätsfaktoren dar. Hierbei wird die Anzahl der Klicks, die eine Webseite erhält, gezählt[177]. Unter einem Klick wird dabei die Auswahl eines Treffers in der SERP verstanden, welche dem Dokument im Ranking als eine positive Relevanzbewertung von Seiten des Nutzers angerechnet wird. Das Klickverhalten als Rankingfaktor agiert nach dem Prinzip der „Weisheit der Vielen"[178], was zur Folge hat, dass je mehr Nutzer auf einen bestimmten Treffer klicken, d.h. das Dokument aufrufen, dieses einen entsprechend größeren Vorteil im Ranking erhält. Die benötigten Daten werden implizit aus den Protokolldaten erhoben[179].

Ob ein Nutzer jedoch auf einen Treffer klickt, ist zunächst abhängig von der Qualität der, durch die Suchmaschine automatisch generierten, Trefferbeschreibung auf der SERP[180]. Erscheint diese dem Nutzer als für die Deckung seines Informationsbedarfs als nicht relevant, wird die Webseite nicht angeklickt. Des Weiteren ist der Erhalt eines Klicks abhängig von der bereits erreichten Rankingposition der Webseite, da überwiegend nur die auf den oberen Rängen platzierten Dokumente aufgerufen werden[181]. Folglich profitieren von diesem Rankingfaktor nur die Dokumente, die bereits hohe Rankingpositionen besitzen[182]. Um auch weiter unten platzierte, relevante Seiten hochzuranken, benötigt es ergänzende Rankingfaktoren, wie beispielsweise die Aktualität (siehe Kapitel 3.3. Aktualität).

Ein Vorteil des Faktors „Klickverhalten", ist hingegen die schnelle Anpassungsfähigkeit der Web-Suchmaschine an aktuelle Änderungen des Informationsbedürfnisses ihrer Nutzer (vgl.

---

[177] Behnert, Christiane; Lewandowski, Dirk (2015).
[178] Lewandowski, Dirk (2015), Kpt. 5.3.2.
[179] Behnert, Christiane; Lewandowski, Dirk (2015).
[180] Joachims, Thorsten; Granka, Laura; Pan, Bing [u.a.] (2005): Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th annual international ACM SIGIR conference […], S. 161.
[181] Nutzerverhalten auf Google-Suchergebnisseiten.
[182] Lewandowski, Dirk (2015), Kpt. 5.3.2.

Kapitel 1.4. Relevanz und Pertinenz). Hierfür werden nicht alle Klicks die ein Dokument je erhalten hat in den Rankingalgorithmus mit einbezogen, sondern nur die Klicks kurzer Zeitabstände. Somit sind die Bevorzugung von Seiten mit Informationen zu einem aktuellen Ereignis und die folgende Herabstufung im Ranking, bei Interessensverlust des Ereignisses, möglich.

Gemäß dem Vorgehen der Web-Suchmaschinen, kann im Bibliothekskontext ein Klick von der SERP des Discovery-Systems, auf die Vollanzeige, d.h. die bibliografische Dokumentrepräsentation, dem Klick von der Trefferliste einer Web-Suchmaschine auf eine Webseite, gleichgesetzt werden. Das Verfahren entspricht somit der Klickinterpretation der Web-Suchmaschinen und wäre somit ohne Modifikation, direkt übertragbar.

In dieser Weise findet der Rankingfaktor Klickverhalten im Katalog der Staats- und Universitätsbibliothek Bremen (SuUB) Anwendung[183]. Mit ihrer titelbezogenen Transaction Log Analysis (TLA) kann das selbst entwickelte Discovery-System häufig nachgefragte Titel identifizieren und an die Spitze der Trefferliste ranken. Haake [u.a.] merken jedoch an, dass Neuerscheinungen „kurzfristig durch fehlende Klickaktivitäten im Verfahren benachteiligt werden“[184]. Dies würde sich jedoch nach weniger Zeit nutzergesteuert wieder relativieren.

Auch das Recommendersystem „Bibtip“ verwendet durch Klickverhalten erhobene Daten, mit denen es Empfehlungen nach dem Motto „Nutzer die sich für diesen Titel interessierten, interessierten sich auch für..." generiert[185]. Dieses Empfehlungssystem wurde bereits durch viele Öffentliche- und Wissenschaftliche Bibliotheken in Deutschland in den jeweiligen Bibliothekskatalog integriert[186].

Im Webkontext verbirgt sich hinter jedem in der SERP aufgelisteten Treffer, direkt das Dokument, d.h. die Webseite. Bei einem Klick auf eine Trefferanzeige in der SERP eines Discovery-Systems, gelangt der Nutzer jedoch zunächst auf die Vollanzeige des Treffers. In der Vollanzeige finden sich zunächst die bibliographischen Angaben eines Dokuments. Im Falle einiger Medien stellen diese die einzige Repräsentation dar, sofern keine zusätzlichen Anreicherungen, wie z.B. Abstracts, Inhaltsverzeichnisse etc. verfügbar sind. Für einen Teil der Dokumente, ist dagegen in der Vollanzeige auch der komplette Volltext hinterlegt. Diese Heterogenität der Erreichbarkeit der Dokumente kann bei der Übertragbarkeit des Rankingfaktors „Klickverhalten“ auf den Bibliothekskontext berücksichtigt werden.

---

[183] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015): Nur die ersten drei zählen! In: o-bib. Das offene Bibliotheksjournal, 2/Heft 2, S. 33-42.

[184] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015), hier: Abschnitt 6.

[185] Vgl. hierzu: Mönnich, Michael; Spiering, Marcus (2007): Bibtip. Recommendersystem für den Bibliothekskatalog. In: EUCOR-Bibliotheksinformationen 30, S. 4-8.;Bibtip. Homepage.

[186] Ebd.

Demnach wird das Klickverhalten zusätzlich in drei Teilfaktoren unterschieden: die Anzahl der Klicks auf einen bibliographischen Datensatz, Klicks auf Anreicherungen, wie Inhaltsverzeichnisse, Abstracts o.ä. und die Anzahl der Klicks auf die Verfügbarkeitsanzeige (Volltextdownload oder Ausleihe)[187]. Deren Gewichtung nimmt gemäß der Aufzählungsreihenfolge zu. So impliziert das Aufrufen eines Abstracts oder gar des Volltextes eine höhere Relevanzbewertung durch den Nutzer, als das bloße Aufrufen der bibliographischen Daten in der Vollanzeige[188]. Ein solches Vorgehen wäre folglich eine Ergänzung des Rankingfaktors Klickverhalten, wie er oben beschrieben wurde, in Form einer tiefergehenden Ausdifferenzierung.

Bedingt durch die höhere Gewichtung von Volltextklicks, entsteht jedoch das Problem, dass, durch die schnelle Verfügbarkeit des gesamten Dokuments, Treffer mit vorhandenem Volltext, im Ranking automatisch den Medien ohne hinterlegten Volltext, i.d.R. die gedruckten Bestände, überlegen wären[189]. Gleichzeitig wäre bei den gedruckt vorliegenden Beständen einer Bibliothek, die tatsächliche Nutzungsfrequenz jedoch nicht ermittelbar, da diese außerhalb des Systems stattfindet[190]. Somit ist eine Übertragbarkeit der Klickzahlen von elektronischen Dokumenten auf den Printmedienbestand einer Bibliothek nicht möglich.

Folglich ist eine dem im Bibliothekskontext anzutreffenden, verschiedenen Dokumentenrepräsentationen entsprechende Ausdifferenzierung des Klickverhaltens nicht möglich. Eine Übertragung des Rankingfaktors aus dem Webkontext ist nur möglich, wenn lediglich die Klicks von der SERP auf die bibliografischen Daten, die für alle Dokumente vorhanden sind, gezählt werden.

3.2.2.2 Verweildauer

Das Klickverhalten wird im Webkontext mit dem Rankingfaktor Verweildauer (dwell time) ergänzt. Unter der Verweildauer wird dabei die Zeit verstanden, in der der Nutzer das ausgewählte Dokument betrachtet[191]. Es wird davon ausgegangen, dass bei einer Rückkehr zur Trefferliste nach nur geringer Zeit, das Dokument durch den Nutzer als doch nicht relevant für das Informationsbedürfnis eingestuft wurde[192]. Somit erhält das Dokument im folgenden Ranking eine Abstufung. Kehrt der Nutzer jedoch nicht, oder erst nach einer bestimmten Zeitspanne, auf die SERP zurück, erhält das Dokument eine Aufwertung im Ranking[193]. Die Verweildauer

[187] Plassmeier, Kim (2016).
[188] Behnert, Christiane (2015): Relevance Ranking. State of the Art in Web Search and Library Catalogs. LibRank Technical Report.
[189] Behnert, Christiane (2015).
[190] Ebd.
[191] Behnert, Christiane; Lewandowski, Dirk (2015).
[192] Lewandowski, Dirk (2015), Kpt. 5.3.2.1.
[193] Ebd.

ergänzt somit das Klickverhalten dahingehend, dass sich nicht jeder Klick zwingend positiv auf das Ranking der Webseite auswirkt.

Bei der Übertragung auf den Bibliothekskontext, kann die Verweildauer, ebenso wie das Klickverhalten, auf nachstehende drei Bereiche bezogen werden: Verweildauer auf der bibliografischen Vollanzeige des Titels, Verweildauer auf den Anreicherungen, wie Inhaltsverzeichnisse, Zusammenfassungen, etc. und Verweildauer auf dem Volltext[194]. Die Forscher des Projekts LibRank, schlossen diesen Rankingfaktor von ihren Untersuchungen aus, da festgestellt werden konnte, dass sich die Zeitspanne einer positiv zu bewertenden nicht wesentlich von der einer negativ zu bewertenden Verweildauer unterscheidet[195]. Die Verweildauer könne somit im Bibliothekskatalog nicht das Interesse eines Nutzers korrekt wiederspiegeln. Da sich die Untersuchungen des Projekts jedoch nur auf die Wirtschaftswissenschaften beziehen, sollten an dieser Stelle für die anderen Forschungsdisziplinen ähnliche Erhebungen durchgeführt werden. Dennoch ist es wahrscheinlich, dass auch bei diesen kein signifikanter Unterschied messbar wäre, da fachunabhängig im Bibliothekskatalog die tabellarisch aufgebauten bibliografischen Daten, sowie evtl. die Kataloganreicherungen, für sowohl eine positive, als auch eine negative Relevanzentscheidung begutachtet werden müssen. Im Web dagegen wird die Relevanzentscheidung durch eine ansprechende bzw. intuitive Webseitengestaltung, wie z.B. große Überschriften, Navigationselemente etc. die mit „einem Blick“ erfassbar sind, unterstützt und somit beschleunigt. In diesem Falle wäre der Rankingfaktor Verweildauer nicht auf Discovery-Systeme übertragbar.

Auch in der untersuchten Literatur konnte in Bezug eine aktuelle Verwendung der Verweildauer in Discovery-Systemen gefunden werden. Da die Verweildauer in Kombination mit dem Klickvergalten im Webkontext einen äußerst wichtigen Rankingfaktor darstellt, könnte die Frage gestellt werden, ob die Anwendung des Klickverhaltens ohne die Verweildauer, überhaupt zielführend ist. Der praktische, alleinige Einsatz des Klickverhaltens scheint dies jedoch durch eine offensichtliche Rankingverbesserung zu wiederlegen (siehe Kapitel 3.2.2.1 Klickverhalten).

3.2.2.3 Explizite Bewertungen

Durch den Nutzer explizit erfolgte Bewertungen stellen einen weiteren Rankingfaktor der nutzungsstatistischen Popularität dar. Hierbei werden die Dokumente direkt durch den Nutzer mittels beispielsweise eines Likes (vgl. soziale Netzwerke), Sternevergabe (vgl. z.B. Amazon.de)

[194] Behnert, Christiane (2015).
[195] Vgl. Plassmeier, Kim (2016), Kpt. 3.2.2.

o.ä. bewertet[196]. Für das Ranking gilt dabei, je mehr Bewertungen ein einzelnes Dokument auf sich vereinen kann, desto höher wird es positioniert. Die Bewertungen werden folglich, gemäß der „Weisheit der Vielen", als nutzerseitigen Ausdruck für Qualität interpretiert.

Analog dazu sind die so genannten Web 2.0 Einbindungen der Bibliothekskataloge als direkte Übertragung der „expliziten Bewertungen" zu sehen. Nutzerseitige Empfehlungen, Rezensionen, Kommentare, Sterne- und Schlagwortvergabe etc. liefern im Bibliothekskontext identische Hinweise auf Popularität[197]. Auch Anschaffungsvorschläge könnten als eine Art der expliziten Bewertung angesehen werden[198].

Die Funktion dieses Rankingfaktors ist jedoch in existentiellem Maße von der Beteiligung der Nutzer abhängig. Diese ist allerdings, sobald die Identität neben einer Bewertung sichtbar ist, relativ gering ist[199]. Da diese Bewertungen nur nach erfolgreicher Anmeldung, d.h. Authentifizierung, im Bibliothekssystem erfolgen können, wäre folglich der Verwendung dieses Rankingfaktors ein Angebot zur Verschleierung der Identität, z.B. mittels Pseudonymen, zuträglich.

Die Notwendigkeit eines hohen Engagements der Nutzer wird insbesondere bei der nutzergesteuerten Schlagwortvergabe, dem social tagging, ersichtlich. Ist die Beteiligung zu niedrig, sind die vorliegenden Daten, neben der für eine empirische Verwertung zu geringen Menge, in schlechter Qualität, da die Nutzer ihre Schlagworte ohne Normierung, d.h. ohne Synonymkontrolle etc. vergeben[200]. Vornehmlich in solchen Fällen, ist eine entsprechend niedrige Gewichtung gegenüber den, durch Fachkräften vergebenen, Schlagworten im Rankingalgorithmus empfehlenswert. Ist die Beteiligung dagegen hoch, bilden sich die populärsten Schlagwörter zu einer Dominanz heraus, was die fehlende Normierung ansatzweise ausgleichen kann[201].

Bezogen auf die expliziten Bewertungen insgesamt, können durch ein hohes Engagement der Nutzer bestimmte Themen, die aktuell von Interesse sind, so genannte „hot topics", identifiziert werden. Somit stellen die Bewertungen auch einen Indikator für die Rankingfaktorengruppe „Aktualität" dar (siehe Kapitel 3.3.2. Aktualisierungsdatum)[202].

Zudem könnten Bewertungen aus externen Systemen, wie Amazon für Bücher, oder iTunes für Musik, etc. mit in das Ranking von Bibliothekskatalogen eingebunden werden. Die an dem

[196] Behnert, Christiane; Lewandowski, Dirk (2015).
[197] Ebd.
[198] Ebd.
[199] Stock, Wolfgang G.; Stock, Mechtild. (2013): Handbook of Information Science.
[200] Kroski, Ellyssa (2005): The Hive Mind: Folksonomies and User-Based Tagging. In: InfoTangle Blogsome.
[201] Ebd.
[202] Haustein, Golov, Luckanus, Reher, & Terliesner, 2010) tecnical report

Projekt LibRank Beteiligten argumentieren dagegen jedoch, dass Bibliotheken nach Möglichkeit keine Schnittstellen zu proprietärer Software betreiben wollten[203].

Eine Herausforderung bei der Anwendung von expliziten Nutzerbewertungen im Ranking von Bibliothekskatalogen, stellt die, durch die direkte Eingriffsmöglichkeit der Nutzer bedingte, Option zum Vandalismus dar[204]. So könnten Nutzer durch gezielt falsche Bewertungen versuchen, das Ranking zu manipulieren. Da diese Gefahr nicht zu verhindern ist, sollte sie in der Gewichtung des Faktors im Algorithmus berücksichtigt werden, indem diese relativ gering gehalten wird.

Die Datenbasis für eine Übertragbarkeit dieses Rankingfaktors ist, durch die Einbindung der Web 2.0 Funktionalitäten, in den Discovery-Systemen gegeben. Folglich wäre eine Übertragbarkeit in derselben Form, wie sie auch im Web verwendet wird, mit entsprechender Gewichtung im Rankingalgorithmus möglich. Jedoch findet dieses Potential im Ranking bibliothekarischer Suchmaschinen noch keine Anwendung[205]. Auch im Projekt LibRank wurde dieser Faktor mit der Begründung "the usage of rating systems is considered low in economics and no source for ratings is known" ausgeschlossen[206].

3.2.2.4 Nutzungshäufigkeit und Erwerbungsverhalten

Da das Erwerbungsverhalten als Rankingfaktor auf dem Faktor der Nutzungshäufigkeit aufbaut, werden beide nutzungsstatistischen Popularitätsindikatoren in einem Teilkapitel behandelt.

Nutzungshäufigkeit:

Die Nutzungshäufigkeit stellt einen weiteren Rankingfaktor den die Web-Suchmaschinen nutzen dar. Er zählt die Downloads von im Web angebotenen (z.B. PDF-)Dokumenten[207]. Da diese Zählung durch einen erfolgten Klick auf den Downloadbutton erfolgt[208], könnte dieser Rankingfaktor als Erweiterung bzw. Intensivierung des Klickverhaltens (siehe Kapitel 3.2.2.1) angesehen werden. Es wird davon ausgegangen, dass mit einem Download eine hohe Nutzungsintention verbunden ist und damit eine aussagekräftige Relevanzbewertung erfolgt ist, was wiederum ein Indikator für Popularität darstellt[209].

[203] Vgl. Behnert, Christiane (2015), S. 16.
[204] Kroski, Ellyssa (2005).
[205] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015).
[206] Plassmeier, Kim (2016).
[207] Ebd.
[208] Ebd.
[209] Behnert, Christiane (2015).

Äquivalent zum Vorgehen im Web, kann im Bibliothekskontext der Volltextdownload mit einer, dem Download von Anreicherungsmaterialien, wie Zusammenfassungen etc. gegenüber, stärkeren Rankinggewichtung angesehen werden[210]. Für die Printmedien könnten als ein Maß für die Nutzungshäufigkeit die Ausleihzahlen herhalten, welche die Anforderungen jedoch nur bedingt erfüllen[211]. So können diese beispielsweise eine eventuelle Präsenznutzung vor Ort nicht wiedergeben. Zudem wäre die Frage zu stellen, ob eine Leihfristverlängerung ebenfalls eine Aussage über die Nutzungsfrequenz, gleichwertig einer Erstausleihe angerechnet werden kann. Für die Durchführung des Projekts LibRank wurde diese Frage verneint[212].

Ebenso könnte die Übernahme von bibliografischen Datensätzen in ein Literaturverwaltungssystem, d.h. Klick auf den Exportbutton, neben dem Download oder der Ausleihe, als ein Indikator für die Popularität im Sinne der Nutzungshäufigkeit angesehen werden[213]. Des Weiteren werden für das Projekt LibRank zusätzlich das Speichern in der Favoritenliste, das Exportieren eines Dokuments per E-Mail, sowie das Zitieren eines Dokuments an sich als Indikatoren für die Nutzungsfrequenz gezählt[214].

Die entsprechenden Daten ergeben sich, wie im Webkontext, bei der technischen Interaktion direkt in den Bibliothekskatalogssystemen[215] und können anonym erhoben werden[216]. Das Sammeln von derartigen Nutzungsstatistiken (Download- und Ausleihzahlen) wird von Bibliotheken bereits zur eigenen Wirtschaftlichkeitsüberprüfung regelmäßig praktiziert[217], somit ist die Datengrundlage gegeben.

Die Bibliothek der North Carolina State University testete diesen Rankingfaktor bereits 2006 erfolgreich in ihrem Katalog[218]. Bei der Verwendung von Ausleihzahlen als Popularitätsfaktor wurden jedoch Benachteiligungen von Büchern festgestellt, die aus verschiedenen Gründen keine oder nur geringe Ausleihzahlen aufweisen können[219].

Somit ist eine Übertragbarkeit der Nutzerhäufigkeit in Bezug auf elektronische Volltextdokumente direkt möglich, jedoch liegt im Bereich der Printmedien eine mangelnde Vergleichbar-

[210] Behnert, Christiane (2015).
[211] Antelman, Kristin; Lynema, Emily; Pace, Andrew K. (2006): Toward a twenty-first century library catalogue. In: Information Technology & Libraries 25/Heft 3, S. 128–139.
[212] Vgl. Behnert, Christiane (2015), Kpt. 2.
[213] Behnert, Christiane; Lewandowski, Dirk (2015).
[214] Plassmeier, Kim (2016), Kpt. 3.2.3.
[215] Plassmeier, Kim; Borst, Timo; Behnert, Christiane [u.a.] (2015).
[216] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015).
[217] Zum Beispiel mittels der Fächerstatistik der Deutschen Bibliotheksstatistik.
[218] Antelman, Kristin; Lynema, Emily; Pace, Andrew K. (2006).
[219] Ebd.

keit von Downloads und Ausleihen vor. Der Rankingfaktor „Nutzungshäufigkeit“ ist folglich nur auf einen Teil der Bibliotheksdokumente problemlos anwendbar.

Erwerbungsverhalten:

Das Erwerbungsverhalten einer Bibliothek orientiert sich nicht nur an dem entsprechenden Erwerbungsprofil bzw. dem Sammelauftrag der jeweiligen Institution, sondern auch an der Nutzungsfrequenz der Medien[220]. Damit basiert das Erwerbungsverhalten, wie einleitend bereits erwähnt, auf dem Rankingfaktor der Nutzungshäufigkeit. Berücksichtigt werden kann zunächst die Anzahl der lokal vorrätigen Exemplare[221], denn das „Beschaffungsverhalten der Bibliothek reflektiert die Nachfragesituation am Standort“[222]. Weiterführend kann die Anzahl der im entsprechenden Bibliotheksverbund zur Verfügung stehenden Exemplare herangezogen werden. Zum standortunabhängigen Vergleich kann die Höhe der Auflage eines Titels[223], ergänzt um die Verkaufszahlen des Verlages[224], in das Ranking miteinbezogen werden.

Dieser Rankingfaktor findet in unterschiedlicher Breite in den Katalogen der Staats- und Universitätsbibliothek Bremen (SuUB) und der UB Heidelberg Anwendung. Während in Letzterem lediglich "die Anzahl der besitzenden Bibliotheken (als indirekter Parameter) berücksichtigt“ [225] wird, geben Haake [u.a.] für die SuUB an, neben der gekauften Exemplarzahl, auch die veröffentlichte Ausgabenzahl zu nutzen[226].

3.2.2.5 Zusammenfassung Nutzungsstatistik

Über die Übertragbarkeit der nutzungsstatistischen Popularitätsfaktoren auf das Ranking von Discovery-Systemen lässt sich zusammenfassend feststellen: Das Klickverhalten ist übertragbar, sofern keine Ausdifferenzierung in Klicks auf Anreicherungen und Volltexte erfolgt, sondern lediglich der Klick auf den auf bibliografischen Datensatz gezählt wird. Die Verweildauer ist nicht übertragbar, sofern tatsächlich kein Unterschied in der Betrachtungszeit zwischen einer positiv und einer negativ bewerteten Trefferansicht besteht. Explizite Bewertungen sind mit entsprechender Gewichtung, um z.B. Manipulationen zu relativieren, übertragbar. Die Nutzungshäufigkeit ist auf elektronische Medien direkt übertragbar, für die Printmedien be-

[220] Behnert, Christiane; Lewandowski, Dirk (2015).
[221] Yang, Sharon Q.; Hofmann, Melissa A. (2011): Next generation or current generation? A study of the OPACs of 260 academic libraries in the USA and Canada. In: Library Hi Tech, 29/Heft 2, S. 266-300.
[222] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015), hier: Abschnitt 5.
[223] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015).
[224] Behnert, Christiane; Lewandowski, Dirk (2015).
[225] Maylein, Leonhard; Langenstein, Anette (2013): Neues vom Relevanz-Ranking im HEIDI-Katalog der Universtitätsbibliothek Heidelberg. In: B.I.T.online - Zeitschrift für Bibliothek, Information und Technologie, Heft 3, S.190-200.
[226] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015), hier: Abschnitt 1.

steht allerdings eine mangelnde Vergleichbarkeit zwischen Downloads und Ausleihen. Sie kann um den Faktor des Erwerbungsverhaltens ergänzt werden.

Das Ranking nach Popularität, d.h. nach der „Weisheit der Vielen“, birgt die Gefahr weniger verbreitete Meinungen zu unterdrücken. Ein Dokument, das viele als relevant empfinden, erhält eine entsprechend hohe Positionierung, was wiederum dazu führt, dass das Dokument von vielen gesehen und angeklickt wird, was erneut seine Popularität steigert[227]. Eine Möglichkeit diesem Selbstverstärkungseffekt zu begegnen, wäre, wie von der SuUB praktiziert, den Relevanzaufschlag für jeden Rankingfaktor nur einmal pro Dokument im aktuellen Rankingprozess zu vergeben[228]. Ebenso wichtig ist in diesem Zusammenhang, dass die Popularitätsfaktoren nur einen Teil des Rankingalgorithmus darstellen und durch weitere Faktorengruppen ergänzt werden sollten.

3.3. Aktualität

Für Web-Suchmaschinen ist es essentiell ihren Nutzern stets aktuell relevante Dokumente anbieten zu können[229]. Hierfür ist zum einen eine möglichst aktuelle Kopie des Webs bzw. ein möglichst aktueller Index (siehe Kapitel 2.3. Bedeutung des Suchmaschinenindexes) entscheidend, zum anderen die Einbindung der Aktualität der Dokumente in das Ranking. In den vorigen Kapiteln wurde die Aktualität als Rankingfaktor z.T. schon tangiert. So kann beispielsweise die Popularität der Dokumente in Abhängigkeit der Zeit variieren. In diesem Kapitel sollen nun die Rankingfaktoren der Gruppe Aktualität hinsichtlich ihrer Übertragbarkeit von Web-Suchmaschinen auf Discovery-Systeme diskutiert werden. Diese sind:

- Erstellungsdatum
- Aktualisierungsdatum
- Aktualisierungsfrequenz

3.3.1. Erstellungsdatum

Das Datum an dem ein Dokument erstellt wurde, ist zunächst der naheliegendste Indikator für das Rankingkriterium Aktualität. Das Erstellungsdatum ist ein statischer Wert, der bereits im Index vermerkt werden kann und ist somit ohne Zeitverlust, anfrageunabhängig im Rankingprozess anwendbar[230]. Die Feststellung des exakten Erstellungsdatums eines Dokuments ist im Webkontext aufgrund der kontinuierlichen Dynamik des Webs jedoch schwierig. Da das Datum der Abspeicherung auf dem Server sich bei jeder kleinsten Veränderung des Dokuments aktua-

[227] Metahaven (2009): Periphere Kräfte. Zur Relevanz von Marginalität in Netzwerken. In: Deep Search.
[228] Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015).
[229] Behnert, Christiane; Borst, Timo (2015).
[230] Lewandowski, Dirk (2015), Kpt. 5.4.

lisiert, kann dieses nicht für diesen Zweck herangezogen werden. Stattdessen wird das erste Auffinden des Dokuments durch den Crawler der Suchmaschine als Erstellungsdatum gewertet[231].

Im Bibliothekskontext entspricht das Erstellungsdatum dem Publikations- bzw. Erscheinungsdatum. Aber auch das Datum der ersten Einarbeitung bzw. die erste Indexierung im Discovery-System kann mit dem Vorgehen der Web-Suchmaschinen korrelieren[232].

Die Anwendung einer Ergebnissortierung nach Aktualität, d.h. Erscheinungsjahr, wird im Bibliothekskontext seit langem praktiziert. Auch vor der Anwendung von Suchmaschinentechnologie durch die Discovery-Systeme, wurde in den Bibliothekskatalogen neben der alphabetischen Ergebnissortierung, auch eine chronologische angeboten[233]. Folglich ist der Teilfaktor der Aktualität „Erstellungsdatum“ auch im Rankingalgorithmus der kommerziellen Discovery-Software[234] und den selbstentwickelten Systemen, wie z.B. dem Katalog der UB Heidelberg[235] integriert.

3.3.2. Aktualisierungsdatum

Aus der bereits angesprochenen kontinuierlichen Veränderung der Webdokumente resultiert im Webkontext das Aktualisierungsdatum als weiterer Indikator für Aktualität. Dieses wird als der Zeitpunkt definiert, indem der Crawler die aktualisierte Version des Dokuments zum ersten Mal auffindet. Dabei muss die Veränderung des Dokuments in signifikantem Umfang stattgefunden haben. Kleinere Veränderungen, wie beispielsweise das Hinzufügen eines Links, fallen nicht darunter. Äquivalent hierzu verhält sich im Bibliothekskontext die Herausgabe und Indexierung einer neuen, überarbeiteten Ausgabe eines Werkes[236].

Ebenso wird das Aktualisierungsdatum durch eine plötzliche Vermehrung der IN-Links definiert. In diesem Punkt steht die Aktualität in direkter Korrelation mit den Rankingfaktoren der Gruppe Popularität. So steigt die Popularität von bestimmten Webseiten, wenn diese für ein aktuelles Ereignis unmittelbar eine Interessenssteigerung erfahren. Folglich steigen die linktopologisch und nutzungsstatistisch ermittelten Werte messbar an. Dies lässt sich über dieselben Indikatoren für den Bibliothekskontext feststellen. So kann beispielsweise, wie in Kapitel 3.2.2.3 (Explizite Bewertungen) bereits erwähnt, die Steigerung nutzerseitig vergebener Be-

[231] Lewandowski, Dirk (2015).
[232] Behnert, Christiane; Lewandowski, Dirk (2015).
[233] Oberhauser, Otto (2010).
[234] OCLC Inc.; EBSCO Inc.; Ex Libris; Ex Libris Ltd. (2015).
[235] Maylein, Leonhard; Langenstein, Anette (2013).
[236] Lewandowski, Dirk (2015).

wertungen für ein spezielles Dokument ein Hinweis auf „hot topics“ darstellen[237]. Ebenso kann die Steigerung von Zitationen eines bestimmten, älteren Werks ein Indikator für wiedererlangte Relevanz sein.

Wie eingangs bereits erwähnt, ist es für Web-Suchmaschinen wichtig, ihren Nutzern aktuelle Ergebnisse zu liefern. Je besser sie sich an aktuelle Ereignisse und der damit einhergehenden Popularitätssteigerung einzelner Dokumente, anpassen kann, desto höher ist die Qualität der Suchmaschine[238]. Im wissenschaftlichen Kontext ist dagegen, deutlich stärker als im Webkontext, die Aktualität der Dokumente nicht immer von Interesse. Die Notwendigkeit der Aktualität variiert in Abhängigkeit der Forschungsdisziplin. So ist beispielsweise in den Naturwissenschaften, in denen häufiger neue Erkenntnisse publiziert werden, die Aktualität im Ranking deutlich stärker zu gewichten, als vergleichsweise in den historischen Wissenschaften, in denen auch immer wieder ältere Literatur von Bedeutung sein kann[239].

3.3.3. Aktualisierungsfrequenz

Aus dem Aktualisierungsdatum geht im Webkontext der Rankingfaktor „Aktualisierungsfrequenz“ hervor. Diese wird aus der Aktualisierungsrate, d.h. wie oft ein Dokument in der Vergangenheit ein neues Aktualisierungsdatum erhalten hat, errechnet. Der entsprechende Wert wird dafür verwendet, um Dokumente zu identifizieren, die besonders oft den aktuellen Ereignissen angepasst werden, was als Indikator für Qualität angesehen wird[240]. Dies betrifft beispielsweise Nachrichtenseiten. Die auf den Servern eines Discovery-Systems abgelegten Dokumente können, im Gegensatz zu den Webdokumenten, jedoch nicht bearbeitet bzw. verändert werden. Eine eventuelle Aktualisierung der Inhalte wird hier, wie oben ausgeführt, lediglich durch das Hinzufügen neuer, überarbeiteter Ausgaben oder durch eine Veränderung der Popularität gemessen. Damit entfällt dieser Rankingfaktor für eine Übertragung auf bibliothekarische Suchmaschinen.

Zusammenfassung:

Das Erstellungsdatum als Rankingfaktor wird in der Bibliothekspraxis bereits seit geraumer Zeit angewendet und ist somit direkt übertragbar. Auch das Aktualisierungsdatum aus dem Webkontext, hinsichtlich der Dokumentaktualisierung in Form von neuen Ausgaben und der plötzli-

[237] Haustein, Stefanie; Golov, Evgeni; Luckanus, Kathleen [u.a.] (2010): Journal evaluation and Science 2.0. In: Book of Abstracts of the 11th Internat. Conference on Science and Technology Indicators, S. 117-119.

[238] Lewandowski, Dirk; Höchstötter, Nadine (2008): Web searching. A quality measurement perspective. In: Web Search.

[239] Roscher, Mieke (2014).

[240] Lewandowski, Dirk (2015).

chen Steigerung der Popularität, ist ohne Probleme auf die Discovery-Systeme übertragbar. Ein Problem besteht jedoch auch hier wieder in den spezifischen Informationsbedürfnissen der differenten Disziplinen. Die Aktualisierungsfrequenz dagegen entfällt für den Bibliothekskontext komplett.

3.4. Lokalität

Geographische Daten können das Ranking ergänzen, da sie Informationen aus dem konkreten Nutzerkontext darstellen und somit zur Ermittlung des Informationsbedürfnisses beitragen[241]. In welchem Umfang, d.h. in welcher Gewichtung, die Lokalität in die Relevanzsortierung einbezogen wird, ist je nach Suchanfrage unterschiedlich. So werden einer Studie[242] zufolge z.B. bei der Suchanfrage „Rechtsanwalt" Dokumente in einem größeren Radius zum Nutzerstandort ausgegeben, als bei der Suchanfrage „Restaurant", da davon ausgegangen wird, dass der Nutzer für einen Rechtsanwalt bereit wäre, eine größere Strecke zurückzulegen. Somit kann nicht gesagt werden, in welchem Umfang oder wann überhaupt die Lokalität im Ranking Anwendung findet. Dennoch stellt sie in manchen Fällen eine wirksame Ergänzung dar.

Die Faktorengruppe Lokalität gliedert sich in folgende Rankingfaktoren:

- Physischer Standort des Nutzers und der Dokumente
- Inhaltlicher Standort des Dokuments

Ihre Übertragbarkeit auf bibliothekarische Suchmaschinen soll in diesem Kapitel diskutiert werden.

3.4.1. physischer Standort des Nutzers und der Dokumente

Der physische Standort beinhaltet den Nutzerstandort und den tatsächlichen Standort der Dokumente. Ersterer bezeichnet den Ort, an dem sich der Nutzer zum Zeitpunkt der Suchanfragestellung befindet[243]. Auf das Ranking wirkt sich dieser insoweit aus, dass Dokumente in der Nähe des Nutzers bevorzug werden, da davon ausgegangen wird, dass diese für das Informationsbedürfnis einen höheren Wert besitzen[244]. So ist etwa für den Suchbegriff „Eisdiele" anzunehmen, dass der Nutzer die Intention hat, eine solche aufzusuchen. Für diesen Zweck wäre eine durch Popularitätsfaktoren auf die oberen Positionen gerankte Eisdiele in Italien für einen Nutzer in Köln irrelevant, da diese aufgrund der Entfernung nicht in nächster Zeit aufge-

[241] Baeza-Yates, Ricardo; Broder, Andrei Z.; Maarek, Yoelle (2011): The new frontier of web search technology. Seven challenges. In: Search Computing. Trends and Developments, S. 3-9.

[242] Jones, Rosie; Zhang, Wei; Rey, Benjamin [u.a.] (2008): Geographic intention and modification in web search. In: International Journal of Geographical Information Science, 22/Heft 3, S. 1-20.

[243] Lewandowski, Dirk (2015), Kpt. 5.5.

[244] Ebd.

sucht werden kann. Folglich wäre es sinnvoller, Webseiten von Eisdielen in der Nähe des Nutzers nach oben zu ranken. Im dezentralen Web kann der Nutzerstandort jeder beliebige Ort sein. Die Lokalisierung erfolgt über GPS oder die IP-Adresse[245].

Im Bibliothekskontext ist jedoch zwischen folgenden drei möglichen Nutzerstandorten zu unterscheiden: im Bibliotheksnetzwerk, d.h. in den Räumlichkeiten der Bibliothek, einem beliebigen Ort außerhalb der Bibliothek und ihres Netzwerks, sowie außerhalb der Bibliothek, aber z.B. mittels VPN innerhalb des Bibliotheksnetzwerks[246]. Äquivalent zum Webkontext kann impliziert werden, dass der Nutzer Dokumente bevorzugt, die seinem Standort entsprechen. Folglich könnten einem Nutzer außerhalb der Bibliothek vermehrt elektronische Dokumente mit Volltextverfügbarkeit hoch gerankt werden, da diese das Informationsbedürfnis schneller befriedigen können, als die sich in der Bibliothek befindlichen gedruckten Bestände. Diese wiederum könnten den Nutzern vor Ort in der Bibliothek auf höheren Positionen angezeigt werden[247]. Hiergegen wäre jedoch anzuführen, dass nicht ermittelt werden kann, ob der Nutzer nicht Medien recherchiert, die er anschließend bei einem Bibliotheksbesuch gezielt begutachten möchte[248]. Dementsprechend sollte der Rankingfaktor „Nutzerstandort“ nicht zu stark gewichtet werden.

Der physische Standort der Dokumente bezeichnet im Webkontext den Ort, an dem der Server, auf dem ein Dokument gespeichert ist, steht. Dieser findet als Rankingfaktor der Web-Suchmaschinen keine Anwendung, da er in keinem Zusammenhang mit dem Inhalt der Dokumente steht. Im Bibliothekskontext muss allerdings zwischen den elektronischen und den gedruckten Medien innerhalb der Bibliothek unterschieden werden[249]. Im direkten Zusammenhang mit der Lokalität der Medien, steht deren Verfügbarkeit. Von der Bibliothek lizensierte bzw. gekaufte Dokumente, sind, sofern sie nicht ausgeliehen wurden, zugänglich. Diese können im Ranking bevorzugt werden. Medien die sich nicht im Besitz der Bibliothek befinden, werden zwar in der SERP eines Discovery-Systems angezeigt, sind jedoch nicht verfügbar und werden dementsprechend im Ranking benachteiligt.

Um dem entgegenzuwirken, bieten die kommerziellen Anbieter von Discovery-Systemen den Bibliotheken die Möglichkeit des Boostings ihrer eigenen lokalen Bestände an, wodurch diese nicht in der großen Datenmenge aus der die Trefferlisten gebildet werden verschwinden[250].

[245] Lewandowski, Dirk (2015).
[246] Plassmeier, Kim (2016).
[247] Behnert, Christiane; Borst, Timo (2015).
[248] Behnert, Christiane; Lewandowski, Dirk (2015).
[249] Plassmeier, Kim (2016).
[250] Neubauer, Karl Wilhelm (2010).

Mit diesem Feature wird den Bibliotheken der direkte Eingriff in den, durch die Anbieter vorgegebenen, Rankingalgorithmus gewehrt, der ansonsten jedoch komplett bestehen bleibt[251].

Der Zusammenhang zwischen Lokalität der Medien und deren Verfügbarkeit, lässt sich auch weiterführend mit den Faktoren der Popularität verknüpfen. So werden beliebte Medien hoch gerankt und dementsprechend oft ausgeliehen. Da hohe Ausleihzahlen ein Maß für Popularität darstellen (siehe Kapitel 3.2.2.4 Nutzungshäufigkeit und Erwerbungsverhalten), erhalten diese Medien weiterhin einen Vorteil im Ranking. Für die Lokalität ist jedoch die Ausleihe eines Mediums, d.h. dessen Nicht-Verfügbarkeit, ein Kriterium das Medium im Ranking abzuwerten. Folglich entsteht bei der Anwendung aller Rankingfaktoren eine potentielle Widersprüchlichkeit.

3.4.2. Standort des Dokuments

Der Standort eines Dokuments bezieht sich im Webkontext auf den im Dokument inhaltlich behandelten Standort, im Sinne der bereits erwähnten Eisdielen in Köln bzw. Italien[252]. Die entsprechenden Daten werden hierbei aus den Inhalten der Dokumente selbst entnommen, so z.B. aus der Adresse im Impressum usw. und dem Klickverhalten in Kombination mit dem Nutzerstandort (von wo wird diese Seite am meisten angeklickt?)[253].

Die Anwendung des inhaltlichen Standorts der Dokumente im bibliothekarischen Ranking ergibt, angesichts der internationalen Forschungsthemen der Wissenschaften, keinen Sinn. Eine Bevorzugung von Publikationen, die von Instituten in der Nähe des suchenden Forschers herausgegeben wurden, würde aufgrund der internationalen Wissenschaftsvernetzung ebenfalls keinen nutzerseitigen Vorteil erbringen. Somit entfällt dieser Rankingfaktor für die Übertragung auf Discovery-Systeme.

Zusammenfassung:

Die Lokalität als Rankingfaktor bietet im Zusammenhang mit dem Nutzerstandort Potential für die Relevanzsortierung in Discovery-Systemen. In Abhängigkeit zum Nutzerstandort können gedruckte oder elektronische Titel bevorzugt werden. Der physische Nutzerstandort ist somit aus dem Webkontext übertragbar, sollte jedoch nicht zu stark gewichtet werden. Der physische Standort der Dokumente spielt im Webkontext keine Rolle, im Bibliothekskontext besteht hiermit jedoch ein direkter Zusammenhang zur Verfügbarkeit der Medien. Der inhaltliche

[251] Vgl. Ex Libris Ltd. (2015), S. 5.
[252] Lewandowski, Dirk (2015).
[253] Ebd.

Standort eines Dokuments ist im Webkontext ein wichtiger Rankingfaktor, im Discovery-System dagegen nicht anwendbar.

3.5. Technische Faktoren

In den vorigen Kapiteln wurden Gruppen von Rankingfaktoren diskutiert, die sich auf die Feststellung der inhaltlichen Qualität beziehen. Die technischen Faktoren ergänzen dies mit der Beurteilung der formalen Eigenschaften der zu rankenden Dokumente. Dies ist im Webkontext durch die Möglichkeit Jedermanns, sich an der Inhaltserstellung des Webs zu beteiligen, nötig, um die technisch schlecht entwickelten Seiten im Ranking abzuwerten. Die Übertragbarkeit folgender technischer Faktoren auf Discovery-Systeme soll in diesem Kapitel diskutiert werden:

- Ladegeschwindigkeit und Adaptierbarkeit auf mobile Endgeräte
- Anreicherungen
- Sprache
- Dateiformat

3.5.1. Ladegeschwindigkeit und Adaptierbarkeit auf mobile Endgeräte

Die Ladegeschwindigkeit und die Adaptierbarkeit auf mobile Endgeräte sind im Webkontext sehr wichtige technische Rankingfaktoren. Unter der Ladegeschwindigkeit wird dabei die Zeit verstanden, die eine Webseite benötigt, um sich nachdem sie angeklickt wurde, aufzubauen. Da der Nutzer bei langen Wartezeiten tendenziell schneller den Ladevorgang abbricht und auf die Trefferliste zurückkehrt, werden Dokumente mit hoher Ladegeschwindigkeit im Ranking bevorzugt[254]. Im Bibliothekskontext dürfte dies jedoch keine Rolle spielen, da hier die mit den Treffern auf der SERP verknüpften Datensätze, d.h. die Vollanzeigen der einzelnen Titel, klein genug sein müssten, um sofort geladen werden zu können. Denkbar wäre hier lediglich eine Benachteiligung von Dokumenten aus externen Datenbanken, deren Volltexte aktuell aus technischen Gründen nicht verfügbar sind.

Die Adaptierbarkeit der Seiten auf mobile Endgeräte, spielt besonders im Zusammenhang mit dem Nutzerstandort (siehe Kapitel 3.4.1. physischer Standort des Nutzers und der Dokumente) eine Rolle. Hierbei werden Seiten bevorzugt, die sich der Bildschirmgröße des verwendeten Endgeräts anpassen können[255]. Auch für diesen Rankingfaktor besteht nicht die Notwendigkeit einer Übertragung in das bibliothekarische Ranking, da die Anpassung an das Endgerät auf Ebene der gesamten Katalogsoftware umgesetzt wird.

[254] Lewandowski, Dirk (2015), Kpt. 5.7.
[255] Ebd.

3.5.2. Anreicherungen

Der Rankingfaktor „Anreicherungen“ ermittelt im Webkontext die Angabe von Metatags für jedes Dokument[256]. Metatags sind dabei Auszeichnungen durch die Programmiersprache, die sich auf das gesamte Dokument beziehen (beispielsweise das <title>Tag, das dem Dokument einen Titel gibt, sowie alle anderen Metatags im header-Bereich eines HTML- Dokuments), aber auch Metadaten die für einzelne Teile der Webseiten vergeben werden, wie z.B. ein Alternativtext für Bilder o.ä.[257].

Die Anreicherungen als Rankingfaktor lassen sich in derselben Interpretationsweise der Web-Suchmaschinen auch auf den Bibliothekskontext übertragen. Datensätze, die nur unvollständige Metadaten besitzen, können so im Ranking abgewertet werden. Die ermöglicht, dass Dubletten, die sich durch eine schlechtere Datenqualität auszeichnen, auf niedere Positionen gesetzt werden können[258].

Der Rankingfaktor lässt sich im bibliothekarischen Ranking jedoch auch als Relevanzsteigerung durch einen Mehrwert, den die Anreicherungen den Dokumenten bieten, interpretieren. Dementsprechend können Dokumente, für die Anreicherungen, wie Inhaltsverzeichnisse, Abstracts, Coverbilder etc. hinterlegt sind, im Ranking bevorzugt werden. Denn diese könnten als Indikator für einen hohen Grad an Indexierungstiefe und somit für eine hohe Zuverlässigkeit des Inhalts, angesehen werden[259]. Problematisch wäre bei dieser Form der Anwendung des Rankingfaktors jedoch, dass lediglich Bücher, d.h. keine Zeitschriftenartikel oder ähnliches, über Anreicherungen wie Inhaltsverzeichnisse etc. verfügen können. Folglich wären diese stets im Vorteil.

Ebenso könnte das Vorhandensein von einer Verknüpfung zu den entsprechenden Forschungsdaten als Anreicherung betrachtet werden. Eine Bevorzugung solcher Dokumente könnte dahingehend begründet werden, als dass Forschungsdaten die Transparenz der angewandten Methoden, sowie eine Reproduzierbarkeit der Studie ermöglichen und damit ein Qualitätsmerkmal darstellen[260]. Hiergegen stellt sich wiederum das Problem, dass nicht in jeder Forschungsdisziplin Forschungsdaten im gleichen Maße anfallen und dass das Forschungsdatenmanagement in einigen Fächern, wie z.B. der Wirtschaftswissenschaft, noch nicht in ho-

[256] Behnert, Christiane; Lewandowski, Dirk (2015).
[257] Lewandowski, Dirk (2015).
[258] Plassmeier, Kim (2016).
[259] Behnert, Christiane; Lewandowski, Dirk (2015).
[260] Ebd.

hem Maße ausgearbeitet ist[261]. Als Lösung bietet sich ein fächergetrenntes Ranking an, indem somit die spezifische, relative Menge an Forschungsdaten berücksichtigt werden kann.

3.5.3. Sprache

Die Sprache eines Dokuments kann ebenfalls für das Ranking Verwendung finden. In Abhängigkeit mit dem Nutzerstandort (siehe Kapitel 3.4.1. physischer Standort des Nutzers und der Dokumente) wird im Webkontext die Frontendsprache[262] automatisch zugeteilt[263]. Im Discovery-System kann diese manuell ausgewählt werden, oder eine bevorzugte Sprache kann im Benutzerprofil festgelegt werden[264]. In Kombination mit der Sprache, in der die Suchanfrage gestellt wird[265], kann die Frontendsprache das Ranking dahingehend beeinflussen, als dass Dokumente in derselben Sprache bevorzugt werden. Folglich wird davon ausgegangen, dass der Nutzer Dokumente in der Sprache bevorzugt, die der der Eingabe und des Frontend entsprechen[266]. In dieser Form wird die Sprache als Rankingfaktor beispielsweise in der UB Heidelberg angewandt[267]. Dies erbringt insbesondere bei Suchanfragen mit Begriffen, die in mehreren Sprachen identisch geschrieben werden, wie z.B. „Computer“, einen deutlichen Vorteil für die Relevanzbewertung[268]. Somit ist auch dieser technische Faktor auf den Bibliothekskontext übertragbar.

3.5.4. Dateiformat

Ein weiterer technischer Rankingfaktor stellt das Dateiformat der Dokumente dar. Im Webkontext werden beispielsweise HTML-Dokumente gegenüber PDF- oder Word-Dokumenten bevorzugt, da „der Benutzer diese Dateien in seinem Browser sehen kann, ohne ein anderes Programm oder Plug-in zu öffnen“[269]. Im wissenschaftlichen Kontext könnten äquivalent hierzu PDF-Dokumente bevorzugt werden, da das PDF-Format „de[n] bevorzugte[n] Dateityp für elektronischen Volltext, der über die lizenzierten Zeitschriften der Bibliothek oder Volltextdatenbanken zur Verfügung gestellt wird“[270], darstellt. Dies würde jedoch wiederum dazu führen, dass elektronische Medien gegenüber den gedruckten Beständen einer Bibliothek, stets im Vorteil wären. Somit wäre im Falle einer Übertragung auf Discovery-Systeme, die Gewichtung des Dateiformats im Rankingalgorithmus für das gewünschte Maß der Bevorteilung von Onli-

[261] Plassmeier, Kim (2016).
[262] Das Frontend ist die Schnittstelle zwischen der Suchmaschine und dem Benutzer. Siehe hierzu Kapitel 2.1. Funktion von Suchmaschinen.
[263] Erlhofer, Sebastian (2016).
[264] Plassmeier, Kim (2016).
[265] Behnert, Christiane; Lewandowski, Dirk (2015).
[266] Maylein, Leonhard; Langenstein, Anette (2013).
[267] Ebd.
[268] Behnert, Christiane; Lewandowski, Dirk (2015).
[269] Lewandowski, Dirk (2005), zit. n. ebd.
[270] Behnert, Christiane; Lewandowski, Dirk (2015).

ne-Ressourcen ausschlaggebend. Als Alternative könnten den Printmedien ein ausgleichender Aufschlag im Ranking hinzugefügt werden.

Zusammenfassung:

Die technischen Rankingfaktoren „Ladegeschwindigkeit“ und „Adaptierbarkeit auf mobile Endgeräte“ können nicht auf das bibliothekarische Ranking übertragen werden. Die Faktoren „Anreicherungen“ und „Sprache“ können dagegen direkt übernommen werden. Im Falle der Anreicherungen, kann der Faktor, zusätzlich zu den Metadaten, um die Kataloganreicherungen im klassischen, bibliothekarischen Sinne erweitert werden. Die Übertragung des Dateiformats auf das Ranking von Discovery-Systemen ist prinzipiell möglich, führt jedoch zu einer stärkeren Gewichtung der Online-Medien, was individuell durch die jeweilige Bibliothek entschieden werden müsste.

3.6. Personalisierung

Die Personalisierung unterscheidet sich von den anderen Rankingfaktorgruppen dahingehend, dass sie keine ihr spezifischen Faktoren zur Datenerhebung verwendet, sondern vielmehr alle anderen Rankingfaktoren auf individueller Ebene anwendet. Hierfür werden, insbesondere die Faktoren der nutzungsstatistisch ermittelten Popularität (siehe Kapitel 3.2.2 Nutzungsstatistik), d.h. implizit erhobene Daten aus dem Surfverhalten und explizit gegebene Bewertungen, eingesetzt. Bei diesen wird versucht, durch die Analyse der Nutzungsdaten Aller, Rückschlüsse auf die Relevanzbewertung einzelner Personen zu ziehen. Im personalisierten Ranking werden diese Nutzungsdaten lediglich für individuelle Nutzer erhoben, um so das Ranking auf einzelne Personen und ihre Interessen zuzuschneiden. In diesem Kapitel soll die Übertragbarkeit des personalisierten Ranking von Web-Suchmaschinen auf Discovery-Systeme untersucht werden.

Web-Suchmaschinen setzten das personalisierte Ranking sehr intensiv ein. So ergab bereits 2011 eine Studie der Universität London, dass bei der Google-Suche ca. 64 % der Treffer auf der ersten SERP-Seite personalisierte Ergebnisanzeigen darstellen[271]. Ebenso konnte festgestellt werden, dass bei Suchen, die sich thematisch deutlich von den zuvor erfolgten Anfragen unterscheiden, im selben Maße personalisierte Treffer angezeigt werden. Hieraus wurde geschlossen, dass die Suchmaschine unmittelbar mit Hilfe der Nutzungsstatistik mit der Anlage eines individuellen Interessenprofils beginnt[272]. Folglich stellt die Personalisierung einen wich-

[271] Feuz, Martin; Fuller, Matthew; Stalder, Felix (2011): Personal Web searching in the age of semantic capitalism. Diagnosing the mechanisms of personalisation. In: First Monday, 16/Heft 2.

[272] Ebd.

tigen Bestandteil des Ranking im Webkontext dar und auch in Bibliotheken werden, wie im Folgenden aufgezeigt werden soll, einige Elemente davon bereits eingesetzt.

Durch das personalisierte Ranking kann die Qualität der Trefferliste für jeden individuell verbessert werden, indem Informationen, die für die Masse interessant sind, nicht jedoch für die jeweilige Person, herausgefiltert werden[273]. Dabei ist lediglich die Reihenfolge der angezeigten Treffer individuell, d.h. es werden weiterhin alle Treffer angezeigt, jedoch die für den Nutzer als besonders relevant ermittelten zu Beginn der Liste. Diese Treffer können Dokumente sein, die bereits bei einer ähnlichen oder gleichen Suche zuvor durch den Nutzer angeklickt wurden, oder Dokumente, die aufgrund des durch die Suchmaschine ermittelten persönlichen Such- und Interessenprofils als relevant identifiziert wurden[274]. Die Such- und Interessenprofile der Nutzer enthalten alle ermittelten Nutzungsdaten und stellen die Basis der Personalisierung dar.

Nutzerdaten:

Im Webkontext werden die Nutzerdaten, wie bereits erwähnt, überwiegend durch die Rankingfaktoren der Nutzungsstatistik ermittelt. Es werden somit mittels Klickverhalten und Verweildauer, sowie z.T. expliziten Bewertungen, Daten zu persönlichen Interessen, wie beispielsweise Hobbies, Sport, politische Orientierung, Gebiete mit Expertenwissen, usw. erhoben[275]. Ebenso können z.B. Lese- und Suchgewohnheiten, Geschlecht, Standort und ein ungefähres Alter ermittelt werden. Diese Daten werden in individuellen Interessensprofilen gespeichert und dienen dem personalisierten Ranking als Datengrundlage[276]. Dies kann dazu führen, dass ein Nutzer, der sich in der Vergangenheit oft für Sport interessiert hat, bei der Suchanfrage „Golf“ keine Webseiten zu Autos oder Meeresausbuchtungen auf den ersten Positionen seiner SERP erhält.

Im Bibliothekskontext können die Nutzerprofile aus individuellen Daten über Such- und Lesegewohnheiten, frühere Ausleihen sowie Klick- und Nutzerverhalten gebildet werden[277]. Letztere können Klicks auf bibliographische Datensätze, Anreicherungen, wie Inhaltsverzeichnisse und Zusammenfassungen etc. oder Volltexte umfassen und auch in Kombination mit der Termgewichtung erhoben werden[278]. Ebenso kann die Zugehörigkeit zu bestimmten Nutzer-

[273] Lewandowski, Dirk (2015), Kpt. 5.6.
[274] Ebd.
[275] Culliss, Gary A. (2001): Personalized Search Methods.
[276] Lewandowski, Dirk (2015).
[277] Behnert, Christiane; Borst, Timo (2015).
[278] Behnert, Christiane; Lewandowski, Dirk (2015).

gruppen mit in das persönliche Ranking einbezogen werden[279]. So gibt es beispielsweise in einer Universitätsbibliothek ein Nutzerspektrum mit verschiedenen Niveaustufen was Recherchekompetenz und akademischen Grad anbelangt[280]. Dementsprechend könnten z.B. bei Nutzern der Gruppe „Erstsemester" verstärkt Grundlagenwerke bevorzugt werden. Der entsprechende Status kann aus der Bibliotheks-ID ausgelesen werden[281]. Zudem kann, durch eine explizite, nutzerseitige Angabe, die jeweilige Forschungsdisziplin im Ranking Verwendung finden. Der selbstentwickelte Katalog der UB Heidelberg etwa, bevorzugt Dokumente, deren Fach mit dem des Nutzers übereinstimmen[282]. In den kommerziellen Discovery-Systemen ist eine solche nutzerseitige Rankingbeeinflussung ebenfalls möglich[283], jedoch gibt es hier, durch die in Kapitel 3.1.6. Übertragbarkeitangesprochene Datenheterogenität bzw. den Mängeln in den Metadaten der aggregierten Datensätze, zum Teil unstimmige Trefferanzeigen. Das Fachboosting könnte dennoch besonders für Bibliotheken profitabel sein, die ein breites Spektrum an Disziplinen abdecken müssen, in denen diverse (Such-) Begriffe jeweils kontextabhängig, unterschiedliche Bedeutungen aufweisen.

Erhebung:

Die aufgeführten Daten, die die Nutzerprofile bilden, werden durch die Suchmaschine sowohl anhand von impliziten aus dem Surfverhalten, als auch durch explizite Daten über Bewertungen erhoben[284]. Im Webkontext erfolgt dies beispielsweise durch die Auslesung der Browser-Cookies, dem die Nutzer explizit oder implizit durch die Nutzungsbedingungen zustimmen müssen. Die Interessenprofile werden durch die Suchmaschinen bevorzugt in den Accounts (z.B. Microsoft-Konto für Bing oder Google Konto für die Google-Suche) abgespeichert[285]. Hierzu ist vor der Suche eine Anmeldung im Konto nötig, jedoch können die Profile so auf lange Zeit angelegt werden und sind damit deutlich umfangreicher. Erfolgt keine Anmeldung, werden die Profile über die IP-Adresse anonymen, temporären Konten zugeordnet[286].

In der Bibliothek ist, aus rechtlichen Gründen, für die Anlegung von Nutzerprofilen ebenfalls eine Zustimmung durch den Nutzer zur Sammlung und Verwendung derartiger Daten einzuholen[287]. Diese kann durch eine Anmeldung im Bibliothekskonto gegeben werden, womit gleich-

[279] Lewandowski, Dirk (2009).
[280] Behnert, Christiane (2015).
[281] Ebd.
[282] Vgl. Maylein, Leonhard; Langenstein, Anette (2013), S.195.
[283] OCLC Inc.; EBSCO Inc.; Ex Libris; Ex Libris Ltd. (2015).
[284] Lewandowski, Dirk (2015).
[285] Behnert, Christiane; Lewandowski, Dirk (2015).
[286] Tißler, Jan (2011): SEO-Studie. In: t3n News.
[287] Behnert, Christiane; Borst, Timo (2015).

zeitig die erhobenen Daten durch die Authentifizierung zuordnungsbar sind. Somit kann das personalisierte Ranking im Bibliothekskontext lediglich nach erfolgreicher Anmeldung im Benutzerkonto angewendet werden.

Kritik:

Die Personalisierung der Suchergebnisse ruft jedoch oft Kritik hervor. Im Bereich des Web ist dies zumeist der Vorwurf, des Entstehens von Filterblasen[288]. Bei diesen werden durch die Algorithmen der Personalisierung lediglich Informationen angezeigt bzw. hochgerankt, die mit den bisherigen Meinungen übereinstimmen. Durch diese würde eine „konstruktive Auseinandersetzung mit politischen Fremdmeinungen verhinder[t]“[289] werden. Hiergegen kann jedoch gesagt werden, dass es dieses Phänomen bereits vor dem Internet in Form vom Lesen immer derselben Zeitschrift usw. gab[290]. Ebenso belegt eine gemeinsame Studie der Landesanstalt für Medien NRW und der Johannes Gutenberg-Universität Mainz, dass die Wirkung der Filterblasen überschätzt wird und die Nutzer sich keineswegs nur einseitig bilden[291].

Im Bereich der Bibliotheken spielen Filterblasen tendenziell eher eine untergeordnete Rolle, da die bibliothekarischen Suchmaschinen lediglich, eventuell mit Ausnahme der Öffentlichen Bibliotheken, für die wissenschaftliche Arbeit und nicht für eine mit den Web-Suchmaschinen vergleichbaren politischen oder freizeitlichen Meinungsbildung genutzt werden. Ein großer Kritikpunkt ist hier dagegen die mangelnde Transparenz des Ranking, die durch die Personalisierung zusätzlich verschärft wird. Durch die individuellen Trefferlisten sinkt das Vertrauen der Wissenschaftler in die Ergebnisreihung, da sie in besonderem Maße darauf angewiesen sind, alle vorhandenen Publikationen zu ihrem Forschungsthema zu finden[292]. Folglich sollte die Personalisierung des bibliothekarischen Ranking nur durch explizit durch den Nutzer gewählte Kriterien erfolgen. So überlegte die UB Heidelberg bereits 2013 die Einführung von Nutzerprofilen, in denen alle Personalisierungsoptionen jederzeit, manuell von den Nutzern aktiviert bzw. deaktiviert werden können[293]. Der bereits implementierte Fachboost wird zudem als aktiviert bzw. deaktiviert in der SERP angezeigt, sodass der Nutzer auch während einer Recherche über die aktuell verwendeten Personalisierungskriterien informiert ist[294].

---

[288] Metahaven (2009).

[289] Geib, Fabian (2017): Filterblasen. In: Silver Tipps.

[290] Ebd.

[291] Stark, Birgit; Magin, Melanie; Jürgens, Pascal (2017): Ganz meine Meinung? In: LfM-Dokumentation, Bd. 55.

[292] Vgl. Roscher, Mieke (2014), S. 54.

[293] Vgl. Maylein, Leonhard; Langenstein, Anette (2013), S. 196.

[294] Ebd.

Zusammenfassung:

Zusammenfassend lässt sich zur Übertragbarkeit des personalisierten Ranking sagen, dass die Discovery-Systeme viele Möglichkeiten bieten, entsprechende Nutzerdaten individuell zu erheben, die dann im Bibliothekskonto als Interessenprofil gespeichert werden können. Eine Hürde zur Anwendung ist in diesem Zusammenhang die notwendige Anmeldung im Konto, da dies nicht bei jeder Recherche praktiziert wird. Ein weiteres Problem stellt die, durch die von Nutzer zu Nutzer verschiedenen Suchergebnisse bedingte, Verstärkung der Intransparenz des Ranking dar. Somit ist das personalisierte Ranking im Bibliothekskontext nur anwendbar, wenn der Nutzer über dessen Einsatz exakt informiert ist und diesen steuern kann.

Zusammenfassung und Schlussfolgerungen

In dieser Arbeit wurde die Übertragbarkeit der Rankingfaktoren, wie sie von Web-Suchmaschinen verwendet werden, auf Discovery-Systeme analysiert. Die bisher überwiegend in Bibliothekskatalogen verwendeten textstatistischen Rankingfaktoren können lediglich eine Relevanzbewertung mittels Textmatching zwischen den Begriffen einer Suchanfrage und den Dokumenten, vornehmen. Um die Ergebnistreffer weiterführend hinsichtlich ihrer inhaltlichen Qualität sortieren zu können, werden weitere Faktoren aus den Gruppen Popularität, Aktualität, Lokalität, Technische Faktoren, sowie dem personalisierten Ranking benötigt. Ausgewählt wurden die entsprechenden Rankingfaktoren, nach ihrer Vorkommenshäufigkeit in der analysierten Literatur und der daraus abgeleiteten Wichtigkeit. Die Ergebnisse der Übertragbarkeitsanalyse für jeden Faktor, nach Gruppe sortiert, sind nochmals kompakt im Anhang wiedergegeben.

Es wurden 23 Rankingfaktoren aus sechs Faktorengruppen untersucht. Von diesen sind 14 (61%) vollständig direkt vom Ranking der Web-Suchmaschinen auf das bibliothekarische Ranking übertragbar. Zu diesen zählen unter anderem das Klickverhalten, das Erstellungsdatum, der Nutzerstandort, sowie die Sprache. Sechs (26 %) der untersuchten Faktoren sind dagegen nicht übertragbar. Die Linktopologie, die Nutzungshäufigkeit, sowie die Aktualisierungsfrequenz sind mit entsprechenden Modifikationen übertragbar.

Für die Anwendung der Textstatistik auf das bibliothekarische Ranking haben sich die unterschiedlichen Datenmengen der Dokumente als größte Herausforderung gezeigt. Da für manche Dokumente lediglich bibliographische Daten, für andere die Volltexte vorliegen, ist eine gleichförmige Anwendung der Textstatistik, wie sie im Web gegeben ist, nicht möglich. So muss in dieser Hinsicht eine Möglichkeit gefunden werden, wie mit der Heterogenität der Metadatenmengen umgegangen werden kann. Vorschläge wie das Erstellen von Gruppen nach vorhandener Datenmenge und deren separates Ranking oder die Implementierung einheitlicher Erschließungsregeln für jede Bezugsdatenbank, könnten bereits in die richtige Richtung gehen.

Die Gewichtung der einzelnen Faktoren ist ebenfalls ein Punkt, der genauer analysiert werden muss. Die ideale Gewichtung ist hierbei zunächst von dem jeweiligen Bibliotheksprofil abhängig. Die Qualität und Menge der vorhandenen Daten spielt jedoch auch hier eine wichtige Rolle. Diese bezieht sich nicht nur auf die Metadaten, sondern auch auf Daten, die durch andere Rankingfaktoren erhoben wurden, wie z.B. die der Nutzungsstatistik, insbesondere die Menge der expliziten nutzerseitigen Bewertungen. Entscheidend ist jedoch, wie sich ebenfalls zeigte, die jeweilige Disziplin für die spezifische Gewichtung der Faktoren im Rankingalgorithmus. Die

unterschiedliche Notwendigkeit an Aktualität, die differenten Mengen an produzierten Forschungsdaten im Bereich der technischen Faktoren, oder die individuellen Zitations- und Publikationskonventionen für die Linktopologie, sind einige Beispiele hierfür. Ob auch für die optimale Faktorengewichtung eine Gruppenbildung und deren separates Ranking eine Lösung sein und wie eine potentielle gegenseitige Neutralisierung einiger Faktoren (vgl. Popularität und Lokalität, Kapitel 3.4.1. physischer Standort des Nutzers und der Dokumente) verhindert werden kann, muss weiterführend diskutiert werden.

Ein weiterer, sich aus der Analyse der Übertragbarkeit der Web-Rankingfaktoren auf Discovery-Systeme ergebender Ansatz, der weiterverfolgt werden sollte, stellt der Umgang mit den elektronischen und den gedruckten Medien dar. Es hat sich gezeigt, dass durch Rankingfaktoren wie das Dateiformat, der Nutzungshäufigkeit, oder die Anreicherungen, d.h. das Vorhandensein von Volltexten, elektronische Medien den Gedruckten gegenüber, im Ranking stets einen Vorteil besitzen. Dies sollte jedoch, in Abstimmung mit dem jeweiligen Bibliotheksprofil, ebenfalls durch eine entsprechende Gewichtung reguliert werden.

Es hat sich zudem gezeigt, dass die Anwendung und Gewichtung der einzelnen Rankingfaktoren nicht pauschal erfolgen kann, sondern vielmehr für jede Forschungsdisziplin, z.T. auch für einige Suchanfragen, individuell geschieht. Dies hat zur Folge, dass sich das Ranking für den Nutzer unberechenbar gestaltet. Diese Intransparenz könnte, wie dargelegt, durch die Personalisierung zusätzlich gesteigert werden. Für die Wissenschaftler ist es jedoch essenziell, dass sie kein Dokument zu ihrem Forschungsthema übersehen. Somit sinkt durch die Intransparenz das Vertrauen in das Ranking und damit in das Discovery-System. Folglich wäre die Relevanzreihung für die ausführliche, professionelle Forschung weniger geeignet, als für den Einstieg in ein Thema und damit eher für Studierende als Zielgruppe. Es wäre daher angebracht, die Sortierung nach Erscheinungsjahr und Alphabet der alten Kataloge als Option beizubehalten.

Um eine Übertragung der Web-Rankingfaktoren auf Discovery-Systeme durchführen zu können, müssten folglich in weiterführenden Analysen der Umgang mit der Heterogenität der Datenmengen geklärt, sowie eine ausgeglichene Gewichtung der einzelnen Faktoren unter Berücksichtigung des Bibliotheksprofils, der verschiedenen Forschungsdisziplinen und ihrer gegenseitigen Beeinflussung erarbeitet werden.

Quellen- und Literaturverzeichnis

Antelman, Kristin; Lynema, Emily; Pace, Andrew K. (2006): Toward a twenty-first century library catalogue. In: Information Technology & Libraries 25/Heft 3, S. 128–139.

Baeza-Yates, Ricardo; Broder, Andrei Z.; Maarek, Yoelle (2011): The new frontier of web search technology. Seven challenges. In: Search Computing. Trends and Developments. Hrsg. von: Stefano Ceri; Marco Brambilla. Berlin [u.a.]: Springer, S. 3-9.

Ball, Rafael (2014): Bibliometrie. Einfach, verständlich, nachvollziehbar. Berlin [u.a.]: De Gruyter Saur.

Ball, Rafael (2015): Bibliometrie im Zeitalter von Open und Big Data. Das Ende des klassischen Indikatorenkanons. Wiesbaden: Dinges & Frick.

Behnert, Christiane (2015): Relevance Ranking. State of the Art in Web Search and Library Catalogs. LibRank Technical Report. Hamburg.

Behnert, Christiane; Borst, Timo (2015): Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen. Das DFG-Projekt LibRank. In: Bibliothek – Forschung und Praxis, 39/Heft 3, S. 384-393.

Behnert, Christiane; Lewandowski, Dirk (2015): Ranking search results in library information systems. Considering ranking approaches adapted from web search engines. In: The Journal of Academic Librarianship, 41/Heft 6, S. 725-735.

Bibtip. Homepage. online unter: http://www.bibtip.com/de/product.html, Zugriff 13.02.2018.

Bidder, Benjamin; Schultz, Stefan (2009): Microsoft und Yahoo verbünden sich gegen Google. Web-Allianz. In: Spiegel Online, Ressort Wirtschaft, 29.07.2009 - 14:04 Uhr, online unter: http://www.spiegel.de/wirtschaft/web-allianz-microsoft-und-yahoo-verbuenden-sich-gegen-google-a-639009.html, Zugriff 12.02.2018.

Breeding, Marshall (2012): Library Web-Scale. In: Computers in Libraries, 32/Heft 1, S. 19-22.

Broder, Andrei (2002): A taxonomy of web search. In: SIGIR Forum, 36/Heft 2, S. 3-10.

Bues, Johannes (2015): Klickwahrscheinlichkeiten in den Google SERPs. In: SISTRIX Blog, 25.10.2015, online unter: https://www.sistrix.de/news/klickwahrscheinlichkeiten-in-den-google-serps/, Zugriff 14.02.2018.

Culliss, Gary A. (2001): Personalized Search Methods / Ask Jeeves, Inc. Patent Nr. US 6,539,377 B1 vom 25.03.2003.

Dellit, Alison; Boston, Tony (2010): Relevance ranking of results from MARC-based catalogues. From guidelines to implementation exploiting structured metadata. National Library of Australia, online unter: https://www.nla.gov.au/content/relevance-ranking-of-results-from-marc-based-catalogues-from-guidelines-to-implementation, Zugriff 13.02.2018.

EBSCO Inc.: Relevance Ranking. Online unter: https://www.ebsco.com/technology/search/relevance-ranking, Zugriff 13.02.2018.

Erlhofer, Sebastian (2016): Suchmaschinen-Optimierung. Das umfassende Handbuch. 8., aktualisierte Aufl. Bonn: Rheinwerk-Verl..

Ex Libris: How does the Summon service determine the order of search results? Online unter: https://knowledge.exlibrisgroup.com/Summon/Product_Documentation/ Searching_in_The_Summon_Service/Search_Results/Summon%3A_Relevance_Ranking, Zugriff 13.02.2018.

Ex Libris Ltd. (2015): Primo Discovery. Search, Ranking, and Beyond. o.O, Online unter: https://knowledge.exlibrisgroup.com/@api/deki/files/26778/Primo_Search_and_Ranking.pdf, Zugriff 12.02.2018.

Feuz, Martin; Fuller, Matthew; Stalder, Felix (2011): Personal Web searching in the age of semantic capitalism. Diagnosing the mechanisms of personalisation. In: First Monday, 16/Heft 2, online unter: http://firstmonday.org/article/view/3344/2766, Zugriff 13.02.2018.

Garfield, Eugene (1972): Citation analysis as a tool in journal evaluation. Journals can be ranked by frequency and impact of citations for science policy studies. In: Science, 178/Heft 4060, S. 471–479.

Garfield, Eugene (2006). The history and meaning of the journal impact factor. In: JAMA – The Journal of the American Medical Association, 295/Heft 1, S. 90-93.

Geib, Fabian (2017): Filterblasen. Medien zwischen Individualität und Manipulation? In: Silver Tipps, 29.09.2017, online unter: http://www.silver-tipps.de/filterblase-medien-zwischen-individualitaet-und-manipulation/, Zugriff 14.02.2018.

Glänzel, Wolfgang; Schubert, András (1988): Characteristic Scores and Scales in Assessing Citation Impact. In: Journal of Information Science, 14/Heft 2, S. 123-127.

Glöggler, Michael (2003): Suchmaschinen im Internet. Funktionsweisen, Ranking-Methoden, Top-Positionen. Berlin [u.a.]: Springer.

Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias (2012): Informationserschließung und automatisches Indexieren. Ein Lehr- und Arbeitsbuch. Berlin [u.a.]: Springer.

Google – Inside search. Algorithms. Online unter: https://www.google.ca/insidesearch/howsearchworks/algorithms.html, Zugriff 12.02.2018.

Haake, Elmar; Blenkle, Martin; Ellis, Rachel [u.a.] (2015): Nur die ersten drei zählen! Optimierung der Rankingverfahren über Popularitätsfaktoren bei der Elektronischen Bibliothek Bremen (E-LIB). In: o-bib. Das offene Bibliotheksjournal, 2/Heft 2, S. 33-42.

Haustein, Stefanie; Golov, Evgeni; Luckanus, Kathleen [u.a.] (2010): Journal evaluation and Science 2.0. Using social bookmarks to analyze reader perception. In: Book of Abstracts of the Eleventh International Conference on Science and Technology Indicators, S. 117-119.

Hirsch, Jorge E. (2005): An index to quantify an individual's scientific research output. In: PNAS, 102/Heft 46, S. 16569-16572.

Hirsch, Jorge E. (2010): An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. In: Journal Scientometrics, 85/Heft 3, S. 741-754.

Joachims, Thorsten; Granka, Laura; Pan, Bing [u.a.] (2005): Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05, New York: ACM Press, S. 154-161.

Jones, Rosie; Zhang, Wei; Rey, Benjamin [u.a.] (2008): Geographic intention and modification in web search. In: International Journal of Geographical Information Science, 22/Heft 3, S. 1-20.

Kaluza, Harald (2013): Google Scholar versus EBSCO Discovery Service. Ein vergleichender Retrieval-Test. In: B.I.T.online – Innovativ Bd. 44, S. 59-80.

Kroski, Ellyssa (2005): The Hive Mind: Folksonomies and User-Based Tagging. In: InfoTangle Blogsome, 12.07.2005 - 11:48 Uhr, online unter: http://web20bp.com/13z2a6019/wp-content/uploads/2013/03/The-Hive-Mind-Folksonomies-2005.pdf, Zugriff 13.02.2018.

Lancaster, Frederick W.; Gale, V. (2003): Pertinence and Relevance. In: Encyclopedia of Library and Information Science. Bearb. von: Miriam A. Drake, Hrsg. von: Marcel Dekker. New York [u.a.], S. 2307-2316.

Leibniz-Informationszentrum Wirtschaft. Homepage/Forschung/LibRank. Online unter: http://www.zbw.eu/forschung/science-2-0/librank/, Zugriff 14.02.2018.

Lewandowski, Dirk; Höchstötter, Nadine (2008): Web searching. A quality measurement perspective. In: Web Search. Hrsg. von: Amanda Spink; Michael Zimmer. Berlin[u.a.]: Springer, S. 309–340.

Lewandowski, Dirk (2005): Web Information Retrieval. Technologien zur Informationssuche im Internet. Hrsg. von: Marlies Ockenfeld; Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V.. Frankfurt/M. (Informationswissenschaft der DGI, Bd. 7).

Lewandowski, Dirk (2009): Ranking library materials. In: Library Hi Tech, 27/Heft 4, S. 584-593.

Lewandowski, Dirk (2010): Der OPAC als Suchmaschine. In: Handbuch Bibliothek 2.0. Hrsg. von: J. Bergmann; P. Danowski. München: Saur/de Gruyter, S. 87-107.

Lewandowski, Dirk (2013): Suchmaschinenindices. In: Handbuch Internet-Suchmaschinen. Suchmaschinen zwischen Technik und Gesellschaft, Bd. 3. Hrsg. von: Dirk Lewandowski. Heidelberg: AKA, S. 143-162.

Lewandowski, Dirk (2015): Suchmaschinen verstehen. Berlin, [u.a.]: Springer.

Lewandowski, Dirk (2016): Suchmaschinenkompetenz als Baustein der Informationskompetenz. In: Handbuch Informationskompetenz. Hrsg. von: Wilfried Sühl-Strohmenger. 2., überarb. Aufl.. Berlin [u.a.]: De Gruyter Saur, S. 115-126.

Maylein, Leonhard; Langenstein, Anette (2013): Neues vom Relevanz-Ranking im HEIDI-Katalog der Universtitätsbibliothek Heidelberg. In: B.I.T.online - Zeitschrift für Bibliothek, Information und Technologie, Heft 3, S.190-200.

Metahaven (2009): Periphere Kräfte. Zur Relevanz von Marginalität in Netzwerken. In: Deep Search. Politik des Suchens jenseits von Google [Beiträge der Deep-Search-Konferenz am 8. November 2008 in Wien]. Hrsg. von: Konrad Becker; Felix Stalder. Innsbruck [u.a.]: Studien-Verl..

Mönnich, Michael; Spiering, Marcus (2007): Bibtip. Recommendersystem für den Bibliotheks-katalog. In: EUCOR-Bibliotheksinformationen 30, S. 4-8.

Nadella, Satya (2010): New Signals in Search. The Bing Social Layer. In: Bing blogs, 13.10.2010, online unter: https://blogs.bing.com/search/2010/10/13/new-signals-in-search-the-bing-social-layer, Zugriff 12.02.2018.

Neubauer, Karl Wilhelm (2010): Die Zukunft hat schon begonnen. Führen neue Dienstleis-tungsprodukte zu neuen Strategien für Bibliotheken? In: B.I.T.online - Zeitschrift für Bibliothek, Information und Technologie, Heft 1, online unter: http://www.b-i-t-online.de/heft/2010-01/fachbeitrag1, Zugriff 12.02.2018.

Nutzerverhalten auf Google-Suchergebnisseiten. Eine Eyetracking-Studie im Auftrag des Ar-beitskreises Suchmaschinen-Marketing des Bundesverbandes Digitale Wirtschaft (BVDW) e.V. (o.V. und o.J.) online unter: http://docplayer.org/10390994-Nutzerverhalten-auf-google-suchergebnisseiten.html#download_tab_content, Zugriff 13.02.2018.

Oberhauser, Otto (2010): Relevance Ranking in den Online-Katalogen der „Nächsten Generati-on“. In: Mitteilungen der VÖB, 63/Heft 1/2, S. 25-35.

OCLC Inc.: How does relevance ranking work in WorldCat Local? Online unter: https://www.oclc.org/support/services/worldcat-local/faq-op2.en.html, Zugriff 13.02.2018.

Pfeffer, Magnus; Wiesenmüller, Heidrun (2016): Resource Discovery Systeme. In: Handbuch Informationskompetenz. Hrsg. von: Wilfried Sühl-Strohmenger. 2., überarb. Aufl.. Berlin [u.a.]: De Gruyter Saur, S. 105-114.

Plassmeier, Kim; Borst, Timo; Behnert, Christiane [u.a.] (2015): Evaluating Popularity Data for Relevance Ranking in Library Information Systems. In: Proceedings of the Association for In-formation Science and Technology [Konferenzschrift vom 6.-10.11.2015, St. Louis, MO, USA].

Plassmeier, Kim (2016): Relevance Model. Working Paper. Volltext online unter: http://www.librank.info/de/articles/, Zugriff 13.02.2018.

Rangordnung. In: Wikipedia – Die freie Enzyklopädie, online unter: https://de.wikipedia.org/wiki/Rangordnung, Zugriff 12.02.2018.

Roscher, Mieke (2014): Fachdisziplinäre Bedürfnisse in der Gestaltung von Discovery-Lösungen. Wirklich ein Katalog für alle? Masterarbeit, Humboldt-Universität zu Berlin. In: Ber-liner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 356.

Stark, Birgit; Magin, Melanie; Jürgens, Pascal (2017): Ganz meine Meinung? Informationsintermediäre und Meinungsbildung. Eine Mehrmethodenstudie am Beispiel von Facebook. In: LfM-Dokumentation, Bd. 55.

Stock, Wolfgang G. (2007): Information-Retrieval. Informationen suchen und finden. München [u.a.]: Oldenbourg.

Stock, Wolfgang G.; Stock, Mechtild. (2013): Handbook of Information Science. Berlin [u.a.]: De Gruyter Saur.

Suchmaschinenranking. In: Wikipedia – Die freie Enzyklopädie, online unter: https://de.wikipedia.org/wiki/Suchmaschinenranking, Zugriff 12.02.2018.

Tillett, Barbara (2003): What is FRBR? A Conceptual Model for the Bibliographic Universe. In: Technicalities, 25/Heft 5.

Tißler, Jan (2011): SEO-Studie. Personalisierte Suche macht Ranking unberechenbar. In: t3n News, Ressort Marketing, 19.04.2011 – 14.25 Uhr, online unter: https://t3n.de/news/seo-studie-personalisierte-suche-macht-ranking-306470/, Zugriff 14.02.2018.

Was ist Rankingfaktor. In: SEO-united.de – Glossar, online unter: https://www.seo-united.de/glossar/rankingfaktor/, Zugriff 12.02.2018.

Yan, Erjia; Ding, Ying (2010): Weighted citation. An indicator of an article's prestige. (o.O.).

Yang, Sharon Q.; Hofmann, Melissa A. (2010): The next generation library catalog. A comparative study of the OPACs of Koha, Evergreen, and Voyager. In: Information Technology & Libraries, 29/Heft 3, S. 141-150.

Yang, Sharon Q.; Hofmann, Melissa A. (2011): Next generation or current generation? A study of the OPACs of 260 academic libraries in the USA and Canada. In: Library Hi Tech, 29/Heft 2, S. 266-300.

Zumstein, Philipp (2011): Suchmöglichkeiten in Primo auf dem Prüfstand. Universitätsbibliothek Trier.

Anhang

Übertragbarkeit	Rankingfaktor	Beschreibung	Anmerkungen
Textstatistik			
Direkt	**Term- und Dokument-häufigkeit**	Zahl des Vorkommens eines Begriffs in den Dokumenten	-
Direkt	**Reihenfolge**	Reihenfolge der Begriffe im Dokument entsprechend der Anfrage	-
Direkt	**Entfernung**	Entfernung der Begriffe aus der Anfrage im Dokument	-
Direkt	**Position**	Vorkommen der Begriffe an markanten Stellen (Überschriften, Seitenanfang, Metatags…)	Die Qualität und der Umfang der Metadaten schwanken im Discovery-System. Dies erschwert die Anwendung.
Entfällt	**Größe der Site**	Umfang der Anzahl von Unterseiten eines Websauftritts	Den Site-Umfang mit dem Publikationsmenge eines Autors gleichzusetzen ergibt keinen Sinn.
Direkt	**Dokumentlänge**	Tatsächliche Länge einer Webseite	Entspricht der Länge des Volltextes oder dem Seitenumfang der Printmedien. Ist jedoch abhängig von der jeweiligen Forschungsdisziplin.
Direkt	**Hervorhebungen**	Hervorhebungen einzelner Begriffe durch Formatierung.	Der Ankertext als Form einer Hervorhebung entfällt.
Popularität			
Mit Modifikation	**Linktopologie**	Ein Wert, der durch die IN-Links die Beliebtheit aus Autorensicht der Dokumente wiederspiegelt.	IN-Links = Zitate. Dem PageRank entsprechen die bibliometrischen Werte der Autoren, Zeitschriften und Artikel.
Direkt	**Klickverhalten**	Anzahl der Klicks auf ein Dokument	-
Entfällt	**Verweildauer**	Zeit, die ein Nutzer auf dem angeklickten Dokument verweilt	Die Zeitspanne eines positiven und eines negativen Dokumentaufrufs scheinen sich nicht zu unterscheiden
Direkt	**Explizite Bewertungen**	Nutzerseitige Bewertungen durch z.B. Sterne, Kommentar, Like…	Damit der Faktor aussagekräftig ist, muss die Beteiligung der Nutzer groß genug sein. Manipulation durch „falsche Bewertungen“.

Übertragbarkeit	Rankingfaktor	Beschreibung	Anmerkungen
Direkt / Mit Modifikation	**Nutzungshäufigkeit**	Downloadhäufigkeit +Ausleihzahlen Erweiterung: Exemplaranzahl (Erwerbungsverhalten)	Direkt auf elektronische Medien anwendbar. Für den Printbestand dienen Ausleihzahlen als Äquivalent. Problem: keine Erfassung der Präsenznutzung, Umgang mit Verlängerungen…
Aktualität			
Direkt	**Erstellungsdatum**	Datum an dem ein Dokument erstellt/publiziert wurde	Erstes Crawling entspricht Erscheinungsjahr/Veröffentlichungsdatum.
Mit Modifikation	**Aktualisierungsdatum**	Erstes Auffinden/Einarbeiten der aktualisierten Version + plötzliche Vermehrung der IN-Links	Erstes Crawling entspricht Herausgabe einer überarbeiteten Ausgabe. IN-Links entsprechen Zitationen.
Entfällt	**Aktualisierungsfrequenz**	Anzahl wie häufig ein Dokument ein neues Aktualisierungsdatum erhält	Findet im Discovery-System nicht statt.
Lokalität			
Direkt	**Nutzerstandort**	Standort des Nutzers: außerhalb der Bibliothek, in der Bibliothek, im Bibliotheksnetzwerk	Problem: Nutzer außerhalb der Bibliothek bevorzugen nicht zwangsläufig elektronische Medien
-	**Physischer Dokumentstandort**[295]	Tatsächlicher Aufenthalt des Dokuments: Serverstandort, ausgeliehen oder verfügbar	Keine Verwendung durch Web-Suchmaschinen.
Entfällt	**Inhaltlicher Dokumentstandort**	Ort, der in einem Dokument thematisiert wird	Für die Wissenschaft irrelevant, wegen Internationalität der Forschung
Technische Faktoren			
Entfällt	**Ladegeschwindigkeit**	Die Dauer bis eine Web-Seite sich aufgebaut hat	Dokumente im Discovery-System können schnell geladen werden.

[295] Der physische Dokumentstandort wird nicht zu den zu Übertragenden Rankingfaktoren gezählt, da er im Web keine Anwendung findet und somit nicht übertragen werden kann. Er wurde dennoch aufgeführt, da seine Anwendung im Discovery-System einen wichtigen Beitrag zum Ranking leisten kann.

Übertragbarkeit	Rankingfaktor	Beschreibung	Anmerkungen
Entfällt	**Adaptierbarkeit auf mobile Endgeräte**	Anpassung der Seite an Größe des Bildschirms	Betrifft nicht das Ranking, sondern das Discovery-System als Software.
Direkt	**Anreicherungen**	Ausführlichkeit der Metadatentags	Direkte Anwendung auf Metadatenfelder (Problem: Heterogenität). Erweiterung auf Kataloganreicherungen, Forschungsdaten
Direkt	**Sprache**	Sprache des Dokuments und der Suchanfrage, ggf. auch Sprache am Standort	-
Direkt	**Dateiformat**	Format des Dokuments	Ohne ausgleichende Gewichtung für Printmedien, erfolgt eine Bevorteilung der elektronischen Medien
Direkt	**Personalisierung**[296]	Anwendung der Rankingfaktoren, insbesondere die Nutzungsstatistik, auf Individuen	Anlegung von individuellen Nutzerprofilen notwendig. Aus diesem Grund nur nach Anmeldung im Account anwendbar. Erhöht die Intransparenz des Relevanzranking. Nutzer sollte über Anwendung einzelner Faktoren jederzeit entscheiden können.

[296] Streng genommen stellt die Personalisierung eine weitere Faktorengruppe dar. Diese Gruppe besitzt jedoch keine eigenen Faktoren, sondern bedient sich derer der anderen Gruppen. Da die Personalisierung bei Verwendung dennoch das Ranking stark beeinflusst, wird sie in der Zusammenfassung als ein einziger Rankingfaktor gewertet.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde.

Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Dies gilt auch für Quellen aus eigenen Arbeiten.

Ich versichere, dass ich diese Arbeit oder nicht zitierte Teile daraus vorher nicht in einem anderen Prüfungsverfahren eingereicht habe.

Mir ist bekannt, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs mittels einer Plagiatserkennungssoftware auf ungekennzeichnete Übernahme von fremdem geistigem Eigentum überprüft werden kann.

Köln, den 16.02.2018

Julia Wolf

Unterschrift