



Hochschule für Angewandte Wissenschaften Hamburg  
*Hamburg University of Applied Sciences*

DEPARTMENT INFORMATION

## *Bachelorarbeit*

### **Known-Item-Suchanfragen im Discoverysystem beluga: Retrievaleffektivität und Empfehlungen**

*vorgelegt von*  
***Imke Rulik***

Studiengang Bibliotheks- und Informationsmanagement

erster Prüfer: Prof. Dr. Dirk Lewandowski  
zweite Prüferin: Prof. Ursula Schulz

Hamburg, Dezember 2014

**Abstract:**

Die Suche nach bekannten Objekten, sogenannten Known-Items, wird in Bibliothekskatalogen häufig durchgeführt. Für solche Suchen ist das gesuchte Known-Item der relevanteste Treffer. Daher sollte das Known-Item in der Trefferliste weit oben gerankt werden, um von der recherchierenden Person wahrgenommen zu werden. In dieser Arbeit wird untersucht, wie erfolgreich die Suche nach Known-Items in einem Discoverysystem mit Relevanzranking ist. Dafür wurden Suchanfragen in der einfachen Suche aus den Logdateien des Discoverysystems beluga intellektuell klassifiziert und die Known-Item-Suchanfragen identifiziert. Fast die Hälfte aller Suchanfragen waren Suchen nach Known-Items. Drei Viertel der gesuchten Known-Items wurden auf dem ersten Rang der Trefferliste gerankt, wobei Zeitschriftentitel meist niedriger gerankt wurden. Da das Relevanzranking für Known-Item-Suchanfragen nicht verbessert werden kann, ohne dass dies negative Auswirkungen auf andere Suchanfragen hat, wurde anhand der Merkmale der Known-Item-Suchen ein Algorithmus entwickelt, der die Known-Item-Suchanfragen erkennt. Dann könnten mit einem auf Suchanfragentyp angepassten Relevanzranking die Known-Items besser gerankt werden. Die Evaluation des Algorithmus in einer Testmenge zeigt, dass mit einer Precision von 0,83 und einem Recall von 0,85 die Known-Item-Suchen erkannt werden können und somit das Ranking von 30% der Known-Item-Suchanfragen verbessert werden könnte.

**Schlagwörter:**

Discoverysystem, Information-Retrieval-Evaluierung, Known-Item-Suchanfrage, Retrievaleffektivität, Suchanfragenerkennung, Suchanfragentyp

# Inhaltsverzeichnis

1	Einleitung .....	1
1.1	Ziele der Untersuchung .....	2
1.2	Aufbau der Arbeit .....	3
2	Discoverysysteme.....	4
2.1	Aufbau und Funktionen eines Discoverysystems.....	5
2.2	Anbieter .....	7
2.3	Darstellungsoptionen zur Integration des Index .....	8
3	Relevanzranking.....	9
3.1	Rankingfaktoren in einem Discoverysystem .....	10
3.2	Relevanzbewertung .....	13
4	Suchverhalten .....	14
5	Die Known-Item-Suche.....	17
5.1	Identifizierung von Known-Item-Suchanfragen.....	18
5.2	Die Known-Item-Suche im OPAC .....	20
5.3	Die Known-Item-Suche im Discoverysystem .....	23
6	Methodik und Datenerhebung.....	27
6.1	Methodisches Vorgehen .....	27
6.2	Kennzahlen zur Evaluierung der Retrievaleffektivität .....	29
6.3	Das untersuchte Discoverysystem: beluga .....	30
6.4	Datenerhebung.....	31
7	Auswertung .....	32
7.1	Identifizierung der Known-Item-Suchanfragen.....	34
7.2	Anteil der Known-Item-Suchanfragen .....	36
7.3	Trefferanzahl .....	37
7.4	Eigenschaften der Known-Item-Suchanfragen.....	38
7.4.1	Wortanzahl .....	38

7.4.2	Metadaten .....	39
7.4.3	Signalwörter .....	41
7.5	Medientypen von Known-Item-Suchanfragen .....	41
7.6	Retrievaleffektivität der Known-Item-Suchanfragen .....	42
8	Algorithmus zur Erkennung der Known-Item-Suchanfragen .....	45
8.1	Geeignete Merkmale auf Grundlage der intellektuell klassifizierten Suchanfragen ..	46
8.2	Ablauf und Funktionen des Algorithmus .....	47
8.3	Klassifikation der Testmenge mit dem Algorithmus.....	50
8.3.1	Evaluierung des Algorithmus .....	51
8.3.2	Retrievaleffektivität für die erkannten Known-Item-Suchanfragen.....	53
9	Diskussion der Ergebnisse .....	54
10	Empfehlungen und Fazit .....	58
11	Literaturverzeichnis.....	61
	Anhang .....	i

## Abbildungsverzeichnis

Abbildung 1: Trefferanzahlen für Known-Item- und sonstige Suchanfragen.....	38
Abbildung 2: Anteil von Known-Item- und sonstigen Suchanfragen pro Wortanzahl.....	39
Abbildung 3: Anzahl der Known-Item- und sonstigen Suchanfragen mit bestimmten Merkmalen .....	40
Abbildung 4: Medientypen der Known-Items .....	42
Abbildung 5: Success @n für Known-Item- und unsichere Suchanfragen.....	43
Abbildung 6: Success @n für Known-Item-Suchanfragen nach bestimmten Medientypen....	45
Abbildung 7: Success @n für Known-Item-Suchanfragen mit bestimmten Merkmalen.....	45
Abbildung 8: Hauptfunktion des Algorithmus zur Erkennung von Known-Item-Suchanfragen .....	49
Abbildung 9: Success @n für die Known-Item-Suchanfragen, die mit dem Algorithmus erkannten wurden .....	54

## **Tabellenverzeichnis**

Tabelle 1: Anzahl und Anteile der Suchanfragentypen.....	37
Tabelle 2: Übereinstimmung zwischen Titel des Known-Items und Wörtern der Suchanfragen, die sich auf den Titel beziehen .....	41
Tabelle 3: Anteil der Suchanfragentypen an den Merkmalen in % .....	47
Tabelle 4: Anzahl der mit dem Algorithmus erfassten Suchanfragen und der intellektuell identifizierten Suchanfragentypen.....	52
Tabelle 5: Anteil der Suchanfragentypen an den Merkmalen, die mit dem Algorithmus erfasst wurden in % .....	53

## **Abkürzungsverzeichnis**

ISBN Internationale Standardbuchnummer

ISSN Internationale Standardnummer für fortlaufende Sammelwerke

KVK Karlsruher Virtueller Katalog

OPAC Online Public Access Catalogue

# 1 Einleitung

Mit dem World Wide Web sind Informationen nicht nur einfacher, schneller und komfortabler recherchierbar und verfügbar geworden, sondern auch die Menge an Informationen und damit auch die Anzahl wissenschaftlicher Publikationen hat überproportional zugenommen. Diese Informationsexplosion führt zu einer Informationsüberflutung, in der die Rolle der Bibliotheken als Vermittler von Informationen sich stark geändert hat (vgl. GANTERT & HACKER 2008, S. 289 f.). Bibliotheken können sich nicht mehr nur auf die Sichtbarmachung und den Nachweis ihres eigenen Bestandes beschränken, da zum einen viele Informationen nur als Zeitschriftenartikel oder in elektronischer Form vorhanden sind und zum anderen die Nutzer Bibliotheken nicht mehr als zentralen Anlaufpunkt für die Suche nach Informationen sehen. Diese Rolle hat längst das Web übernommen, wo man auf (gefühlte) alle Fragen und Informationsbedürfnisse mit Suchmaschinen wie Google schnell und komfortabel eine Antwort/die passende Webseite und auch das wissenschaftliche Dokument findet, das früher nur über die Vermittlung von Bibliotheken verfügbar war.

Zudem ist das Rechercheverhalten der Nutzer längst durch Erfahrungen und Gewohnheiten auf Webseiten wie Google oder Ebay geprägt. Diese Erfahrungen übertragen sie auch auf Bibliothekskataloge, womit sie jedoch im klassischen Online Public Access Catalogue (OPAC) weit weniger Sucherfolg haben. Lewandowski spricht in diesem Zusammenhang davon, dass „Suchmaschinen [...] die Nutzer zu einem 'schlechten' Rechercheverhalten erziehen“ (vgl. LEWANDOWSKI 2010, S. 88).

Eine zeitgemäße Bibliothek sollte es dem Nutzer ermöglichen, bequem, schnell und einfach sowie seinen Gewohnheiten entsprechend zu recherchieren. Dafür braucht er einen zentralen Sucheinstieg der „wie Google“ funktioniert – sowohl was die Technik im Hintergrund als auch die Gebrauchstauglichkeit, also die Usability, anbelangt.

Statt weiterhin zu versuchen, den Nutzer durch Schulungen und Tutorials zur korrekten Bedienung des OPACs zu befähigen, wurden Discoverysysteme wie beluga eingeführt. Discoverysysteme sollten verschiedenen Suchanfragetypen gerecht werden. Bei OPACs hat der Nutzer zumindest theoretisch die Möglichkeit, seine Treffermenge durch Einschränkung auf bestimmte Suchkategorien (z.B. Titel und/oder Autor für Suchen nach spezifischen Titeln bzw. Schlagwörter oder Klassifikationen für thematische Suchen) für einen Suchanfragetyp zu optimieren – auch wenn er es praktisch kaum tut. Bei Discoverysystemen ist dies zwar im



Nachhinein über das Filtern mit bestimmten Facetten möglich, aber auch das ist für den Nutzer mit Aufwand verbunden. Lewandowski sieht ein Problem bei Discoverysystemen „in der mangelhaften Unterstützung der zielgerichteten Suche und dort vor allem in einem mangelhaften Ranking der Ergebnisse“ (vgl. LEWANDOWSKI 2010, S. 90).

Discoverysysteme werden von vielen Bibliothekaren kritisch gesehen. Sie sind irritiert, dass die Suche mit Boole'schen Operatoren und Suchschlüsseln in dem Suchschlitz der Startseite nicht funktioniert und sie andere (und höhere) Treffermengen als im gewohnten OPAC bekommen, da die Suche nicht mehr dem Prinzip des „exact match“ folgt. Die Reduktion auf einen Suchschlitz wird als problematisch wahrgenommen (vgl. CHRISTENSEN 2014). Zudem wird vielfach anekdotenhaft berichtet, dass die Suche nach Known-Items schwieriger geworden ist, da Suchkategorien (wie z.B. Titel, Autor, Titelanfang) im einfachen Suchschlitz nicht mehr angewählt werden können, die Suchergebnisse nicht mehr so präzise sind und das Relevanzranking besonders thematische Suchanfragen unterstützt (vgl. z.B. CMOR & LITWIN 2014, S. 214; FRY 2011; GRANT 2014; HÄNGER 2013, S. 7; SLAVEN et al. 2011, S. 15). Nutzer erwarten aber, dass relevante Treffer weit oben auf der Trefferliste erscheinen und beachten Treffer weiter unten kaum (vgl. SCHNEIDER 2009, S. 10; vgl. JANSEN & POOCH 2001, S. 243).

## **1.1 Ziele der Untersuchung**

Diese Arbeit geht der Frage nach, ob die vielfach geäußerte Unzufriedenheit von Bibliothekaren mit dem Discoverysystem beluga in Bezug auf einen bestimmten Suchanfragentyp – nämlich die Suche nach einem spezifischen Titel – gerechtfertigt ist. Dafür soll zum einen festgestellt werden, welchen Anteil der Suchanfragentyp „Known-Item-Suche“ an allen Suchanfragen in einem Zeitraum an das System hat. Denn ein häufiges Informationsbedürfnis sollte ein System zufriedenstellend beantworten können. Zum anderen soll herausgefunden werden, wie die Retrievaleffektivität für die Known-Item-Suchanfragen ist, d.h. wie „gut“ das System diese Suchanfragen beantwortet. Das Problem ist, dass Nutzer einfache Suchanfragen formulieren, die in Discoverysystemen vermutlich meist eine große Treffermenge produzieren, auch wenn nur ein bestimmter Treffer gesucht wird. Umso wichtiger ist es, dass dieser auf der ersten Position in der Trefferliste gerankt wird. Wenn die Ergebnisse „schlecht“ sind, muss überlegt werden, wie das System verbessert werden kann. Möglicherweise sind nur bestimmte Arten von Known-Item-Suchanfragen oder Suchen nach bestimmten Medientypen problematisch. Dafür werden die Merkmale der Known-Item-Suchanfragen ermittelt.

Die Fragestellung lautet also: Wie erfolgreich sind Known-Item-Suchanfragen in beluga und – falls notwendig – wie kann die Retrievaleffektivität für diese verbessert werden?

„Erfolgreich“ ist ein Informationssystem, wenn es auf eine Suchanfrage für den Nutzer relevante Treffer liefert, wobei die relevantesten Treffer ganz oben auf der Trefferliste stehen sollten. Da es für Known-Item-Suchen (meist) nur einen relevanten Treffer gibt, lässt sich die Retrievaleffektivität einfach und so gut wie eindeutig feststellen. In dieser Arbeit wird nur ein Suchanfragentyp untersucht, deshalb dürfen Verbesserungsvorschläge und Lösungsansätze keine Auswirkungen auf andere, nicht untersuchte Suchanfragentypen – also größtenteils thematische Suchen – haben. Ein Lösungsweg, um die Retrievaleffektivität für Known-Item-Suchanfragen zu verbessern, ist das System die Known-Item-Suchanfragen mittels eines regelbasierten Algorithmus automatisch erkennen zu lassen. Dann kann für diesen Suchanfragentyp ein angepasstes Relevanzranking verwendet werden. Für solch einen Algorithmus sollen Kriterien gefunden werden, nach denen Suchanfragen sich eindeutig in Suchen nach Known-Items und in „sonstige“/nicht Known-Item-Suchen klassifizieren lassen.

Aus den obigen Überlegungen, dem bekannten Suchverhalten von Nutzern und den Eigenschaften von Discoverysystemen ergeben sich für diese Arbeit folgende Hypothesen:

1. Etwa 50% der Suchanfragen in beluga lassen sich als Suche nach Known-Items identifizieren.
2. Die Known-Item-Suchanfragen werden einfach formuliert: Am häufigsten werden Titelstichwörter gesucht, Metadaten werden selten kombiniert.
3. Die Suchanfragen führen in 50% der Fälle auch bei Known-Item-Suchen zu mehr als 10 Treffern.
4. Ein Drittel der Known-Items steht nicht an erster Position.
5. Eine Mehrzahl der Known-Item-Suchanfragen lässt sich mittels eines regelbasierten Algorithmus identifizieren.

## **1.2 Aufbau der Arbeit**

Die Arbeit ist wie folgt aufgebaut: Im Kapitel 2 werden Aufbau und Funktionen von Discoverysystemen vorgestellt. Dann wird eines der Hauptmerkmale von Discoverysystemen behandelt, nämlich Relevanzranking und welche Rankingfaktoren in Discoverysystemen eingesetzt werden können (Kapitel 3). Es folgt Kapitel 4 zum Suchverhalten von Nutzern in

Bibliothekskatalogen und im Web. Die bisherigen Studien zu Known-Item-Suchanfragen in OPACs und Discoverysystemen werden in Kapitel 5 behandelt. Dann werden die Methodik und die Datenerhebung dieser Arbeit erläutert (Kapitel 6). Kapitel 7 enthält die Auswertung der Ergebnisse der intellektuell klassifizierten Known-Item-Suchanfragen und ihrer Retrievaleffektivität. In Kapitel 8 wird ein möglicher Algorithmus zur Erkennung von Known-Item-Suchanfragen vorgestellt. Nach einer Diskussion der Ergebnisse (Kapitel 9) werden im letzten Kapitel Empfehlungen ausgesprochen und ein Fazit gezogen.

## **2 Discoverysysteme**

Die Erkenntnis, dass Bibliothekskataloge an die Erwartungen und Bedürfnisse der mehrheitlich unerfahrenen Nutzer angepasst werden müssen und nicht andersrum die Nutzer zu Experten im Umgang mit Bibliothekskatalogen geschult werden müssen, sieht Sadeh als Ausgangspunkt für die Entwicklung von Discoverysystemen (vgl. SADEH 2013, S. 3).

Schon seit den 2000er Jahren gab es Bemühungen, neue Kataloge zu entwickeln, die besser an das Rechercheverhalten und die Nutzererwartungen angepasst sind (vgl. SADEH 2013). Mit Katalog 2.0, Next Generation Catalogue, Discoverysystem und Web-Scale-Discovery gibt es verschiedene Bezeichnungen, die austauschbar verwendet werden oder unterschiedliche Aspekte betonen.

Ein Next Generation Catalogue bezeichnet einen Katalog, der die nächste Version einer Entwicklung darstellt und grundlegende Neuerungen enthält (vgl. BARTON & MAK 2012, S. 97). Darunter lassen sich demzufolge auch Katalog 2.0 oder Discoverysystem verstehen.

Ein Katalog 2.0 und ein Discoverysystem haben prinzipiell ähnliche Funktionen und Eigenschaften. Bei der Bezeichnung Katalog 2.0 wird betont, dass der Katalog Elemente des Web 2.0 enthält, die ebenso teilweise auch in OPACs umgesetzt werden können, z.B. Relevanzranking und Kataloganreicherungen (vgl. KNEIFEL 2010, S. 38). Discoverysysteme enthalten ebenso Elemente des Web 2.0, sind aber vom integrierten Bibliothekssystem unabhängige Suchoberflächen mit Suchmaschinentechnologie, die von verschiedenen Anbietern entwickelt werden (vgl. HOFMANN & YANG 2012, S. 256). Discoverysysteme sollen das entdeckende Suchen unterstützen. Dies bezieht sich zum einen auf die Funktionen der Suchoberfläche wie Browsing und Filtern mit Facetten, zum anderen auf den

durchsuchten Inhalt, der nicht nur durch Kataloganreicherungen wie Inhaltsverzeichnisse, Abstracts und Rezensionen, sondern auch durch die Einbindung eines kommerziellen Indexes mit Artikeldaten erweitert werden kann.

Web-Scale-Discovery bezeichnet meist die gleichzeitige Suche in allen Informationen einer Bibliothek, seien es lokale Bestände, Artikeldaten, Bibliotheks-Webseite oder Datenbanken (vgl. BARTON & MAK 2012, S. 98).

In dieser Arbeit wird mit dem Begriff Discoverysystem vorrangig die Suchoberfläche eines Anbieters bezeichnet, unabhängig davon, ob ein kommerzieller Index eingebunden ist oder nicht.

## **2.1 Aufbau und Funktionen eines Discoverysystems**

Laut Barton sind die wichtigsten Anforderungen an einen Katalog der Zukunft ein zentraler Sucheinstieg, eine moderne Weboberfläche und Anbindung an externe Anwendungen (vgl. BARTON & MAK 2012, S. 98 ff.). Discoverysysteme haben den Anspruch, diese drei Anforderungen umzusetzen.

Der zentrale Sucheinstieg wird über einen Index aus den Metadaten der verschiedenen lokalen und nicht-lokalen Bestände einer Bibliothek generiert (vgl. BARTON & MAK 2012, S. 98). Die Daten für den Index können sich aus verschiedenen Quellen zusammensetzen und werden aggregiert indexiert. Dies können lokale Bibliotheksbestände, elektronische Zeitschriften, Datenbanken, Webseiten und Open-Access-Medien aus digitalen Repositorien sein. Bei einer Suchanfrage werden so die verschiedenen Datenbanken nicht wie bei einem Metakatalog einzeln abgefragt, sondern gleichzeitig, da alle Metadaten sich in einem Index befinden. Da die Metadaten von verschiedenen Anbietern stammen (Bibliothek, Verlag, Aggregatoren), besteht jedoch das Problem einer großen Datenheterogenität, sowohl was die Vollständigkeit als auch die Qualität der Metadaten angeht (BARTON & MAK 2012, S. 101).

Der Nutzer kann also über eine Suchoberfläche viele verschiedene Bestände gleichzeitig durchsuchen. Discoverysysteme legen den Schwerpunkt auf eine benutzerfreundliche und intuitiv bedienbare Oberfläche, die an das Rechercheverhalten der Nutzer im Web angepasst sein soll. Dies wird über verschiedene Funktionen unterstützt, die je nach Discoverysystem unterschiedlich vollständig umgesetzt werden, z.B. einfacher Suchschlitz,

Rechtschreibkorrektur, Relevanzsortierung, Suchergebnisfilter mit Facetten der Metadaten und Empfehlungsdienste (vgl. KNEIFEL 2010, S. 42 ff.)<sup>1</sup>.

Die Anbindung an externe Anwendungen ist wichtig, um die Sichtbarkeit des Discoverysystems zu erhöhen und den Nutzer da zu treffen, wo er sich bereits aufhält. Das kann die Integration in Suchmaschinen, in Lernplattformen an Universitäten oder auch in andere Bibliothekssysteme für die Fernleihe sein (vgl. BARTON & MAK 2012, S. 100).

Neben den offensichtlichen Unterschieden bei der Suchoberfläche und dem Inhalt des Index ist einer der größten Unterschiede von Discoverysystemen zu OPACs die verwendete Retrievalmethode. Während der OPAC nach dem Prinzip des „exact match“ sucht, funktioniert Discovery nach dem Prinzip des „best match“. Im OPAC muss der Treffer die Suchbegriffe genau enthalten. Dadurch gibt es eine geringe Fehlertoleranz. Außerdem kann eine spezielle Anfragesprache (z.B. Boole'sche Operatoren) verwendet werden, mit der genau festgelegt werden kann, welche Begriffe in welchem Feld (nicht) vorhanden sein sollen. Das Ziel ist es, entweder bekannte Dokumente zu finden oder eine überschaubare Treffermenge zu generieren. Die Standardsortierung ist nach Chronologie des Erscheinungsjahres.

Dahingegen entsprechen die Treffer im Discoverysystem der Suchanfrage nur so gut wie möglich. Mittels Verfahren wie Stemming, bei dem Wörter automatisch auf ihre Stammform reduziert werden, Kompositazerlegung der indexierten Terme und/oder der Suchanfrage und Synonymwörterbücher werden Discoverysysteme fehlertoleranter gemacht und die Treffermenge erhöht, was jedoch auch zu falschen Treffern führen kann (vgl. MAYLEIN & LANGENSTEIN 2013, S. 193). Das Gewichtungungsverfahren erlaubt eine Sortierung der Trefferliste nach Relevanz, in der auch Treffer enthalten sein können, in denen möglicherweise nicht alle Suchbegriffe enthalten sind (siehe auch Kapitel 3). Das Ziel ist es, eine große Treffermenge zu generieren, bei der die relevantesten Treffer am Anfang der Trefferliste stehen und die durch Filter eingeschränkt werden können (vgl. STEILEN 2012, S. 16).

---

<sup>1</sup> Auf weitere Einzelelemente soll hier nicht eingegangen werden, da sie für diese Untersuchung nicht relevant sind. Eine nahezu vollständige Darstellung von Elementen eines Katalog 2.0 findet sich bei Kneifel (vgl. KNEIFEL 2010).

## 2.2 Anbieter

Bei den Produkten der Anbieter von Discoverysystemen muss zwischen den Komponenten Suchoberfläche, Index und Suchmaschinentechnologie unterschieden werden. Die drei wichtigsten kommerziellen Anbieter von Discoverysystemen sind in Deutschland EBSCO, ExLibris und Serials Solutions (vgl. JANSEN et al. 2010). Eine Open-Source-Lösung ist VuFind.

Die Discoverylösung von EBSCO besteht aus einer eigenen Rechercheoberfläche, dem aggregierten Suchindex EBSCO Discovery Service und einer proprietären Suchmaschinentechnologie von EBSCO. Ein Beispiel für die Verwendung der EBSCO Suchoberfläche und dem Index EBSCO Discovery Service in Deutschland ist das „Wissensportal der SULB“<sup>2</sup> der Saarländischen Universitäts- und Landesbibliothek.

Die Komponenten der Discoverylösung von ExLibris bestehen aus der Suchoberfläche Primo, dem aggregierten Index Primo Central und der Open Source Suchmaschinentechnologie von Solr/Lucene (vgl. JANSEN et al. 2010, S. 12). Ein Beispiel einer Anwendung von Primo und Primo Central in Deutschland ist das „Primus – Suchportal“<sup>3</sup> an der HU zu Berlin.

Serials Solutions bietet den Suchindex Summon mit gleichnamiger Suchoberfläche an. Die Suchmaschinentechnologie ist Solr/Lucene. Im Gegensatz zu den beiden vorherigen Anbietern werden bei Summon zuerst alle Ergebnisse einer Suchanfrage im gesamten Suchindex Summon angezeigt – unabhängig davon, ob die betreffende Einrichtung die Inhalte lizenziert hat. Eine Zugriffsüberprüfung erfolgt erst nach Aufruf des Treffers. In Deutschland werden Suchoberfläche und Suchindex Summon von vielen Bibliotheken eingesetzt. Ein Beispiel ist KonSearch<sup>4</sup> der Universität Konstanz.

VuFind wurde 2006 an der Falvey Memorial Library der Villanova University entwickelt und ist eine Open-Source-Discovery-Suchoberfläche mit Suchmaschinentechnologie auf Basis von Solr/Lucene. VuFind kann zum einen als reiner Bibliothekskatalog mit Discoveryoberfläche verwendet werden, in dem dann also nur die lokalen Bestände recherchierbar sind. Zum anderen kann in VuFind auch ein Suchindex eines kommerziellen Anbieters (also z.B. Primo Central, Summon oder EBSCO Discovery Service) integriert werden (vgl. SCHMITT & STEHLE 2010, S. 23 f.). In Deutschland gibt es inzwischen schon

---

<sup>2</sup> <http://www.sulb.uni-saarland.de/de/>

<sup>3</sup> <http://www.ub.hu-berlin.de/de>

<sup>4</sup> <http://konstanz.summon.serialssolutions.com/de-DE/>

über 30 Bibliotheken mit VuFind-basiertem Discoverykatalog<sup>5</sup>. Die meisten davon funktionieren als Bibliothekskatalog, aber viele haben inzwischen auch schon einen aggregierten Index integriert. VuFind wird auch bei dem hier untersuchten Discoverysystem beluga eingesetzt (siehe Kapitel 6.3).

### **2.3 Darstellungsoptionen zur Integration des Index**

Für die Umsetzung eines Discoverysystems bei einer Einrichtung gibt es verschiedene Möglichkeiten, welche Komponente welchen Anbieters in welcher Kombination verwendet werden. So können zum einen wie oben beschrieben, die Komplettlösungen der Anbieter übernommen werden. Häufig lizenzieren Einrichtungen aber nur den Suchindex eines kommerziellen Anbieters und integrieren ihn dann in ihr vorhandenes Portal bzw. ihre Discoveryoberfläche<sup>6</sup>. Beispiel hierfür ist der Katalog plus!<sup>7</sup> der Universität Bielefeld mit dem EBSCO Discovery Service.

Für die Darstellung der Treffer aus dem kommerziellen Index und den lokalen Beständen existieren zwei verschiedene Ansätze: Zum einen können die Treffer wie bei den oben genannten Beispielen in einer Liste zusammengeführt werden. Vorteile sind hier Übersichtlichkeit, einfaches und gleichzeitiges Filtern aller Treffer und die Konformität mit den Erwartungen der Nutzer. Nachteile sind die mangelnde Sichtbarkeit der lokalen Bestände und – besonders in Bezug auf die Facetten – die Heterogenität der Metadaten. Zum anderen können über Reiter die Treffer in zwei verschiedenen Listen dargestellt werden – eine Variante dieser Lösung beinhaltet eine parallele Zwischenansicht beider Listen, z.B. bei der Lösung Karla II<sup>8</sup> an der Universität Kassel. Für die Reiterlösung sind eine bessere Sichtbarkeit der lokalen Bestände von Vorteil sowie die Möglichkeit, für jede Liste an die vorhandenen Metadaten angepasste Facetten anzubieten. Außerdem können Bibliotheken bei der Reiterlösung weiterhin das Relevanzranking für die lokalen Bestände nach ihren Bedürfnissen anpassen (vgl. MAYLEIN & LANGENSTEIN 2013, S. 190).

---

<sup>5</sup> Für eine (unvollständige) Übersicht über VuFind-Anwender siehe: [https://vufind.org/wiki/installation\\_status](https://vufind.org/wiki/installation_status)

<sup>6</sup> Für eine ausführliche Beschreibung verschiedener Einsatzszenarien siehe JANSEN/KEMNER-HEEK/SCHWEITZER 2010, S. 21 ff.

<sup>7</sup> [http://katalogplus.ub.uni-bielefeld.de/cgi-bin/new\\_search.cgi](http://katalogplus.ub.uni-bielefeld.de/cgi-bin/new_search.cgi)

<sup>8</sup> <https://hds.hebis.de/ubks/index.php>

### 3 Relevanzranking

Ein wichtiges Merkmal von Discoverysystemen ist die Standardsortierung der Trefferliste nach Relevanz. Relevanz ist ein wichtiges Konzept im Information Retrieval: „IR systems [...] are primarily designed to provide potentially relevant information or information objects to people“ (SARACEVIC 2007, S. 1916). Dies gilt also auch für Discoverysysteme: Sie sollen dem Nutzer relevante Informationen auf eine Suchanfrage liefern.

Was Relevanz ist, ist nicht einfach zu bestimmen und hängt von vielen verschiedenen Faktoren ab. Mizzaro definiert Relevanz anhand vier verschiedener Dimensionen, die jeweils verschiedene Ausprägungen haben können, so dass man am Ende auf 36 Ausprägungen von Relevanz kommt. Dabei ist Relevanz immer als Beziehung zwischen zwei Entitäten zu sehen. Die erste Entität entstammt der Gruppe *surrogate, document* oder *information*, die zweite Entität der Gruppe *query, request, information need* oder *problem*. Alle Entitäten werden außerdem von *topic, task* und/oder *context* bestimmt und sind abhängig von der vierten Dimension, nämlich *time* (vgl. MIZZARO 1997, S. 812). Beispielsweise ist ein bibliographischer Datensatz (*surrogate*) in Bezug auf eine thematische Suchanfrage (*query/topic*) relevant, wobei sich das Informationsbedürfnis des Nutzers während des Suchprozesses laufend ändern kann und die Relevanz somit vom Zeitpunkt abhängig ist (*time*).

Saracevic beschreibt ähnliche Beziehungen, berücksichtigt aber als weiteren Aspekt noch das Maß der Relevanzbeziehung (*measure of relatedness*), da Relevanz verschieden stark ausgeprägt sein kann (vgl. SARACEVIC 2007, S. 1918).

Relevanz wird oft zwischen den Polen der Systemsicht und der Nutzersicht gesehen, wobei Relevanz für Systeme als objektiv und für Nutzer als subjektiv beschrieben wird. Ein Informationssystem stellt die Ausprägung einer Übereinstimmung zwischen Suchanfrage und Datensatz fest – größtenteils unabhängig<sup>9</sup> von Kontext und Zeitpunkt im Suchprozess des Nutzers, während der Nutzer die Relevanz eines Treffers daran misst, wie stark die Beziehung ist zwischen seinem Informationsbedürfnis und der Information im Dokument, für das der Datensatz steht – und das im Kontext seiner Aufgabe und dem Verlauf seines Suchprozesses. Aus Nutzersicht besteht die Relevanz eines Treffers also darin, in welchem Maße der Treffer zur Befriedigung seines Informationsbedürfnisses beiträgt. Dieses Informationsbedürfnis

---

<sup>9</sup> Manche Systeme berücksichtigen vorherige Suchanfragen und Umformulierungen, so dass ein gewisser Kontext entstehen kann.



übersetzt der Nutzer in eine Suchanfrage an das System, wobei diese meist nur aus wenigen Wörtern besteht. Das System gleicht die Ähnlichkeit der Suchanfrage mit Inhalten des Datenbestandes ab und berechnet die Relevanz so als Maß der Übereinstimmung zwischen Suchanfrage und Treffern. Das System kann keine Informationen über das subjektive Informationsbedürfnis und Erwartungen des Nutzers an die Treffer berücksichtigen. Zudem kann die gleiche Suchanfrage vielfältige Informationsbedürfnisse verschiedener Nutzer ausdrücken. Das bedeutet auch, dass in Discoverysystemen ein Rankingalgorithmus für möglichst viele verschiedene Informationsbedürfnisse und Suchanfragentypen relevante Ergebnisse liefern muss. Das Ziel des Rankingalgorithmus ist es, die Relevanz anhand objektiver Kriterien zu bestimmen (vgl. BADE 2007, S. 837). Nichtsdestotrotz liegen einem Rankingalgorithmus immer Vorannahmen über Erwartungen und Informationsbedürfnisse des Nutzers an das System zugrunde, aufgrund derer ein Entwickler sich für oder gegen die Aufnahme von bestimmten Rankingfaktoren in den Algorithmus entschieden hat (vgl. HJØRLAND 2010, S. 218). Im folgenden Kapitel sollen nun Rankingfaktoren für bibliographische Metadaten vorgestellt werden.

### **3.1 Rankingfaktoren in einem Discoverysystem**

Es existieren verschiedene Faktoren, anhand derer der Rankingalgorithmus eines Discoverysystems die Relevanz eines Dokumentes bewerten kann. Die Faktoren lassen sich einordnen in die Bereiche Textstatistik, Popularität, Aktualität und Lokalität (vgl. LEWANDOWSKI 2010, S. 98).

Das grundsätzliche Verfahren zur Bestimmung der Relevanz eines Dokuments ist die Berechnung der Ähnlichkeit einer Suchanfrage mit den Ergebnisdokumenten. Dies ist ein textstatistisches Verfahren, bei dem zum einen das Vorkommen der Suchwörter in den Ergebnissen (*term frequency*/TF) sowie die Häufigkeit der Suchwörter im gesamten Datenbestand berücksichtigt werden (*inverse document frequency*/IDF) (vgl. BAEZA-YATES 2011, S. 66 ff.). Das Problem ist, dass dieses Verfahren vor allem für Volltextdatenbanken entwickelt wurde. Bibliographische Datensätze enthalten sehr viel weniger Text und sind in ihrer Textbasis zudem noch unterschiedlich lang, vor allem wenn manche Titel mit Abstracts oder Inhaltsverzeichnissen angereichert sind. Daher ist eine Beschränkung auf das TF/IDF-Verfahren in Bibliothekskatalogen nicht sinnvoll (vgl. DELLIT & BOSTON 2007, S. 5)(vgl. LEWANDOWSKI 2010, S. 101). Da die Informationen eines bibliographischen Datensatzes in Feldern strukturiert ist, kann zusätzlich das Vorkommen eines Suchbegriffes je nach Feld

unterschiedlich gewichtet werden. Mögliche Grundsätze für die Feldgewichtung nennen Dellit und Boston (vgl. DELLIT & BOSTON 2007, S. 6):

- Übereinstimmungen in Feldern der bibliographischen Beschreibung: Vor allem Titel, Autor und Schlagwörter werden höher gewichtet, ebenso wenn die Suchbegriffe in mehreren dieser Felder vorkommen.
- Grad der Übereinstimmung von Suchbegriffen mit einem Feld: Eine exakte und vollständige Übereinstimmung wird höher gewichtet als eine Teilübereinstimmung als Phrase oder als einzelnes Wort. Dabei wird dieses Prinzip unterschiedlich stark auf die Felder im vorherigen Prinzip angewandt.
- Art des Feldes: Übereinstimmungen mit Hauptfeldern sind höher gewichtet als Übereinstimmungen mit anderen Feldern.
- Übereinstimmung des Suchbegriffs: Die exakte Übereinstimmung mit dem Suchbegriff wird höher gewertet als Varianten des Suchbegriffs, die durch Stemming oder Rechtschreibkorrektur gefunden werden.
- Wenn durch die oberen Faktoren keine Rangfolge in der Gewichtung erreicht werden kann, werden Treffer aus bestimmten Sammlungen oder mit höherer Exemplarzahl höher gewichtet.

Die textstatistischen Verfahren gleichen also nur die Suchanfrage und die Ergebnisdokumente miteinander ab und liefern darüber hinaus keinen Hinweis auf die Qualität der Ergebnisse. Es wurden schon einige andere Faktoren vorgeschlagen, anhand derer die Relevanz eines Titels für einen Nutzer steigen könnte (vgl. LEWANDOWSKI 2010, S. 96).

Popularität ist ein Faktor, der anhand verschiedener Kriterien feststellbar ist. So können viele Auflagen oder viele Exemplare in Bestand bzw. viele besitzende Bibliotheken Hinweis auf Standardwerke sein. Außerdem lassen sich aus Nutzungsstatistiken im Katalog oder aus Zitationsstatistiken Rückschlüsse auf die Popularität eines Titels ziehen. Oder aber es werden im Voraus Titeln bestimmter Autoren, Verlagen oder Buchreihen ein bestimmter Popularitätsfaktor zugeschrieben (vgl. LEWANDOWSKI 2010, S. 101). In manchen Katalogen werden auch Bewertungen durch Nutzern zugelassen. Solch ein user-generated-content wird ebenso wie Tags inzwischen allerdings aufgrund der fehlenden Beteiligung und der geringen Anzahl an Nutzern kritisch gesehen (vgl. CHRISTENSEN 2009, S. 33 ff.). Zuletzt kann auch aus Ausleihstatistiken ein Rückschluss auf die Popularität von Titeln bei den Nutzern gezogen werden (vgl. DELLIT & BOSTON 2007, S. 10).

Die Aktualität von Titeln wird in den OPACs ohne Relevanzranking meist als Standardsortierung für die Trefferliste herangezogen. Als Faktor im Relevanzranking kann die Aktualität dafür sorgen, dass neuere Auflagen vor älteren Auflagen bevorzugt werden, was auch bei Known-Item-Suchanfragen sinnvoll ist. Außerdem muss beachtet werden, dass es je nach Fachrichtung Unterschiede bei der Bevorzugung von aktuellen Titeln geben kann (vgl. LEWANDOWSKI 2010, S. 101). Als weiterer Aspekt kann das Anschaffungsdatum im Ranking berücksichtigt werden (vgl. LEWANDOWSKI 2010, S. 100).

Auch kann der Nutzungskontext des Nutzers insofern berücksichtigt werden, als dass sein physischer Standort und der des Buches sowie der Ausleihstatus bzw. die elektronische Verfügbarkeit miteinander abgeglichen werden. Zuletzt können auch manche Medientypen stärker gewichtet werden, der Umfang eines Titels berücksichtigt werden oder für bestimmte Nutzergruppen ein Relevanzranking angepasst werden (vgl. LEWANDOWSKI 2010, S. 100).

Die Rankingfaktoren können zudem danach unterschieden werden, ob sie statisch oder dynamisch zur Anwendung kommen. Statische Maßnahmen sind von der jeweiligen Suchanfrage unabhängig und werden beim Aufbau des Indexes festgelegt. Dazu gehört das Boosten von Dokumenten, die bestimmte Eigenschaften aufweisen. Das können beispielsweise Zeitschriften, Auflagenwerke, Erscheinungsjahr, online verfügbare Dokumente und Anzahl der besitzenden Bibliotheken sein (vgl. LANGENSTEIN & MAYLEIN 2009, S. 410). Mit dynamischen Maßnahmen wird für jede Suchanfrage der Einfluss der Parameter auf das Ranking berechnet. Dies bezieht sich auf das Boosten von bestimmten Feldern und Übereinstimmungsarten der Suchanfrage mit den Feldern. Beispielsweise werden Treffer mit Einworttitel oder Phrasen bzw. nah beieinander stehende Wörter bevorzugt (vgl. LANGENSTEIN & MAYLEIN 2009, S. 410).

Zusammenfassend kann man feststellen, dass neben den textstatistischen Verfahren zum einen Daten, die durch den Nutzer in Bezug auf ein Dokument generiert werden (Ausleihzahlen, Downloads, Zitationshäufigkeit), zum anderen Daten, die auf Annahmen über die Präferenzen von Nutzern (Aktualität, Verfügbarkeit) beruhen, genutzt werden können. Nutzergenerierte Faktoren werden kritisch gesehen, da hier die Gefahr der „Bevorzugung des immergleichen“ (LEWANDOWSKI 2010, S. 103) durch den Algorithmus besonders hoch ist.

Die einzelnen Faktoren müssen an die Voraussetzungen eines Bibliothekskatalogs angepasst werden in Hinsicht auf ihre Nutzergruppe, die fachliche Ausrichtung und den Datenbestand (vgl. MAYLEIN & LANGENSTEIN 2013, S. 190). Roscher stellt dar, aufgrund welcher

Überlegungen sich bei verschiedenen Discoverysystemen für oder gegen manche Rankingfaktoren entschieden wurde, die auf Vorannahmen über den Nutzer beruhen. Es wird deutlich, dass die meisten Discoverysysteme das Boosten von ein oder mehreren Kriterien wie Jahr, lokaler Bestand, Elektronische Dokumente/Verfügbarkeit, Urheberliteratur und Sprache einsetzen, Popularitätsfaktoren jedoch nur von einem Discoverysystem genutzt werden (vgl. ROSCHER 2014, S. 42 ff.). Genaue Angaben zu der Einstellung der verschiedenen Parametern gibt es zu HEIDI<sup>10</sup>, dem Discoverysystem der Universität Heidelberg, das die Suchmaschinentechologie von Solr/Lucene für die lokalen Bestände verwendet (vgl. LANGENSTEIN & MAYLEIN 2009; MAYLEIN & LANGENSTEIN 2013).

Mit den verschiedenen Einstellungsmöglichkeiten bei den Parametern des Boostens können in beschränktem Maße Suchen nach Known-Items unterstützt werden. Das Auffinden von bestimmten Titeln wird durch das Boosten von Einworttiteln und die Bevorzugung von Dokumenten, deren Titel die Suchwörter als exakte oder ähnliche Phrase enthalten, erleichtert (vgl. MAYLEIN & LANGENSTEIN 2013, S. 192). Zeitschriften lassen sich besser dadurch finden, dass der Dokumenttyp geboostet wird. Außerdem können in Solr/Lucene mit dem QueryElevationComponent Suchanfragen festgelegt werden, für die immer ein bestimmter Treffer an erster Stelle gerankt werden soll (vgl. MAYLEIN & LANGENSTEIN 2013, S. 198). Bis auf die letzte Maßnahme werden durch die Einstellungen immer auch Auswirkungen auf das Ranking von thematischen Suchen die Folge sein.

### **3.2 Relevanzbewertung**

Den Nutzer beeinflussen auf verschiedenen Ebenen vielfältige Kriterien bei der Bewertung der Relevanz von Treffern: persönliche Voraussetzungen des Nutzers, Inhalt der Anfrage, Eigenschaften des Dokuments, Eigenschaften des Systems, Bewertungsbedingungen und Bewertungsskala (insgesamt 80 Faktoren nach SCHAMBER in HJØRLAND 2010, S. 218). Die wenigsten Faktoren können von einem Rankingalgorithmus berücksichtigt werden, da sie zum einen von Nutzer zu Nutzer und von Suchanfrage zu Suchanfrage sehr unterschiedlich ausgeprägt sein können.

Um einen Rankingalgorithmus evaluieren zu können, müssen die relevanten Dokumente auf eine Suchanfrage bekannt sein, bzw. es muss festgestellt werden können, wie relevant die gefundenen Dokumente für die Suchanfrage sind. Daher werden Juroren eingesetzt, die

---

<sup>10</sup> [heidi.ub.uni-heidelberg.de/](http://heidi.ub.uni-heidelberg.de/)

anhand der Suchanfrage und einer kurzen Beschreibung des Informationsbedürfnisses die Ergebnisse in zufälliger Reihenfolge auf einer Skala nach Relevanz bewerten sollen. Ein Dokument muss möglichst von mehreren Juroren bewertet werden, um eine möglichst objektive Aussage treffen zu können (vgl. LEWANDOWSKI 2011a, S. 211).

Dennoch gibt es Suchanfragen, bei denen nahezu objektiv und unabhängig vom Kontext des Nutzers eindeutig die Relevanz eines Ergebnisses bewertet werden kann. Dies sind Suchanfragen, die als gewünschtes Ergebnis eine spezifische Webseite bzw. eine bibliographische Titelaufnahme haben. Umso wichtiger ist es, dass die gewünschte Webseite/das Known-Item auch an Rang 1 positioniert werden. Lewandowski nimmt an, dass die Nutzerakzeptanz für ein System auch stark von der Retrievaleffektivität für Suchanfragen mit eindeutigem Trefferbewertung abhängt, da die Ergebnisse sofort als gut oder schlecht bewertet werden können (vgl. LEWANDOWSKI 2010, S. 96; LEWANDOWSKI 2014, S. 10).

## **4 Suchverhalten**

In einem Bibliothekskatalog – seien es OPACs oder Discoverysysteme – sind anhand der formalen und inhaltlichen Erschließung die Daten einer Titelaufnahme durchsuchbar und auffindbar.

Die Benutzeroberfläche eines OPACs ist daraufhin ausgelegt, dass der Nutzer die komplexen Recherchemöglichkeiten in der Katalogdatenbank der Bibliothek fast genau so vollständig ausschöpfen kann wie die Experten. Je nach System können Suchkategorien ausgewählt werden, wie z.B. Autor, Titel, Stichwort oder Schlagwort, die in der erweiterten Suche auch kombinierbar sind. Die meisten OPACs verwenden als Standardeinstellung eine Suche über alle Felder. Mithilfe der Suchkategorien und Boole'schen Operatoren sowie unter Kenntnis der zugrundeliegenden Metadaten kann präzise in den verschiedensten Feldern nach Themen bzw. Known-Items gesucht werden. Da die Ergebnisse meist nach Jahr oder Autor sortiert werden, ist es je nach Suchintention zweckmäßig, eine überschaubare Treffermenge zu genießen.

Studien zum Suchverhalten von Nutzern in OPACs zeigen jedoch, dass von Nicht-Experten einerseits die Funktionalitäten kaum ausgeschöpft werden und andererseits die Suchanfragen an sich einfach gehalten sind. So verwenden Nutzer bevorzugt die Standardeinstellung, die

ihnen auf der Startseite angeboten werden (vgl. SCHNEIDER 2009, S. 12; vgl. HENNIES & HOCHSTEIN 2006, S. 7). Dies ist meist eine Freitextsuche über alle Felder, wie z.B. in den OPACs des GBV<sup>11</sup>. Wenn auf der Startseite mehrere Suchfelder zur Verfügung stehen, so wird trotzdem überwiegend nur ein Suchfeld ausgefüllt (80,2% der Suchanfragen bei HENNIES & HOCHSTEIN 2006, S. 7). Fortgeschrittenere Recherchemethoden wie Verwendung von Operatoren oder Trunkierung von Suchwörtern werden äußerst selten genutzt und bewegen sich im unteren einstelligen Prozentbereich (vgl. HENNIES & HOCHSTEIN 2006, S. 7 f.; JANSEN & POOCH 2001, S. 243; SCHNEIDER & BUGNON 2008, S. 19). Suchanfragen bestehen zum Großteil aus ein bis zwei Wörtern (68,2% bei HENNIES & HOCHSTEIN 2006, S. 7; 75% bei JANSEN & POOCH 2001, S. 243; ähnlicher Anteil bei SCHNEIDER 2009, S. 12). Viele OPACs verfügen über Stoppwortlisten mit (nicht allen der) deutschen, englischen und französischen Artikeln, Pronomen, Adverbien und Präpositionen. Diese Wörter werden bei Suchanfragen ignoriert, da sie kaum bedeutungstragend sind. Dennoch sind bei Wolfram die acht meistgesuchten Wörter Artikel, Konjunktionen und Präpositionen (vgl. WOLFRAM 2008, S. 1286). Die Suche in einem OPAC birgt für den Nutzer viele Stolperfallen. Jones et al. berichten eine Fehlerquote von 30% bei der korrekten Formulierung von Suchanfragen (vgl. JONES et al. 2000, S. 159). Dementsprechend hoch ist der Anteil an Nulltreffer-Suchen (27% bei HENNIES & HOCHSTEIN 2006, S. 10; 30% bei SCHNEIDER 2009, S. 13). Schneider ermittelte, dass von diesen 30% Nulltreffer-Suchen 41% am inexistenten Bestand der Bibliothek liegen (also Known-Item-Suchen waren) und 55% bei thematischen Suchen auftreten oder aufgrund von Rechtschreibfehlern passieren (vgl. SCHNEIDER 2009, S. 13). Die erfolgreichen Suchen stellen den Nutzer eher vor das Problem, dass er mit wenigen Suchwörtern – die teils auch Stoppwörter sind – im Stichwort- bzw. Freitextfeld eine relativ große Treffermenge erzielt. Die wenigsten Nutzer sind aber bereit, eine große Menge an Treffern zu sichten. Dies lässt sich an der Anzahl der aufgerufenen Trefferseiten und des Rangs der aufgerufenen Vollanzeigen ablesen. Hennies und Hochstein stellen fest, dass nur 11% der aufgerufenen Vollanzeigen hinter dem zehnten Treffer liegen (vgl. HENNIES & HOCHSTEIN 2006, S. 10). Zusammenfassend lässt sich feststellen, dass Nutzer die Möglichkeiten des OPACs als präzises Rechercheinstrument kaum ausschöpfen bzw. Probleme mit der korrekten Formulierung komplexerer Suchanfragen haben.

Während also für den erfahrenen Nutzer der Vorteil des OPACs in seiner Eigenschaft als präzisiertem Rechercheinstrument liegt, haben unerfahrene Nutzer – also der Großteil –

---

<sup>11</sup> Wie beispielsweise beim Campus-Katalog der SUB Hamburg <https://kataloge.uni-hamburg.de/>.

Schwierigkeiten mit einer effektiven Benutzung. Besonders dem Suchverhalten der zweiten Gruppe an Nutzern sollen Discoverysysteme entgegenkommen, indem ein einfacher Suchschlitz, Relevanzsortierung, fehlertolerante Verarbeitung der Suchanfrage, Rechtschreibvorschläge, Filter und Titelempfehlungen die Suche vereinfachen sollen, so dass der Nutzer „wie bei Google“ suchen kann und relevante Ergebnisse erhält. In Studien zum Suchverhalten im Web zeigt sich, dass Nutzer nicht bereit sind, aufwendig zu suchen. Suchanfragen sind kurz und nur wenige Ergebnisse werden auf relevante Treffer durchgeschaut. So bestehen englischsprachige Suchanfragen zu über 50% aus 1-2 Wörtern (vgl. JANSEN & POOCH 2001, S. 241). Mit einem Durchschnitt von 1,6 bis 1,8 Wörtern sind deutschsprachige Suchanfragen sogar kürzer als englischsprachige (vgl. HÖCHSTÖTTER 2007, S. 138). Boole'sche Operatoren und Phrasensuche werden kaum verwendet (vgl. HÖCHSTÖTTER 2007, S. 138). Bei mehr als der Hälfte der Suchanfragen wird nur die erste Trefferseite angeschaut und bei knapp drei Vierteln aller Suchanfragen werden nur zwei Treffer aufgerufen (vgl. JANSEN & SPINK 2003, S. 67). Jansen und Spink sehen in ihren Ergebnissen Hinweise darauf, dass die meisten Nutzer nur wenig Zeit für die Sichtung der Treffer aufbringen und auf eine Suchanfrage nur zwei Treffer ansehen müssen, um ein relevantes Dokument zu finden (vgl. JANSEN & SPINK 2003, S. 68).

Die Suchanfragen, die ein Nutzer aufgrund seines Informationsbedürfnisses formuliert, lassen sich nach Broder in der Web-Suche auf drei Intentionen zurückführen, nämlich eine bestimmte Webseite zu erreichen (*navigational queries*), Informationen in den Suchergebnissen zu finden (*informational queries*) sowie weitere Handlungen im Web durchzuführen (*transactional queries*) (vgl. BRODER 2002, S. 5). Von navigationsorientierten Suchen sind Nutzer gewöhnt, dass der relevanteste Treffer auf Rang 1 steht. So zeigen Studien zum Klickverhalten, dass Suchmaschinen auf navigationsorientierte Suchanfragen so gute Ergebnisse liefern, dass meist der erste Treffer angeklickt wird (vgl. LEWANDOWSKI et al. 2012, S. 24). Lewandowski sieht die gute Retrievaleffektivität von Google für navigationsorientierte Suchanfragen auf den ersten drei Rängen auch als einen Erklärungsansatz dafür, dass Google die beliebteste Suchmaschinen ist (LEWANDOWSKI 2011b, S. 360).

## 5 Die Known-Item-Suche

Im Kontext von Bibliothekskatalogen entsprechen navigationsorientierte Suchanfragen den Known-Item-Suchanfragen, informationsorientierte Suchanfragen den thematischen Suchanfragen und transaktionsorientierte Suchanfragen der Suche nach einer Quelle (vgl. LEWANDOWSKI 2010, S. 94).

In Einführungswerken zur Bibliothekswissenschaft werden meist nur die ersten beiden Typen von Suchen unterschieden: die „Titelsuche, d.h. die Suche nach einer Publikation, deren Titeldaten genau oder teilweise bekannt sind“ (GANTERT & HACKER 2008, S. 205) und die „sachliche Suche nach Literatur zu einem bestimmten Thema“ (GANTERT & HACKER 2008, S. 207). Chowdhury spricht von „known-item search“ bzw. „unknown-item search“ (vgl. CHOWDHURY et al. 2008, S. 131), was verdeutlicht, dass nicht alle Suchanfragen außerhalb der *known-item searches* mit thematischen Suchanfragen gleichzusetzen sind.

Wie es der englische Begriff *known-item search* schon verrät, zeichnet sich die Titelsuche dadurch aus, dass der Nutzer weiß, dass ein bestimmtes Medium existiert, er Titelinformationen zu diesem Medium hat und im Bibliothekskatalog prüft, ob dieses Medium im Bestand der Bibliothek nachgewiesen ist. Während es bei der thematischen Suche mehrere relevante Treffer geben kann (z.B. alle Bücher, die Fragestellung xy behandeln), so gibt es bei der Known-Item-Suche normalerweise nur einen spezifischen Treffer<sup>12</sup>.

Einige Studien beschäftigen sich vorrangig mit der Known-Item-Suche, wobei hier die unterschiedlichsten Fragestellungen behandelt werden: An welche Art von (bibliographischen) Informationen erinnern sich Nutzer bei der Suche nach Known-Items? (vgl. ALLEN 1989); Welche Fehlerquellen gibt es bei erfolglosen Suchen nach Known-Items? (vgl. DWYER et al. 1991); Anhand welcher bibliographischen Informationen werden Known-Items am besten gefunden? (vgl. KILGOUR 1995 2001 2004; KILGOUR et al. 1999; KILGOUR & MORAN 2000); Wie lassen sich Known-Item-Suchanfragen automatisch erkennen? (vgl. KAN & POO 2005); Welches konzeptionelle und operationelle Verständnis haben Autoren von Known-Item-Suchen? (vgl. LEE et al. 2006); Wie formulieren Nutzer Suchanfragen nach Known-Items? (vgl. KREBS 2013). Einen Überblick über verschiedene Studien und ihre Ergebnisse geben Wildemuth und O'Neill für den Zeitraum bis in die 90er Jahre des letzten Jahrhunderts (vgl. WILDEMUTH & O'NEILL 1995).

---

<sup>12</sup> Je nach Spezifität der Suche können Werke mit mehreren Auflagen oder mit verschiedenen Format Ausnahmen bilden.



Außerdem wird die Known-Item-Suche oft im Kontext von Studien zum Suchverhalten von Nutzern in OPACs (z.B. HAMILTON & THURLOW 2005; MALLIARI & KYRIAKI-MANESSI 2007; MCKAY & BUCHANAN 2011; SLONE 2000) oder in Discoverysystemen (z.B. HESSEL & FRANSEN 2012; MEADOW & MEADOW 2012; SLAVEN et al. 2011) betrachtet.

Im Folgenden werden die unterschiedlichen Herangehensweisen zur Identifizierung von Known-Item-Suchen angerissen und dann die Ergebnisse aus bisherigen Studien zu Known-Item-Suchanfragen in OPACs und Discoverysystemen zusammengefasst.

### **5.1 Identifizierung von Known-Item-Suchanfragen**

Wortwörtlich gesehen ist die Known-Item-Suche die Suche nach einem (dem Nutzer) bekannten Objekt. Es existieren aber verschiedene Ansätze, was genau als Known-Item-Suche gesehen wird. Lee et al. bemängeln, dass es in der Literatur noch kein vollständiges Verständnis vom Konzept der Known-Item-Suche gibt (vgl. LEE et al. 2006). Sie unterscheiden zwei Perspektiven, die in der Diskussion von Known-Item-Suchen meist eingenommen werden: zum einen die Sichtweise auf das Konzept der Known-Item-Suche und zum anderen die Umsetzung von Known-Item-Suchen in einem bestimmten System (LEE et al. 2006, S. 2). Die Autoren untersuchen auf der Ebene des Konzepts, welcher Art die Beziehung zwischen dem Nutzer und dem bekannten Objekt ist und diskutieren mehrere Fragen, unter welchen Voraussetzungen man von einer Known-Item-Suche sprechen kann: Muss der Nutzer wissen, dass ein Objekt existiert, oder reicht es, wenn er vermutet oder glaubt, dass ein Objekt existiert? Und muss das Objekt überhaupt existieren? (vgl. LEE et al. 2006, S. 3 ff.). Eine Sichtweise kann sein, dass nur solche Suchen Known-Item-Suchen sind, bei denen das Objekt auch existiert. Wenn aber nur die Erfahrung des Nutzers wichtig ist, dann ist eine „known-item search [...] a characteristic of the user’s cognitive state“ (LEE et al. 2006, S. 4). Die „Geistesverfassung“ oder die Absicht des Nutzers bei einer Suchanfrage macht also eine Known-Item-Suche aus, unabhängig davon, ob das Known-Item existiert oder nicht.

In den meisten Studien wird die Known-Item-Suche aus der zweiten Perspektive untersucht, nämlich wie die Known-Item-Suche in einem Katalog umgesetzt wird. Eine Schwierigkeit bei der Untersuchung von Known-Item-Suchanfragen liegt in der verlässlichen Identifikation derselben, da sich ein Informationsbedürfnis mit vielen verschiedenen Suchanfragen ausdrücken lässt und umgekehrt einer Suchanfrage je nach Nutzer verschiedene

Informationsbedürfnisse zugrunde liegen können. Je nach verwendeter Methode kann die Known-Item-Suche unterschiedlich weit gefasst werden.

Eine Methode zur Erforschung von Informationsbedürfnissen sind Umfragen, bei denen die Nutzer vor oder nach der Katalogbenutzung zu Motiven und Ziel der Katalogsuche befragt werden. Dabei können auch Suchstrategien und Suchanfragen abgefragt werden. Der Vorteil dieser Methode ist, dass der Nutzer selbst Auskunft geben kann und nutzerbezogene Daten wie Alter, Geschlecht, Beruf/Tätigkeit oder Fachbereich erhoben werden können. Der Nachteil liegt zum einen in der relativ geringen Erhebungszahl und darin, dass Nutzer oft nicht korrekt das zugrundeliegende Informationsbedürfnis benennen können (vgl. LEWANDOWSKI et al. 2012, S. 33). Nach dieser Sichtweise sind Known-Item-Suchanfragen alle, die aus der Absicht – dem „state of mind“ – des Nutzers gestellt wurden, ein Known-Item zu finden. Das bedeutet, dass auch Suchen mit Schlagwörtern oder andere Recherchestrategien als Known-Item-Suche verstanden werden (vgl. LEE et al. 2006, S. 7).

Des Weiteren können die Suchanfragen aus den Logdateien klassifiziert werden. Ein Vorteil von Logdateianalysen ist, dass die Gesamtheit aller Suchanfragen untersucht werden kann. Allerdings liegen keine Hintergrundinformationen zu den einzelnen Nutzern vor, so dass nur bedingt auf die Suchintention rückgeschlossen werden kann. Je nach System stehen aber aus den Logdateien verschiedene Informationen für eine Identifikation der Known-Item-Suchanfragen zur Verfügung. In OPACs können die Suchanfragen anhand der verwendeten Suchkategorien eingeteilt werden. Gantert und Hacker benennen für die Known-Item-Suche als Suchmöglichkeiten Titelstichwörter, Phrasensuche im Titel, Personennamen, Körperschaftsnamen, Erscheinungsort, Verlag, ISBN, ISSN und Signatur sowie für die sachliche Suche Schlagwörter, Stichwörter und Notationen (vgl. GANTERT & HACKER 2008, S. 205 ff.). Auch Chowdhury et al. sehen die Verwendung von solchen Suchschlüsseln als Hinweis dafür, dass eine „known-item search“ bzw. eine „unknown-item search“ durchgeführt wird (vgl. CHOWDHURY et al. 2008, S. 131). Dementsprechend werden in den Studien, die Known-Item-Suchanfragen in einem OPAC untersuchen, meist alle Suchanfragen, in denen diese Suchkategorien verwendet werden, als Known-Item-Suchanfragen bezeichnet – unabhängig davon, ob ein konkretes Known-Item auch identifizierbar ist.

In Discoverysystemen ist solch eine Definition von Known-Item-Suchanfragen nicht möglich, da hier im einfachen Suchschlitz eine Freitextsuche ausgelöst wird und Suchkategorien nicht als Auswertungsmerkmal zur Verfügung stehen. Daher muss in Discoverysystemen die

Suchanfrage anhand anderer Merkmale klassifiziert werden. Eine Möglichkeit ist, die Known-Item-Suchanfragen intellektuell anhand ihrer Eigenschaften zu identifizieren. In diesen Studien ist meist Voraussetzung für eine Klassifizierung als Known-Item-Suchanfrage, dass das konkrete Known-Item erkannt werden kann, um einigermaßen zweifelsfrei auf eine Known-Item-Suche rückschließen zu können: „known item searches were for specific and identifiable texts“ (MCKAY & BUCHANAN 2011, S. 266). Dafür müssen zwischen der Suchanfrage und dem Known-Item gewisse Übereinstimmungen vorhanden sein, wobei hier wieder die Wörter der Suchanfrage mit formalen Feldern der Titelaufnahme verglichen werden – z.B. Titel und/oder Autor: Known-Item-Suchanfragen sind „Queries in which the patron entered the title of something that was recognized as a unique, identifiable object“ (CHAPMAN et al. 2013, S. 411).

Im Kontext von Web-Suchen haben verschiedene Studien gezeigt, dass die Anfragetypen anhand bestimmter Eigenschaften und Merkmale nicht nur intellektuell, sondern auch automatisch mit Verfahren des maschinellen Lernens klassifiziert werden können (siehe BRENES et al. 2009; ZAMORA et al. 2014 für einen Überblick über die möglichen Verfahren). Für Bibliothekskataloge ist nur eine solche Studie bekannt (siehe KAN & POO 2005 im nächsten Kapitel).

## **5.2 Die Known-Item-Suche im OPAC**

Es lässt sich vermuten, dass OPACs gut dafür geeignet sind, etwas zu finden, wenn man genau weiß, in welchen bibliographischen Feldern welche Begriffe stehen. Dies gilt besonders für Known-Items, da man zumindest einige bibliographische Angaben hat, die man einer Suchkategorie in einem OPAC zuordnen kann. Wie jedoch das Kapitel 4 gezeigt hat, hat der durchschnittliche Nutzer Schwierigkeiten, korrekte Suchanfragen nach den Regeln des Systems und entsprechend der bibliographischen Erschließung zu stellen.

Die meisten Studien zum Suchverhalten in OPACs definieren Suchen in den Feldern Autor, Titel, Signatur, ISBN und ISSN als Known-Item-Suchen (KREBS 2013; MALLIARI & KYRIAKI-MANESSI 2007; SLONE 2000; WILDEMUTH & O'NEILL 1995). Suchen nach ISBN und ISSN werden meist Mitarbeitern der Bibliothek zugeschrieben und deshalb nicht berücksichtigt. Bei Malliari und Kyriaki-Manessi liegt der Gesamtanteil von Known-Item-Suchen bei 68% (vgl. MALLIARI & KYRIAKI-MANESSI 2007, S. 116), ebenso bei Krebs (vgl. KREBS 2013, S. 93). Wildemuth und O'Neill berichten für verschiedene Studien eine hohe

Spannweite des Anteils an Known-Item-Suchanfragen von 25% bis 79%. In Umfragen liegt der Anteil der Known-Item-Suchen mit einer Spanne von 33% bis 67% etwas niedriger (vgl. WILDEMUTH & O'NEILL 1995, S. 268). Diese großen Unterschiede in der Nutzung von OPACs lassen vermuten, dass das Suchverhalten von vielen Faktoren beeinflusst wird und Ergebnisse kaum verallgemeinerbar sind. Es ist zumindest feststellbar, dass Known-Item-Suchanfragen nach dieser formalen Betrachtungsweise in den meisten OPACs einen bedeutenden Anteil – oft sogar den Hauptanteil – ausmachen.

Laut den genannten Studien sind folgende Merkmale für das Suchverhalten nach Known-Items charakteristisch:

Nutzer, die nach Known-Items suchen, sind in ihrem Recherchevorgehen sehr ausdauernd und formulieren mehrmals neue Suchanfragen, um das gewünschte Known-Item zu finden (vgl. KREBS 2013, S. 97; vgl. SLONE 2000, S. 764).

Misserfolge bei Known-Item-Suchen werden meist anhand der Null-Treffer-Suchen erforscht. Dabei wird selten berücksichtigt, ob null Treffer auftraten, weil das Known-Item nicht im Bestand war, oder weil die Suchanfrage fehlerhaft formuliert wurde. Misserfolg ist auf Rechtschreibfehler (vgl. SLONE 2000, S. 763; vgl. WILDEMUTH & O'NEILL 1995, S. 269), auf Schwierigkeiten, die bibliographischen Informationen so zu formulieren, wie sie im Katalog vorliegen sowie auf die Verwendung einer falschen Suchkategorie oder die Nichtbeachtung der Syntax des Systems zurückzuführen (vgl. KREBS 2013, S. 97 ff.; WILDEMUTH & O'NEILL 1995, S. 269 f.). Zusammengenommen sind etwa 25% der Known-Item-Suchanfragen laut der Studien nicht erfolgreich (vgl. WILDEMUTH & O'NEILL 1995, S. 270).

An bibliographischen Informationen wussten Nutzern bei der Suche nach Büchern in 91% der Fälle den Titel (oder Teile davon), in 77% den Autor, in 56% das Erscheinungsjahr, in 42% den Verlag und in 32% das Thema. Bei Zeitschriften kannten die Nutzer neben dem Titel (100%) Erscheinungsjahr (92%) und Seitenzahlen (80%) (vgl. WILDEMUTH & O'NEILL 1995, S. 274).

Problematisch ist bei diesen Studien, dass die Verwendung einer bestimmten Suchkategorie nicht automatisch einen Rückschluss auf die Suchintention des Nutzers zulässt, was den meisten Autoren bewusst ist. Auch wenn es aus bibliothekarischer Sicht am zielführendsten ist, die Felder der formalen Erschließung für die Known-Item-Suche und die inhaltlichen Suchkategorien für die thematische Suche zu verwenden, so spiegelt sich dies – wie in Kapitel 4 beschrieben – nicht im Suchverhalten der Nutzer wider: Meist wird die Standardeinstellung

verwendet, die beispielsweise im OPAC des GBV eine Suche über alle Felder ist. Die Suchanfragen aus diesem Feld werden jedoch in keiner der Studien weiter daraufhin untersucht, zu welchem Suchanfragentyp sie gehören. Hinzu kommt, dass eine Suchanfrage in der Titel- oder Autorkategorie durchaus eine Suche zu einem Thema sein kann, weil der Nutzer vermutet, dass der Titel Hinweise auf den Inhalt eines Mediums enthält oder ein Autor mehrere Bücher zu einem Thema geschrieben hat. So berichten Wildemuth und O'Neill von verschiedenen Studien, in denen bis zu 43 % aller Autor- bzw. Titelsuchen eigentlich als thematische Suchen zu werten seien. Ebenso ist es möglich – wenn mit 6% auch weniger häufig, dass ein Known-Item über Schlagwörter gesucht wird (vgl. WILDEMUTH & O'NEILL 1995, S. 266 f.).

Die intellektuelle Klassifikation von Known-Item-Suchanfragen an einen OPAC beschreiben Kan und Poo (KAN & POO 2005). Dabei wurde jede Suchanfrage von Juroren auf einer Skala daraufhin bewertet, ob es sich um eine Known-Item-Suchanfrage handelt. In einem zweiten Schritt wurden die Suchanfragen nochmals, diesmal mitsamt der Suchergebnisse als zusätzliche Information bewertet. Im dritten Schritt wurden die Suchergebnisse daraufhin bewertet, ob es sich um das Known-Item handelt. Es zeigte sich, dass es zwischen den Juroren eine hohe Übereinstimmung bei der Bewertung der Suchanfragen gab, und dass die übereinstimmende Bewertung sich mit Hinzunahme der Suchergebnisse verbesserte (vgl. KAN & POO 2005, S. 93).

Bei den intellektuell klassifizierten Known-Item-Suchanfragen wurden folgende Eigenschaften festgestellt: Sie haben eine höhere Wortanzahl und enthalten Artikel, Eigenwörter, Groß- und Kleinschreibung, bestimmte Suchschlüssel und Signalwörter wie „journal“ oder „textbook“ (vgl. KAN & POO 2005, S. 93 f.). Zusätzlich zu diesen Merkmalen wurde auch ein Sprachmodell entwickelt, das die Wahrscheinlichkeit einer Wortsequenz wiedergibt (vgl. KAN & POO 2005, S. 94). Diese mit verschiedenen Merkmalen annotierten Known-Item-Suchanfragen dienten dann als Trainingsmenge für die automatische Klassifikation weiterer Suchanfragen mittels maschinellen Lernens. Die Autoren stellten fest, dass die automatische Klassifikation am stärksten auf die Merkmale des Sprachmodells zurückgreift, gefolgt von der Anzahl der Wörter und der Großschreibung von Wörtern. Weniger wichtig waren Suchschlüssel und Signalwörter (vgl. KAN & POO 2005, S. 95).

Danach wurde ein Verfahren zur Identifikation des Known-Items auf eine Known-Item-Suchanfrage entwickelt. Dafür wurde ermittelt, wie Suchanfrage und Suchergebnis, wenn nach dem Titel eines Known-Items gesucht wurde, übereinstimmten. Wesentlich war, dass die

Wortreihenfolge sowie ein großer Teil des Titels insbesondere die ersten und letzten Wörter eines Titels übereinstimmten. Außerdem sollten Signalwörter wie „journal“ und „atlas“ im Titel des Suchergebnis vorhanden sein (vgl. KAN & POO 2005, S. 96). Bei dieser Studie wurden keine Aussagen über den Anteil der gefundenen Known-Item-Suchen an allen Suchen getroffen.

Der Sucherfolg von Known-Item-Suchen wurde bisher für OPACs meist aus der Perspektive des Misserfolgs einer Suche anhand der Fehler in der Suchanfrage erforscht (s.o., S. 21). Da bei OPACs die Sortierung nach Erscheinungsjahr als Standardeinstellung verwendet wird, hat der Rang des Known-Items keine Aussagekraft über die Retrievaleffektivität. Stattdessen kann die Treffermenge als Indiz für Sucherfolg genommen werden. Bei einer Known-Item-Suche gibt es idealerweise einen einzigen Treffer, der das gesuchte Known-Item ist. Hierzu gibt es jedoch keine Ergebnisse in den Studien. Laut Krebs ist „positiv [...], dass die known-item search im Großen und Ganzen den Nutzern meist schnell ein zufriedenstellendes Ergebnis liefert. So findet ein Großteil der Recherchierenden oft schon nach der ersten Suchanfrage das gewünschte Medium“ (KREBS 2013, S. 99). Er wertet unter zehn Treffer als gutes Ergebnis, das bei 49% der Known-Item-Suchen erzielt wird. Null Treffer treten bei 29% der Known-Item-Suchanfragen auf. Zudem kann nicht davon ausgegangen werden, dass Nutzer selbst bei einer Treffermenge von unter zehn Treffern das gesuchte Known-Item identifizieren, da sie von Google gewöhnt sind, dass der relevanteste Treffer weit oben steht.

### **5.3 Die Known-Item-Suche im Discoverysystem**

Mehrere Studien beschäftigen sich mit den Suchanfragen, die Nutzer an (Web-Scale-) Discoverysysteme stellen. Von Interesse ist meist, was und wie Nutzer über den einfachen Suchschlitz suchen und ob sich das Suchverhalten mit der Einführung eines neuen Systems geändert hat. In diesen Studien werden die Suchanfragen intellektuell klassifiziert und Known-Item-Suchanfragen mehr oder weniger ausführlich berücksichtigt.

#### **Chapman et al. 2013**

Chapman et al. untersuchten die Suchanfragen in einem Web-Scale-Discoverysystem, in dem Datenbanken, Katalog, elektronische Zeitschriften, Rechercheführer, Webseiten und Sammlungen der Bibliothek gleichzeitig durchsucht werden. Die Suchanfragen wurden auf

zwei Ebenen klassifiziert: was gesucht wird (*item-type*, d.h. Medientyp) und wie gesucht wird (*query type*, d.h. Suchanfragentyp). Suchanfragen wurden nur einem Medientyp zugeordnet, wenn fünf von sieben Juroren in ihrer Bewertung übereinstimmten. (vgl. CHAPMAN et al. 2013, S. 411).

Beim Medientyp – dem „Was“ einer Suche – wurde unterschieden zwischen Known-Item-Suchen und anderen, d.h. *non-known-item searches*. Known-Item-Suchanfragen wurden definiert als Suchen, bei denen der Nutzer nach dem Titel eines „unique, identifiable object“ suchen (CHAPMAN et al. 2013, S. 411). Zur Identifizierung eines Known-Items waren daher weitere Recherchen, wie z.B. in Google oder WorldCat, erlaubt. Known-Items waren: Artikel, Bücher, Datenbanken, Multimedia, mehrfache Formate (wenn das Medium in mehreren Formaten vorliegt, z.B. Film und Buch) und Zeitschriften. Suchen nach Personennamen wurden nicht als Known-Item-Suchanfragen eingeordnet, da unklar war, ob Medien über die Person oder Medien von der Person gesucht wurden. Im Zweifelsfall wurden Suchen mit Stichwörtern eher der thematischen Suche zugeordnet, auch wenn der Titel eines Buches mit der Suchanfrage übereinstimmte. Großschreibung, Signalwörter (z.B. „introduction“), Präpositionen und Konjunktionen in der Suchanfrage wurden jedoch als Hinweise auf eine Known-Item-Suche gewertet (vgl. CHAPMAN et al. 2013, S. 412 f.).

Das „Wie“ einer Suche beschrieb, welcher Inhalt eines Known-Items mit den Suchwörtern angegeben wurde, z.B. Titel, DOI, Signatur, ISBN, ISSN oder eine Kombination aus mehreren Metadaten wie Autor und Titel, was als „Multi“ bezeichnet wurde (vgl. CHAPMAN et al. 2013, S. 412).

Von den knapp 1000 untersuchten Suchanfragen waren 44% Known-Item-Suchanfragen (vgl. CHAPMAN et al. 2013, S. 414), von denen wiederum 46% nach einer Datenbank 26% nach einem Buch und 18% nach einer Zeitschrift waren. Die restlichen 10% verteilten sich auf Suchen nach Artikeln<sup>13</sup>, Multimedia und mehrfachen Formaten (vgl. CHAPMAN et al. 2013, S. 416). Der Sucherfolg oder die Retrievaleffektivität wurden in dieser Studie nicht erhoben.

## **McKay und Buchanan 2011**

McKay und Buchanan untersuchten ein Discoverysystem, bei dem über Radio-Buttons verschiedene Informationsressourcen ausgewählt werden konnten, u.a. der

---

<sup>13</sup> Die Artikelsuche befindet sich in einem anderen Reiter. Daher fällt der Anteil so gering aus.

Bibliothekskatalog, der Index mit Artikeldaten von EBSCO und Google Scholar (MCKAY & BUCHANAN 2011). Für 600 Suchanfragen wurde ähnlich wie bei Chapman et al. (s.o.) untersucht, wie, d.h. mit welchen Metadaten, und was, d.h. welche Known-Items, gesucht wurden. Als Metadaten wurden hier genauere Informationen als bei Chapman et al. (s.o.) erhoben, nämlich Titel, Autor, Datum, Verlag, Quelle, Auflage/Ausgabe, Seiten, Stichwörter (vgl. MCKAY & BUCHANAN 2011, S. 265). Known-Item-Suchanfragen waren Suchen mit dem Titel, „[...] or if a combination of other metadata (such as author and date) identified an individual document“ (MCKAY & BUCHANAN 2011, S. 266). Als Titelsuche wurden solche Suchanfragen klassifiziert, bei denen drei Wörter oder mehr in der gleichen Reihenfolge oder ein Neologismus im Titel eines Werkes enthalten waren, oder bei denen weniger als drei Wörter mit dem Titel eines Werkes exakt übereinstimmten (vgl. MCKAY & BUCHANAN 2011, S. 265). Als gesuchte Medientypen wurden Buch, Artikel, Zeitschrift, Datenbank, DVD und andere erfasst (vgl. MCKAY & BUCHANAN 2011, S. 266).

Da in dieser Studie zu der Häufigkeit der Known-Item-Suchanfragen, der Verteilung der Metadaten und den gesuchten Medientypen nur absolute Zahlen genannt werden, ohne dass die jeweiligen Bezugsgrößen genannt werden, werden hier nur die wenigen relativen Ergebnisse oder Tendenzen in den Ergebnissen wiedergegeben. Genannt wird zumindest, dass Known-Items mit dem Titel am häufigsten gesucht werden, wobei 75% dieser Suchanfragen exakt mit dem Titel des Treffers übereinstimmten (vgl. MCKAY & BUCHANAN 2011, S. 265). Nur 8% aller Known-Item-Suchanfragen wurden nicht mit dem Titel, sondern einer Kombination aus anderen Metadaten gesucht (vgl. MCKAY & BUCHANAN 2011, S. 266), wobei Suchanfragen nach einem Autor am dritthäufigsten und Suchanfragen mit Angaben zu Jahr, Verlag und Auflage/Ausgabe sehr viel seltener sind (vgl. Table 2 in MCKAY & BUCHANAN 2011, S. 265). Außerdem wird in den Ergebnissen deutlich, dass die Medien dem Inhalt der Informationsressourcen entsprechend gesucht wurden, also Suchen nach Büchern im Katalog am häufigsten waren und Artikel vor allem in EBSCO und Google Scholar gesucht wurden. Zeitschriften und Datenbanken wurden kaum gesucht (vgl. MCKAY & BUCHANAN 2011, S. 266).

In dieser Studie werden solche Known-Item-Suchanfragen als erfolglos bezeichnet, bei denen das Known-Item nicht in den ersten fünf Treffern auftrat (vgl. MCKAY & BUCHANAN 2011, S. 267).



## Meadow und Meadow 2012

Meadow und Meadow untersuchten ebenfalls Suchanfragen an ein Web-Scale Discoverysystem, ohne allerdings die obige Unterscheidung zwischen „Wie“ und „Was“ zu machen. Es wurden sieben *query types* erfasst, wobei diese teilweise durch die gesuchte Medienart (z.B. „database/journal“), teilweise die Merkmale der Suchanfrage (z.B. „natural language“) charakterisiert wurden. Des Weiteren wurden die Suchanfragentypen danach unterschieden, ob das System den Suchanfragentyp unterstützt: „high quality“ waren Suchen der Kategorien „subject, Boolean, known item and database/journal“, „low quality“ waren die Kategorien „URL [...] invalid and natural language“ (MEADOW & MEADOW 2012, S. 169). Darunter sind Known-Item-Suchanfragen für die Autoren Suchen „for a particular item in which the author or title is known“ (MEADOW & MEADOW 2012, S. 169). Demnach sind Suchen nach einer Person Known-Item-Suchanfragen, wenn zusätzlich die Suchkategorie „Author“ ausgewählt war. Die Autoren identifizierten Known-Item-Suchanfragen dadurch, dass jede Suchanfrage mit mehr als drei Wörtern im WorldCat daraufhin geprüft wurde, ob es Treffer gab. Allerdings wird keine Aussage darüber getroffen, wie hoch die Übereinstimmung zwischen Suchanfrage und Titel sein musste, um als Known-Item-Suchanfrage klassifiziert zu werden. Bei Suchanfragen mit zwei oder weniger Wörtern lag es im Ermessen des Jurors, ob eine Known-Item-Suchanfrage vorlag oder nicht, wobei Suchanfragen, die breite Themengebiete umfassen, im Zweifelsfall als thematische Suchen und „obvious titles or journal/database titles“ als Known-Items klassifiziert wurden (MEADOW & MEADOW 2012, S. 169). Die Autoren erfassten keine Merkmale der Known-Item-Suchanfragen und unterschieden nur zwischen Datenbanken und Zeitschriften sowie Known-Items. Es ist zudem nicht ersichtlich, ob Suchen mit anderen Informationen als Autor oder Titel – z.B. ISBN oder Signatur – auch als Known-Item-Suchanfragen klassifiziert wurden.

Im Vergleich zu den vorherigen Studien nennen Meadow und Meadow einen überraschend geringen Anteil von 21,7% für die Kategorie „Known Item“ und 5% für „Database/Journal“ innerhalb der sogenannten „high quality“ Kategorien, die im Schnitt gut 91% aller Suchanfragen ausmachen (vgl. MEADOW & MEADOW 2012, S. 171). Zu Anfang zweier Semester (August 2010 und Januar 2011) waren die Suchanfragen der Kategorie „Known Item“ mit einem Anteil von 30% an allen Suchanfragen häufiger als in den jeweils darauf folgenden Monaten (etwa 15-20%) (vgl. Figure 1 in MEADOW & MEADOW 2012, S. 170).

## 6 Methodik und Datenerhebung

Für die Evaluation von Informationssystemen gibt es verschiedene Ansätze, anhand welcher Aspekte ein Informationssystem untersucht werden kann (vgl. PETRAS 2013, S. 369). Diese Arbeit fokussiert sich auf die system-zentrierte Evaluation, berücksichtigt aber nur einen Suchanfragentyp, nämlich Known-Item-Suchanfragen. Untersucht wird also die Effektivität des Systems in Hinblick auf ein bestimmtes Informationsbedürfnis des Nutzers. Die Effektivität eines Systems misst sich an seiner Fähigkeit, korrekte Ergebnisse auszugeben bzw. „relevante Dokumente zu finden und zu ranken“ (vgl. PETRAS 2013, S. 372).

### 6.1 Methodisches Vorgehen

Im Folgenden soll auf die Kriterien eingegangen werden, anhand derer die Entscheidungen für den Aufbau der Untersuchung getroffen wurde.

Die Auswahl der Suchmaschine hängt vom Zweck der Untersuchung ab (vgl. LEWANDOWSKI 2011, S. 208). In dieser Arbeit sollte ein Bibliothekskatalog mit Suchmaschinentechnologie, nämlich ein Discoverysystem untersucht werden. Für beluga sprach, dass es zum einen zukünftig der Hauptkatalog im Bibliothekssystem der Universität Hamburg werden soll und dass es zum anderen als konsortiales Recherchesystem unter der Beteiligung vieler Hamburger wissenschaftlichen Bibliotheken eine für einen regionalen Bibliothekskatalog relativ große Abdeckung und Reichweite hat.

Bei der Anzahl der untersuchten Suchanfragen gibt es in den verschiedenen Studien eine sehr große Spannweite, als Minimum gelten 50 Suchanfragen (vgl. LEWANDOWSKI 2011a, S. 208). Da in dieser Arbeit nur ein Suchanfragentyp berücksichtigt werden soll, ist eine vorherige intellektuelle Klassifikation notwendig. Bei derartigen Studien reicht die Anzahl der untersuchten Suchanfragen von 320 bis 1200. Da die Anzahl der Suchanfragen bei beluga nicht sehr hoch ist, können Suchhäufigkeiten nicht berücksichtigt werden<sup>14</sup>.

Für die Untersuchung der Retrievaleffektivität von Known-Item-Suchanfragen in beluga ist es sinnvoll, die tatsächlichen Suchanfragen der Nutzer an beluga zu nutzen (vgl. LEWANDOWSKI

---

<sup>14</sup> Dies ist auch nicht notwendig, schließlich sollen vorrangig eindeutige Known-Item-Suchanfragen identifiziert werden, die auch für einen Algorithmus als solche erkennbar sind. Suchanfragen, die je nach Nutzer einem sonstigen Suchanfragentyp zugeordnet werden können, sind zwar wichtig für die Erhebung der Häufigkeit von Known-Item-Suchanfragen, können aber ohne weitere Kontextinformationen nicht mit einem Algorithmus klassifiziert werden.

2010, S. 92). Da die Suchanfragen aus den Logdateien extrahiert wurden, beschränken sich die Kontextinformationen auf die weiteren Informationen aus den Logdateien. Kontextinformationen wie von den Nutzern beschriebene Informationsbedürfnisse und Erwartungen an die Suche sind vor allem zur Bewertung der Relevanz von Suchergebnissen zu einer thematischen Suchanfrage hilfreich, in dieser Arbeit aber nicht zwingend notwendig. Die Bewertung der Relevanz eines Treffers gestaltete sich bei Known-Item-Suchanfragen relativ einfach: Ein Suchergebnis ist entweder das gesuchte Known-Item, oder auch nicht. Deshalb wurden für diese Untersuchung auch keine Juroren benötigt, wie sie sonst in Standardtests üblich sind (vgl. LEWANDOWSKI 2011a, S. 211) und die Bewertung der Relevanz konnte durch eine Person – die Autorin – erfolgen (vgl. LEWANDOWSKI 2011b, S. 358).

Für die Untersuchung wurden zwei Datensets eines vergleichbaren Zeitraums benötigt: Anhand des ersten Sets wurden anhand eines Schemas intellektuell der Suchanfragentyp erkannt und Merkmale für Known-Item-Suchanfragen erfasst. Außerdem wurde festgehalten, an welcher Position in der Trefferliste das gesuchte Known-Item steht. Dann wurde anhand der Kennzahlen Success @n und Mean Reciprocal Rank (siehe Kapitel 6.2) die Retrievaleffektivität von beluga für die Known-Item-Suchen berechnet.

Da es momentan nicht möglich ist, das Relevanzranking nur für einen Suchanfragentyp (hier: Known-Item-Suchanfrage) bei beluga zu verändern, ohne dass dies gleichzeitig unvorhersehbare Auswirkungen auf die Retrievaleffektivität für den sonstigen Suchanfragentyp (thematische Suche) hat, ist es wünschenswert, die Suchintention des Nutzers schon bei der Eingabe automatisch zu erkennen. Dann könnte ein angepasstes Ranking für diesen Suchanfragentyp greifen. Daher wurde ein Algorithmus entwickelt, der anhand bestimmter Merkmale die Known-Item-Suchanfragen erkennen soll. Da die Gefahr bestand, dass der Algorithmus zu sehr auf die Merkmale der Suchanfragen abgestimmt wurde, anhand derer er entwickelt wurde, wurde er an einem zweiten Datenset getestet. Für die Berechnung der Kennzahlen Precision und Recall, die für die Evaluierung dieses Algorithmus geeignet sind (siehe Kapitel 6.2), wurden aus Zeitgründen nur in einer zufälligen Anzahl (50% des zweiten Datensets) wiederum nach dem vorherigen Schema intellektuell die Known-Item-Suchanfragen identifiziert. So konnte festgestellt werden, wie viele Known-Item-Suchanfragen vom Algorithmus erfasst und wie viele nicht erfasst wurden. Außerdem wurde die Retrievaleffektivität für die erfassten Known-Item-Suchanfragen ermittelt, denn wenn der

Algorithmus nur gut gerankte Suchanfragen erfassen sollte, kann er nicht zur Verbesserung der Retrievaleffektivität eingesetzt werden.

## **6.2 Kennzahlen zur Evaluierung der Retrievaleffektivität**

Die klassischen Kennzahlen für die Evaluierung eines Algorithmus sind Precision und Recall. Diese beiden Maße geben für binäre Klassen an, wie genau (Precision) und vollständig (Recall) die Ergebnisse der Klassifikation sind.

Bezogen auf die Ergebnisse einer Suchmaschine bezeichnet Precision den Anteil an relevanten Dokumenten in der Ausgabemenge und Recall den Anteil der ausgegebenen relevanten Dokumenten an der Gesamtmenge aller relevanter Dokumente (vgl. BAEZA-YATES 2011, S. 135). Ein Relevanzranking-Algorithmus kann beispielsweise danach beurteilt werden, wie viele relevante Ergebnisse in der Trefferliste sind (Precision) und welcher Anteil an allen relevanten Ergebnissen in der Trefferliste ist (Recall). Diese Maße sind jedoch nur für die Suchanfragen sinnvoll, für die es mehrere relevante Treffer gibt (vgl. LEWANDOWSKI 2011a, S. 214). Daher eignen sich die beiden Kennzahlen nur zur Evaluierung des Algorithmus zur Erkennung von Known-Item-Suchanfragen, nicht aber zur Evaluierung der Retrievaleffektivität von beluga auf Known-Item-Suchanfragen.

Für Suchanfragen, bei denen nur der erste korrekte Treffer gewertet werden soll, eignen sich vor allem die Retrievalmaße Success @n und Mean Reciprocal Rank (vgl. CRASWELL et al. 2004, S. 4; LEWANDOWSKI 2011a, S. 214).

Success @n gibt den „Erfolg“ für einen Rang n an. Es wird für eine Menge an Suchanfragen ermittelt, welchen Rang der korrekte Treffer einnimmt. Dann wird der Anteil an korrekten Treffern bis zu einem Rang n berechnet. Üblich ist eine Erhebung für die ersten 10 Treffer, da viele Suchmaschinen 10 Treffer auf einer Trefferseite präsentieren und Nutzer kaum mehr Treffer betrachten. Genauer betrachtet werden meist der Rang 1, 5 und 10 (vgl. LEWANDOWSKI 2011b, S. 357). Der Web Track von TREC gibt nur Success @10 an (vgl. CRASWELL et al. 2004, S. 4). Mit dieser Kennzahl lassen sich also für verschiedene Trefferpositionen die Erfolgsquoten benennen.

Ähnlich zu Success @n ist der Mean Reciprocal Rank (vgl. BAEZA-YATES 2011, S. 142). Der Mean Reciprocal Rank gibt jedoch einen relativen Durchschnittswert für die korrekten Ergebnisse in den ersten fünf Treffern an. Hier wird der erste korrekte Treffer bis Rang 5

berücksichtigt und diesem mit absteigender Position ein niedrigerer Wert zugeteilt. Daraus wird dann ein Durchschnittswert ermittelt. Üblich ist einem korrekten Treffer an Position 1 den Wert 1 und dann für jede weitere Position bis einschließlich 5 den halbierten Wert der vorigen Position zuzuweisen – für Position 2 also 0,5, für Position 3 den Wert 0,125 usw. (vgl. MACFARLANE 2007, S. 354). Alle weiteren Treffer bekommen den Wert 0. Je näher der Wert an 1 liegt, desto höher wurde ein korrektes Ergebnis im Durchschnitt gerankt.

### **6.3 Das untersuchte Discoverysystem: beluga**

In dieser Arbeit wird die Retrievaleffektivität von dem Discoverysystem beluga untersucht. Angefangen hat beluga als Eigenentwicklung, die als gemeinsames Projekt sechs wissenschaftlicher Bibliotheken in Hamburg während der Projektlaufzeit vom 01.11.2007 bis 31.10.2010 durch ELCH (E-Learning Consortium Hamburg) gefördert wurde. Dazu wurde ein „Hamburg-Index“ mithilfe der Open-Source-Suchmaschinentechologie Solr/Lucene entwickelt. Das Ziel belugas war von Beginn an, einen Katalog 2.0 umzusetzen, der sich konsequent an den Erwartungen und Wünschen der Nutzer orientiert, um eine hohe Gebrauchstauglichkeit zu gewährleisten. Dafür wurden in der Projektphase mehrere Fokusgruppen mit Studierenden und Lehrenden durchgeführt. Im Oktober 2008 ging der erste Prototyp belugas online und mehrere Usabilitytests wurden durchgeführt, um die Gebrauchstauglichkeit der Benutzeroberfläche zu testen. Als Resultat der Fokusgruppen und der Usabilitytests wurden z.B. aufgrund von Vorbehalten Lehrender und Studierender Merklisten nicht veröffentlicht, Tagging nicht angeboten und das Schreiben eigener Rezensionen nicht ermöglicht (vgl. CHRISTENSEN 2010, S. 327).

Nach einer kompletten technischen Remodellierung mit VuFind als Softwarebasis und dem GBV-Discovery-Index<sup>15</sup> als Datenbasis wird beluga seit Dezember 2012 auf der Startseite der SUB Hamburg als Alternative zum Campus-Katalog angeboten. Die Usability nimmt immer noch einen hohen Stellenwert in der Weiterentwicklung des Systems ein. So wurden die Nutzerführung der vorherigen Version beibehalten und weitere Usabilitytests durchgeführt (beluga 2013).

Vermutlich ab Sommersemester 2015 wird beluga den aggregierten Index Primo Central einbinden. Dieser wird über die sogenannte Reiterlösung angeboten werden, d.h. dass über

---

<sup>15</sup>[http://www.gbv.de/wikis/cls/VuFind#VuFind-Service\\_der\\_VZG](http://www.gbv.de/wikis/cls/VuFind#VuFind-Service_der_VZG)

den einfachen Suchschlitz beide Indizes gleichzeitig durchsucht werden, die Suchergebnisse jedoch in eine Trefferliste mit lokalen Beständen und in eine Trefferliste mit den Ergebnissen aus dem kommerziellen Index aufgeteilt werden. Ergebnisse dieser Untersuchung könnten also unter gleichen Voraussetzungen wie zurzeit Verwendung finden.

Zum Zeitpunkt dieser Untersuchung ergab eine leere Suche in beluga knapp 17 Millionen Treffer. Davon sind über 7 Millionen Titel lokale Bestände aus den zehn teilnehmenden wissenschaftlichen Bibliotheken Hamburgs, 6 Millionen Titel aus den Nationallizenzen sowie Aufsatzdaten aus der Datenbank Online Contents. Die Medientypen setzen sich zusammen aus 9 Millionen Artikel und E-Artikel, 6,5 Millionen Bücher und E-Books, 640.000 Zeitschriften- und Reihenbände sowie 300.000 Zeitschriften und Reihen.

Auf der Startseite belugas befindet sich ein einfacher Suchschlitz, in dem Suchanfragen eine Freitextsuche auslösen. Die Suchergebnisse werden standardmäßig nach Relevanz sortiert. Auf einer Ergebnisseite werden 20 Suchergebnisse angezeigt, die über Facetten gefiltert werden können.

## **6.4 Datenerhebung**

Jede Interaktion eines Nutzers mit einer Webseite wird vom Webserver in sogenannten Logdateien aufgezeichnet. Je nach Webserver werden verschieden umfangreiche Informationen protokolliert. Die Logdateien von beluga stammen von einem Apache Server, der das Combined Log Format verwendet (vgl. Apache 2014). Protokolliert werden unter anderem IP-Adresse, Datum und Uhrzeit, HTTP-Anforderung mit HTTP-Befehl, Name der betroffenen Webseite/des Dokuments, übertragenes Datenvolumen, zuvor besuchte Webseite des Clients (Referrer) sowie Browserinformationen des Clients.

Um die Suchanfragen an beluga aus den Logdateien zu extrahieren wurde die Open-Source-Software Piwik<sup>16</sup> verwendet. Piwik ist ein Programm zur Webanalytik, das zum einen mittels Zählpixel oder JavaScript Tracking eine kontinuierliche Kontrolle und Analyse des Nutzerverhaltens auf einer Webseite erlaubt. Zum anderen können die Logdateien eines Webservers in Piwik hochgeladen werden. Da nur Logdateien eines vergangenen Zeitraums untersucht werden sollten, wurde das zweite Verfahren genutzt. Dafür wurden die Logdateien von beluga zunächst von automatisierten Anfragen (Bots) bereinigt.

---

<sup>16</sup> Piwik.org

Für das Datenset zur intellektuellen Klassifikation der Suchanfragen wurde ein Zeitraum von drei zusammenhängenden Tagen (Montag bis Mittwoch) in einer zufälligen Woche zu Beginn des Sommersemesters 2014 (April bis Mai) ausgewählt. Wochenbeginn und Semesterbeginn wurden ausgewählt, da aus den insgesamt importierten Daten deutlich wurde, dass einerseits in den ersten Wochen des Semesters beluga relativ gleichmäßig genutzt wurde und die Nutzung in der ersten Wochenhälfte relativ hoch ist. Es wurden alle Suchanfragen vom 07.-09. April 2014 berücksichtigt. Für die Evaluierung des Algorithmus wurden als Testmenge die Suchanfragen der gleichen Wochentage eine Woche später verwendet.

Piwik stellt viele verschiedenen Werkzeuge zur umfassenden Analyse der Logdateien bereit. Für diese Arbeit wurden mit der Funktion „interne Suchen“ die Suchanfragen in beluga in eine Excel-Tabelle exportiert. Neben der Suchanfrage werden dabei folgende Informationen mit heruntergeladen:

- Suchanfragen (wie oft die Suchanfrage insgesamt gestellt wurde)
- Seitenansichten (Anzahl der besuchten Ergebnisseiten)
- Von Besuchern verbrachte Gesamtzeit (in Sekunden)
- Ausstiege
- Besuchte Ergebnisseiten
- Durchschnittszeit pro Seite
- Absprungrate
- Ausstiegsrate nach Suche

Piwik teilt unter der Funktion „Besucher-Log“ die Logdateien in Benutzer-Sessions ein. Hier werden die aufeinanderfolgenden Interaktionen eines Nutzers mit dem System dargestellt, u.a. die eingegebenen Suchanfragen, die besuchten Webseiten und die Dauer der Seitenansichten in chronologischer Ansicht. Die Dauer einer Besucher-Session ist in Piwik standardmäßig 30 Minuten (vgl. Piwik 2014).

## **7 Auswertung**

Aus den Suchanfragen wurden doppelte Werte, wie sie von Excel erkannt werden, entfernt. Danach enthielt das erste Datenset 1174 Suchanfragen.

Angelehnt an die Studien in Kapitel 5, die die Suchanfragen intellektuell klassifizierten (CHAPMAN et al. 2013; KAN & POO 2005; MCKAY & BUCHANAN 2011; MEADOW & MEADOW 2012), wurde für die Suchanfragen zum einen erfasst, welche Eigenschaften die Suchanfragen haben und, im Falle von Known-Item-Suchen, was für ein Medientyp gesucht wurde.

Für jede Suchanfrage wurde die Anzahl der Wörter erfasst, die Trefferanzahl und – falls es aus der Treffermenge in beluga nicht sofort ersichtlich war – ob eine Suche im Metakatalog Karlsruher Virtueller Katalog (KVK) oder bei Google Scholar Treffer erzielte, die mit der Suchanfrage übereinstimmten. Als Metadaten wurde bezeichnet, wie gesucht wurde, d.h. welchen Kriterien die Suchwörter entsprachen. Dafür wurde zunächst intellektuell und einzeln für jede Suchanfrage erfasst, ob eines oder eine Kombination der folgenden Kriterien enthalten war:

- **Personenname:** Es wurde vermutet, dass die Suche nach Personennamen auf die Angabe von Autoren und sonstigen Urhebern hinweist. Diese wurden daran erkannt, dass ein Wort als Verfasser in einem möglichen Treffer vorkam.
- **Jahr:** Eine Zahl mit vier Ziffern wurde als Jahreszahl erfasst, da vermutet wurde, dass damit das Erscheinungsjahr gesucht wurde.
- **Zahl:** Eine Zahl mit ein bis drei Ziffern wurde als Hinweis auf die Angabe einer Bandnummer, einer Auflage oder einer Heftnummer verstanden.
- **Ort, Verlag:** Diese Wörter wiesen auf den Erscheinungsvermerk als Suchkriterien hin.
- **Standardnummern:** Dies konnten ISBN, ISSN oder die Pica-Produktionsnummer sein.
- **Signatur.**
- **Phrasenmarkierung, sonstige Suchschlüssel** (z.B. „OR“, „\*“, „?“ , „!“).

Alle anderen Wörter wurden auf die Art ihres syntaktischen Zusammenhalts hin geprüft, um so Hinweise auf Suchen nach Titeln zu erhalten:

- **Stichwörter:** Dies waren einzelne Wörter ohne syntaktischen Zusammenhang.
- **Wortverbindungen:** Dies waren vor allem zusammenhängende Stichwörter aus dekliniertem Adjektiv und Substantiv.
- **Titeltypische Wörter innerhalb von Stichwörtern oder Wortverbindungen:** Damit waren Funktionswörter wie Artikel, Präpositionen, Pronomen und Konjunktionen sowie Verben gemeint.



In einem zusätzlichen Kommentarfeld wurde das Vorhandensein von Signalwörtern (z.B. Zeitschrift, Handbuch, Lexikon), die auf den Medientyp des Known-Items hinweisen, festgehalten. Die Trefferanzahl musste über eine erneute Suchanfrage in beluga neu erhoben werden. Für Known-Item-Suchanfragen wurden außerdem der Rang Known-Items, der Medientyp des gesuchten Known-Items und der Grad der Übereinstimmung zwischen Titel und Suchanfrage in drei Abstufungen (vollständig, teilweise oder fehlerhaft) ermittelt sowie ob das Known-Item von dem Nutzer angeklickt wurde. Die Merkmalsausprägungen wurden für jede Suchanfrage in einer Excel-Tabelle notiert bzw. kodiert und mithilfe von Excel-Formeln ausgewertet.

Die Parameter des Rankingalgorithmus sind zwar seit dem Erhebungszeitraum der Suchanfragen und während dieser Untersuchung nicht modifiziert worden, der Umfang des Indexes jedoch ändert sich kontinuierlich. Deshalb können manche Ergebnisse nur als Momentaufnahme gesehen werden.

## **7.1 Identifizierung der Known-Item-Suchanfragen**

Um entscheiden zu können, welche Suchanfragen Suchen nach Known-Items sind, wurden möglichst objektive Kriterien festgelegt. Ziel war es, eine Suchanfrage unabhängig vom jeweiligen Nutzer und seinem Suchkontext einstufen zu können. Zum einen sollten Eigenschaften der Known-Item-Suchanfragen festgestellt werden, anhand derer eine Suchanfrage für einen Algorithmus erkennbar ist, zum anderen sollte die ungefähre Häufigkeit von Known-Item-Suchanfragen festgestellt werden, um besser einschätzen zu können, ob und wie hoch die Notwendigkeit einer Anpassung von beluga an diesen Suchanfragentyp ist. Für ersteres mussten die Suchanfragen zweifelsfrei und unabhängig vom weiteren Suchverhalten eines einzelnen Nutzers klassifiziert werden können, für letzteres wurde in begrenztem Umfang das Nutzerverhalten mitberücksichtigt.

Voraussetzung dafür, dass eine Suchanfrage als Known-Item-Suchanfrage klassifiziert wurde, war, dass ein spezifisches Known-Item ermittelbar war. Deshalb wurde für jede Suchanfrage geprüft, ob mindestens ein Ergebnis in beluga, im KVK oder in Google Scholar (in seltenen Fällen in der Google Websuche) gefunden wurde, das den Suchkriterien entsprach.

Zusätzlich musste eine der folgenden Richtlinien zutreffen:

- Die Suchanfrage entspricht einer Signatur, ISBN oder ISSN.

- Die Suchanfrage enthält eine Kombination aus verschiedenen Metadaten (für genauere Angaben siehe Kapitel 7.4). Bei manchen Metadatenkombinationen ist das exakte Known-Item nicht feststellbar, wie z.B. Autor und Jahr oder Autor und Ort. Hier wurde dennoch angenommen, dass der Nutzer einen spezifischen Treffer sucht, daher wurden diese als Known-Item-Suchanfragen gewertet. Nicht alle Metadatenkombinationen sind Known-Item-Suchanfragen, z.B. wenn Informationen über eine Person gesucht wurden.
- Die Suchanfrage besteht aus zwei Personennamen, die als Autor identifiziert wurden. Auch hier ist das spezifische gesuchte Known-Item nicht immer eindeutig identifizierbar.
- Die Suchanfrage enthält ein Signalwort oder eine eindeutige Zeitschriftenabkürzung aus mindestens drei Buchstaben.
- Wenn nur mit Stichwörtern gesucht wird, müssen diese möglichst vollständig mit dem Titel übereinstimmen oder einen Großteil des Titels ausmachen. Außerdem dürfen maximal drei unterschiedliche Ergebnisse mit fast vollständiger Titelübereinstimmung gefunden werden. Dadurch wurde vermieden, dass Suchanfragen zu oft behandelten Themen, zu denen es viele Titelerklärungen gibt, als Known-Item-Suchanfragen klassifiziert werden. Suchanfragen aus weniger als drei Wörtern müssen zudem entweder ungewöhnliche Wörter (z.B. „Mondspielerin“), eine ungewöhnliche Kombination aus Wörtern (z.B. „Kino spüren“) oder titeltypische Wörter enthalten. Die Wörter dürfen nicht ausschließlich Eigennamen, Namen von Gesetzen, Wissenschafts- oder Themengebiete oder Schlagwörter sein.

Das letzte Kriterium ließ einen gewissen Ermessensspielraum zu, da „ungewöhnliche“ Wörter und Wortkombinationen nicht objektiv definiert wurden. Im Zweifelsfall wurde eine Suchanfrage jedoch der Kategorie der „unsicheren“ Suchanfragen zugeordnet. Diese wurde eingeführt, da es viele Suchanfragen gab, die nicht eindeutig und unabhängig von einem individuellen Nutzer als Known-Item-Suchanfragen klassifiziert werden konnten, bei denen es jedoch unterschiedliche Hinweise darauf gab, dass es sich um Suchen nach einem Known-Item handelte. Voraussetzung war wiederum, dass möglichst genaue Titelübereinstimmungen in beluga, KVK oder Google Scholar gefunden wurden. Hier waren jedoch mehrere mögliche Treffer zulässig, da diese Suchanfragen meist kürzer und inhaltlich Themengebieten zuzuordnen waren.

Folgende Kriterien wiesen auf eine unsichere Known-Item-Suchanfrage hin:

- Die Suchanfrage stimmt mit einem Zeitungs- oder Zeitschriftentitel überein (z.B. „Die Zeit“, „Perception“, „Grundschulunterricht Deutsch“).
- Eine Kombination aus zwei Großbuchstaben stimmt mit einer oder mehreren geläufigen Zeitschriftenabkürzungen überein.
- Die Suchanfrage enthält ein Signalwort (z.B. „handbuch textil“), ein titeltypisches Wort oder ungewöhnliche Großschreibung (z.B. „Die Industrie Russlands“, „The institutions of the European Union“), jedoch keine genauen oder zu viele Übereinstimmungen in beluga, KVKV oder Google Scholar.
- Ohne dass eine Auswahl in der Trefferliste getroffen wurde, wurde die vorliegende Stichwort-Suchanfrage direkt nachfolgend um spezifischere Angaben wie Autor oder Jahr erweitert oder ein Titel vervollständigt, so dass vermutet werden kann, dass bereits die vorherige Suchanfrage auf das Known-Item abzielte.

Suchanfragen mit einer Kombination aus Stichwort und Autor oder Signalwort, zu denen keine Titelübereinstimmungen gefunden wurden, wurden ebenfalls als unsichere Suchanfragen eingestuft. Hier ist anzunehmen, dass der Nutzer die Existenz eines spezifischen Titels vermutete oder die Titelinformationen nicht genau erinnerte. Außerdem ist es natürlich möglich, dass die Titelinformationen mit meinen begrenzten Bemühungen nicht online recherchierbar waren.

Als „sonstige“ Suchanfragen werden die Suchanfragen bezeichnet, die nicht als Known-Item oder unsichere Suchanfragen eingestuft wurden. Dies sind vor allem thematische Suchanfragen und solche, in denen das Informationsbedürfnis ohne weitere Informationen des Nutzers nicht identifizierbar ist.

## **7.2 Anteil der Known-Item-Suchanfragen**

Es wurden 523 von insgesamt 1174 Suchanfragen eindeutig als Known-Item-Suchanfragen klassifiziert, dies entspricht einem Anteil von 44,55% (siehe Tabelle 1). Als unsicher wurden 6,56% der Suchanfragen eingestuft. Insgesamt sind also knapp über 50% aller Suchanfragen wahrscheinlich Suchen nach Known-Items. Wo nicht explizit anders erwähnt, werden die unsicheren Known-Item-Suchanfragen im Folgenden zu den sonstigen Suchanfragen gezählt.

	Anzahl	Prozent
<b>Known-Item</b>	523	44,55
<b>Unsicher</b>	77	6,56
<b>Sonstige</b>	574	48,89
<b>Summe</b>	1174	100

**Tabelle 1: Anzahl und Anteile der Suchanfragentypen**

Es gab zwei Gründe, aus denen ein Known-Item in beluga nicht gefunden wurde. Der erste Grund ist dem Umfang des Indexes geschuldet, wenn eine Titelaufnahme für das gesuchte Objekt nicht vorhanden war. Dies konnte zum einen daran liegen, dass für den gesuchten Medientyp in beluga keine Einzelaufnahme vorhanden ist, z.B. für Artikel aus Zeitschriften oder Buchaufsätze. Zum anderen sind viele Medien nicht im Bestand einer der beteiligten Bibliotheken. Dies traf auf 30,2% der Known-Item-Suchanfragen zu. Diese Suchanfragen sind nicht im eigentlichen Sinne erfolglos, da der Nutzer überprüfen konnte, dass das Known-Item nicht vorhanden ist<sup>17</sup>. Bei den unsicheren Suchanfragen wurde bei zwölf Stück aufgrund ihrer Eigenschaften vermutet, dass ein Known-Item gesucht wird, auch wenn kein spezifischer Titel gefunden werden konnte.

Der zweite Grund ist dem Sucherverhalten des Nutzers geschuldet, wenn das Known-Item zwar vorhanden war, der Nutzer jedoch Rechtschreibfehler oder falsche Angaben in der Suchanfrage gemacht hat. Die Suchanfragen mit Rechtschreibfehlern wurden auch als Known-Item-Suchanfrage gewertet, da die Absicht des Nutzers, einen spezifischen Titel zu finden, gegeben war. 8,4% der Known-Item-Suchanfragen blieben aufgrund von Rechtschreibfehlern oder ungeeigneten Angaben erfolglos. Oft wurden diese Suchanfragen direkt im Anschluss korrigiert und erneut durchgeführt.

Insgesamt konnten also nur 321 (61,4%) der Known-Item-Suchanfragen zu dem gesuchten Known-Item führen. Dies entspricht 27,3% aller Suchanfragen.

### **7.3 Trefferanzahl**

Betrachtet man die Anzahl der Treffer, die auf eine Suchanfrage ausgegeben werden, so fällt auf, dass mehr Known-Item-Suchanfragen kleinere Treffermengen generieren als sonstige Suchanfragen (siehe Abbildung 1). Dies ist allerdings auch dadurch bedingt, dass nur solche

---

<sup>17</sup> Für über die Hälfte dieser Suchen wurden allerdings trotzdem Treffer gefunden, d.h. dass es dann wiederum nicht so einfach ist zu überprüfen, ob ein Titel im Bestand ist oder nicht.

Suchanfragen, die konkretere Begriffe oder mehr Wörter enthielten, als Known-Item-Suchanfragen identifiziert werden konnten. Keine Treffer wurden bei 102 Known-Item-Suchanfragen und 76 sonstigen Suchanfragen gefunden (insgesamt 15% aller Suchanfragen). Dies ist ein fast halb so großer Prozentsatz wie in den genannten OPAC-Studien (siehe Kapitel 5.2). Ein bis zehn Treffer wurden bei 35,75% und mehr als zehn Treffer bei 44,7% der Known-Item-Suchanfragen gefunden.

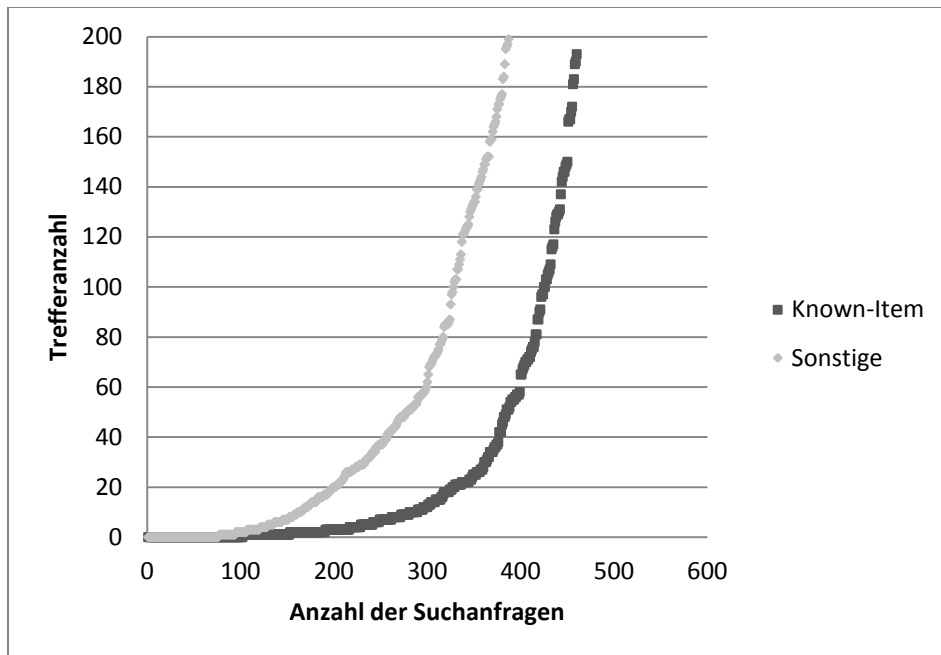


Abbildung 1: Trefferanzahlen für Known-Item- und sonstige Suchanfragen

## 7.4 Eigenschaften der Known-Item-Suchanfragen

Die Eigenschaften der Known-Item-Suchanfragen setzen sich aus der Anzahl der Wörter sowie den Merkmalen, die auf formale Suchkriterien hinweisen, zusammen.

### 7.4.1 Wortanzahl

Known-Item-Suchanfragen sind mit einer durchschnittlichen Anzahl von 4,7 Wörtern und einem Median von 4 länger als sonstige Suchanfragen, die durchschnittlich 2,2 Wörter lang sind. Der Durchschnitt für alle Suchanfragen beträgt 3,2 Wörter mit einem Median von 2. Wie aus Abbildung 2 ersichtlich ist, überwiegt ab einer Suchanfragenlänge von drei Wörtern der Anteil der Known-Item-Suchanfragen.

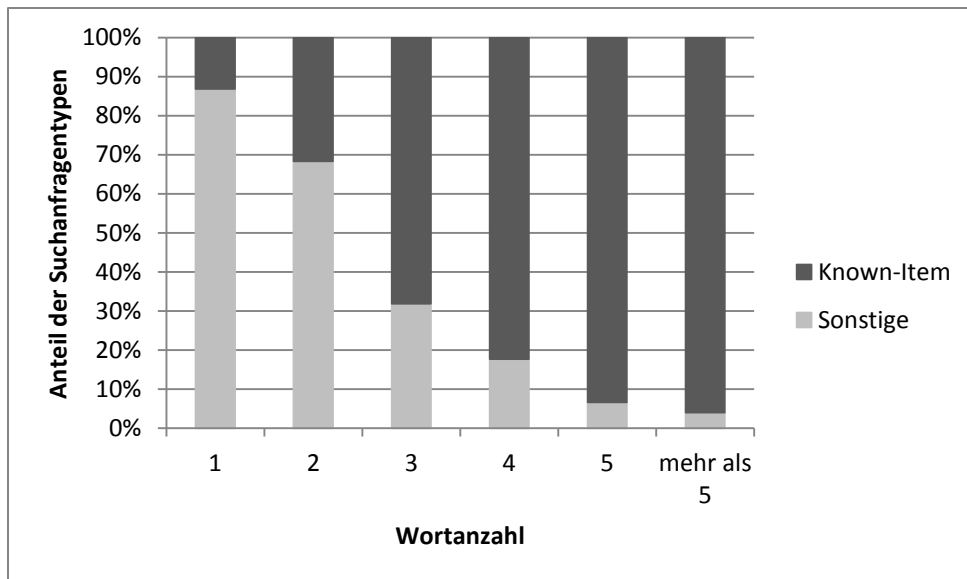


Abbildung 2: Anteil von Known-Item- und sonstigen Suchanfragen pro Wortanzahl

Jedoch ist dieses Ergebnis auch dadurch bedingt, dass es bei wenigen Wörtern selten möglich war, ohne weitere Kontextinformationen des Nutzers, der die Suchanfrage gestellt hat, eine Suchanfragen eindeutig zuzuordnen. Das Ergebnis ist also zum Teil Artefakt der Untersuchungsmethode.

## 7.4.2 Metadaten

Wie beschrieben (siehe S. 33 f.), wurde jede Suchanfrage auf zwölf mögliche Metadaten geprüft. Dies waren: Personennamen, Jahr, Zahl, Ort, Verlag, Signatur, Standardnummer, Phrasenmarkierung, sonstige Suchschlüssel, Stichwörter, Wortverbindungen und titeltypische Wörter. Die letzten drei Kriterien wurden nicht in Kombination erfasst, d.h. eine Suchanfrage mit drei Wörtern, bestehend aus einem Stichwort und aus einer Wortverbindung mit zwei Wörtern, wurde nur als Wortverbindung kodiert.

Diese zwölf Kategorien traten in insgesamt 33 verschiedenen Ausprägungen auf. Mit 82,2% wies der Großteil der Suchanfragen nur eine einzelne Ausprägung auf 15,8% der Suchanfragen enthielten eine Kombination aus zwei und 2,4% eine Kombination aus drei Kriterien. Am häufigsten waren die Kategorien Stichwort (30,9%), titeltypisches Wort (22,5%), Personennamen (14,5%) sowie Wortverbindungen (9,4%).

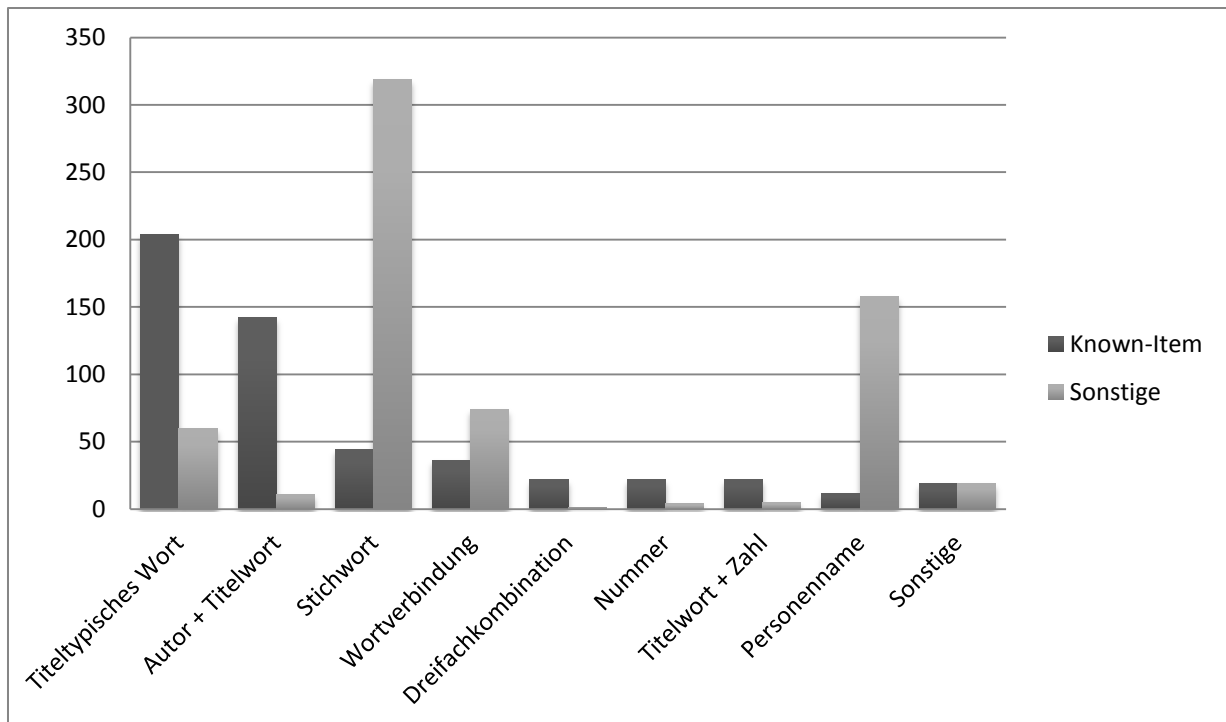


Abbildung 3: Anzahl der Known-Item- und sonstigen Suchanfragen mit bestimmten Merkmalen<sup>18</sup>

Abbildung 3 zeigt, welcher Suchanfragentyp wie oft mit welchen Metadaten gesucht wurde. Für Known-Item-Suchanfragen sind titeltypische Wörter am häufigsten (39%), gefolgt von der Kombination aus Personenname und weiterem Titelwort (27%). Mit diesen beiden Suchmöglichkeiten wurden zwei Drittel aller Known-Item-Suchanfragen gestellt.

Mit 36,5% enthielten relativ viele der Known-Item-Suchanfragen zwei oder drei der zwölf erhobenen Kriterien. Zu der häufigsten Kombination aus Personenname und Titelwort wurde bei Dreifachkombinationen zu diesen noch eine Zahl oder ein Jahr hinzugefügt. Unter Dreifachkombination wurden auch die Suchanfragen eingeordnet, die vermutlich anderswo kopiert und ohne Anpassungen in den Suchschlitz eingefügt wurden. Diese enthielten zudem oft Satzzeichen und Abkürzungen. Zwölf Personensuchen wurden als Known-Item-Suchanfrage klassifiziert. Hier handelt es sich um Suchanfragen mit zwei verschiedenen Personennamen.

Aus Abbildung 3 wird auch deutlich, dass Suchanfragen mit Kombinationen von Metadaten fast immer Known-Item-Suchanfragen sind. Titeltypische Wörter jedoch sind nicht alleine für Known-Item-Suchanfragen charakteristisch. Sie treten auch bei 60 sonstigen Suchanfragen auf, von denen jedoch 24 Stück unsichere Suchanfragen sind.

<sup>18</sup> Zusammengefasst wurden: Titelwort: titeltypisches Wort, Stichwort oder Wortverbindung. Nummer: Standardnummer oder Signatur. Zahl: auch Jahr. Sonstige: Suchschlüssel, Kombinationen aus Personennamen und Zahlen, Phrasenmarkierung.

Außerdem wurden 54% aller Known-Item-Suchanfragen nur mit Wörtern aus dem Titel gesucht. Wenn nur mit Titelwörtern gesucht wurde, stimmten sie in 73,2% der Fälle exakt mit dem (Haupt-) Titel des gesuchten Known-Items überein. Wenn alle Suchanfragen, in denen Titelwörter vorkamen, betrachtet werden, so stimmten immer noch 64,7% der Suchwörter exakt mit dem Titel des Known-Items überein (siehe Tabelle 2).

<b>Titelübereinstimmung</b>	<b>Anteil der Known-Item-Suchanfragen (n=487)</b>	<b>Anteil der unsicheren Suchanfragen (n=62)</b>
Exakter Titel	64,68	62,90
Stichwörter	11,49	14,51
Titel mit fehlendem Wort am Anfang oder Ende	9,85	3,22
Zeitschriftenabkürzung	1,64	4,83
Exakter Titel und Zeitschriftenabkürzung	1,43	0
Titel mit Fehlern in der Suchanfrage (R)	10,88	8,06

**Tabelle 2: Übereinstimmung zwischen Titel des Known-Items und Wörtern der Suchanfragen, die sich auf den Titel beziehen**

### **7.4.3 Signalwörter**

In knapp 15% der Known-Item-Suchanfragen wurden Wörter festgestellt, die als Hinweise auf das Format oder den Medientyp des Known-Items gewertet wurden.

Dies waren: *Bericht, Bulletin, Festschrift, handbook, Handbuch, Heft, Jahrbuch, journal, Lexikon, Magazin, magazine, Mitteilungen, Nachrichten, Report, Rundschau, Wörterbuch, Zeitschrift, Zeitung*. Teilweise wurden diese Wörter nur rechts- und/oder linkstrunkiert gefunden. Am häufigsten waren die Begriffe *Zeitschrift* (14 mal), *Bericht* (11 mal), *Handbuch* (8 mal), *Magazin* (5 mal) und *Zeitung* (5 mal).

## **7.5 Medientypen von Known-Item-Suchanfragen**

Alle Known-Item-Suchanfragen wurden daraufhin geprüft, welchen Medientyp das gesuchte Known-Item hat. Dies war bei vier Suchanfragen nicht möglich, da verschiedene Medientypen auf die Kriterien zutrafen. Die Anteile der Medientypen an den verbleibenden 519 Known-Item-Suchanfragen werden in Abbildung 4 dargestellt. Der Hauptfunktion belugas als Nachweisinstrument für lokale Bestände entsprechend wurden am häufigsten



Bücher gesucht, gefolgt von 15% Zeitschriften und Reihen sowie 13% Artikeln. In geringer Zahl wurden zudem Buchaufsätze und Zeitschriftenbände bzw. -hefte gesucht. Mit jeweils 1-3 Suchanfragen gab es Suchen nach grauer Literatur, Datenbanken und Mikroformen (Sonstige).

Bei den Known-Items konnten 27,9% der Bücher, 65,6% der Artikel und 15,8% der Zeitschriften nicht gefunden werden, da diese nicht im Bestand waren bzw. keine Einzeltitelaufnahme hatten. Bei der Hälfte der Suchanfragen nach Buchaufsätzen wurde über das Inhaltsverzeichnis zumindest das entsprechende Buch gefunden.

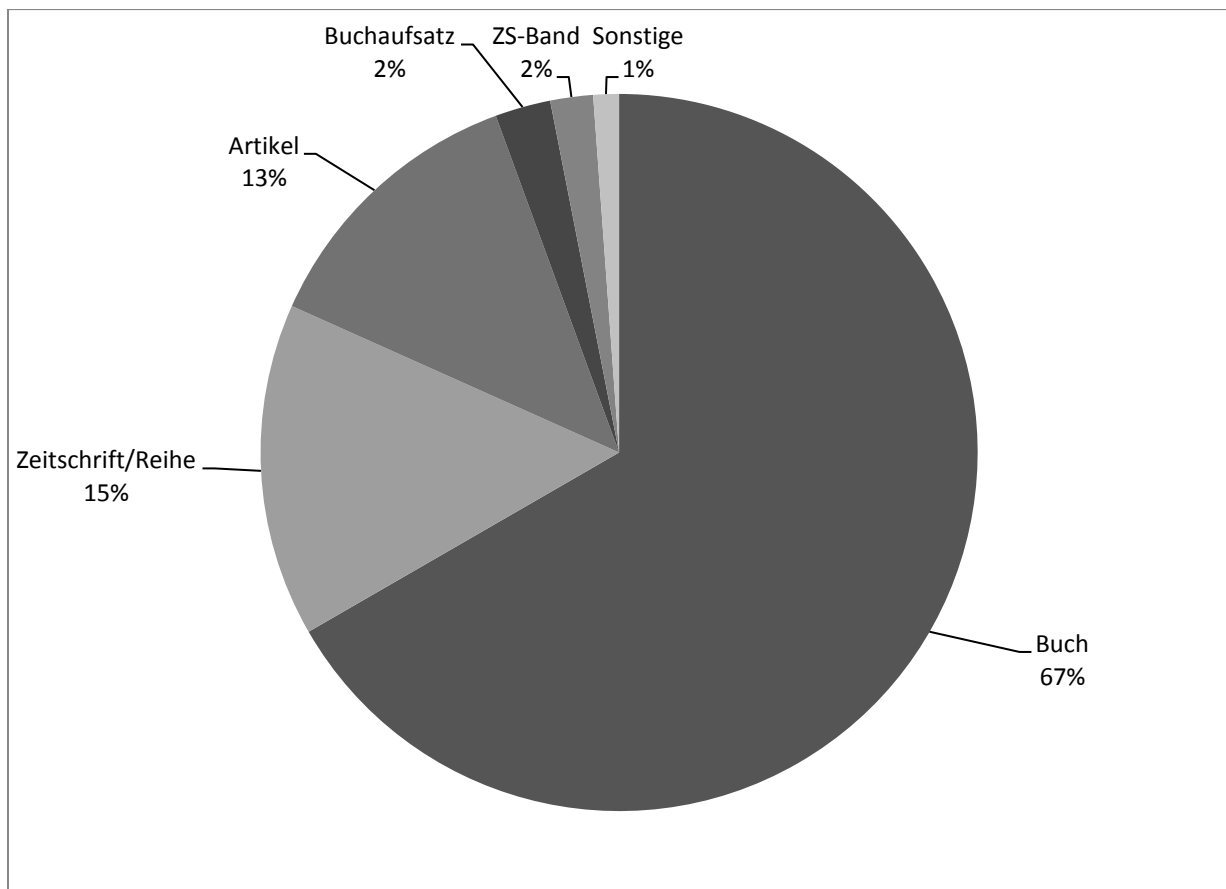


Abbildung 4: Medientypen der Known-Items

## 7.6 Retrievaleffektivität der Known-Item-Suchanfragen

Um die Retrievaleffektivität feststellen zu können, wurden der Rang der Known-Items ermittelt. Es wurde der erste Treffer, der allen Angaben in der Suchanfrage am genauesten entsprach, als Known-Item gewertet. Das heißt, dass bei einer Suche nach einem Zeitschriftentitel die Gesamtaufnahme als Known-Item gewertet wurde, nicht aber einzelne Stücktitel, die vor der Gesamtaufnahme gerankt waren. Bei verschiedenen Formaten eines

Known-Items, z.B. E-Zeitschrift und Printzeitschrift, E-Book und Printausgabe sowie mehreren Auflagen eines Buches, wurde die Position des höchstgerankten Titels berücksichtigt. Üblicherweise werden nur Treffer bis zum Rang 10 berücksichtigt (siehe Kapitel 6.2), da die meisten Bibliothekskataloge und Suchmaschinen auf der ersten Trefferseite nur zehn Suchergebnisse anzeigen. Bei beluga werden 20 Ergebnisse angezeigt, daher wird die Position des Known-Items bis Rang 20 ermittelt. Wegen der geringen Anzahl wurden in den folgenden Abbildungen die Ergebnisse zu den Known-Items zwischen Rang 11 und 20 zusammengezählt.

Abbildung 5 zeigt den Anteil der unsicheren und der Known-Item-Suchanfragen, bei denen das Known-Item bis zu einem Rang n gefunden wurden. Die Angaben beziehen sich nur auf die für die Retrievaleffektivität auswertbaren Known-Item-Suchanfragen, d.h. wenn keine Rechtschreibfehler gemacht wurden und für das Known-Item in beluga eine Titelaufnahme vorhanden ist. Dies sind 321 Known-Item-Suchanfragen und 50 unsichere Suchanfragen. Die fehlenden Prozentpunkte bis 100 bei Rang 11-20 geben an, wie viele Known-Items erst ab der zweiten Trefferseite aufzufinden waren.

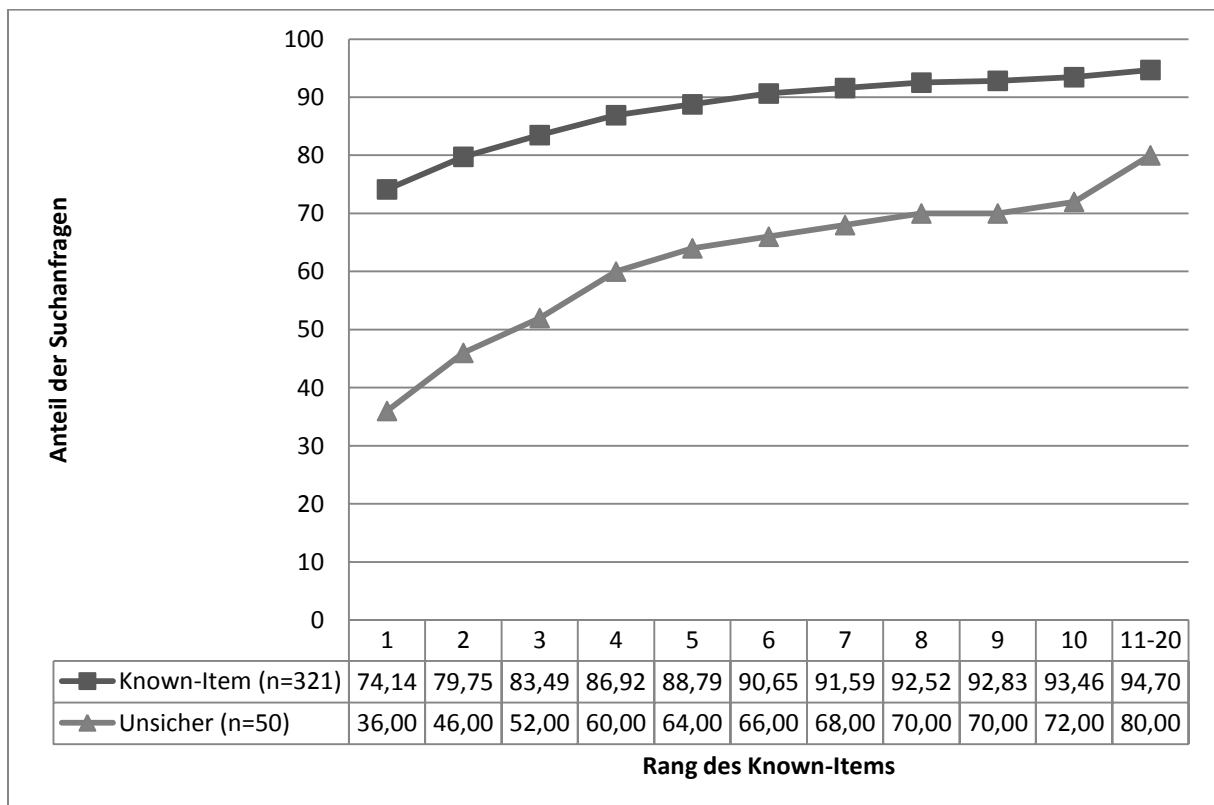


Abbildung 5: Success @n für Known-Item- und unsichere Suchanfragen

Bei 74,1% der Known-Item-Suchanfragen stand das gesuchte Known-Item auf Rang 1, jedoch nur bei 36% der unsicheren Known-Item-Suchanfragen. Auf der ersten Trefferseite, d.h. bis

Rang 20, wurde bei 80% der unsicheren und bei 94,7% der sicheren Known-Item-Suchanfragen das Known-Item gefunden. Der Mean Reciprocal Rank beträgt für Known-Item-Suchanfragen 0,78 und für unsichere Suchanfragen 0,44. Der durchschnittliche Rang eines Known-Items auf der ersten Ergebnisseite beträgt bei Known-Item-Suchanfragen 1,8 und bei unsicheren Suchanfragen 3,7. Die unsicheren Suchanfragen sind also in ihrer Retrievaleffektivität vor allem auf den ersten Rängen deutlich schlechter. Außerdem wird deutlich, dass mit jedem weiteren betrachteten Treffer die Wahrscheinlichkeit zunimmt, dass das Known-Item gerankt wird. Vor allem bei unsicheren Suchanfragen nimmt der Anteil an gefundenen Known-Items pro Rang stark zu, so dass der Anteil von Known-Items von Rang 1 auf Rang 20 um 44 Prozentpunkte steigt. Bei Known-Item-Suchanfragen beträgt der Zugewinn nur 20 Prozentpunkte.

In Abbildung 6 wird Success @n für die drei häufigsten Medientypen von Known-Item-Suchanfragen gezeigt. Etwa 90% der gesuchten Bücher und Artikel sind auf Rang 1, 2 oder 3 platziert. Problematisch ist die Suche nach Zeitschriften und Reihen. Diese werden nur für 42% der entsprechenden Suchanfragen auf Rang 1 und für knapp 20% erst ab der zweiten Trefferseite oder später gefunden. Hier muss allerdings beachtet werden, dass nur die Gesamtaufnahme als Known-Item gewertet wurde. Oftmals wurden Stücktitelaufnahmen für Jahrgänge vor der Gesamtaufnahme gerankt.

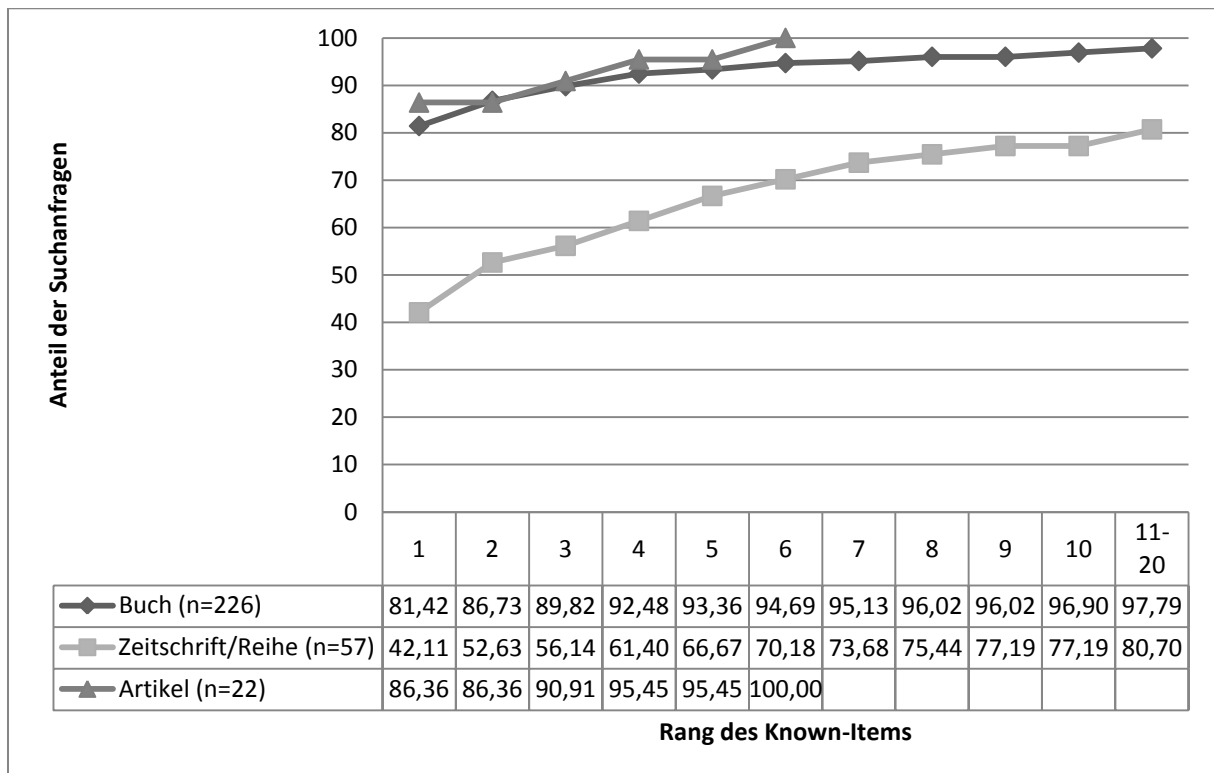


Abbildung 6: Success @n für Known-Item-Suchanfragen nach bestimmten Medientypen

Zuletzt wurde ermittelt, wie erfolgreich Known-Item-Suchanfragen unter Berücksichtigung ihrer Metadaten-Eigenschaften sind. Dies wurde für die vier häufigsten Metadaten von Known-Item-Suchanfragen geprüft. Wenn die Suchanfragen aus einer Kombination von Autor und Titelwort besteht, wird das Known-Item bis zum zehnten Treffer in jedem Fall gefunden (siehe Abbildung 7). Suchanfragen, die nur aus Wörtern des Titels bestehen, sind weniger erfolgreich. Wenn ein titeltypisches Wort enthalten ist, beträgt der Success an Rang 1 gut 70%.

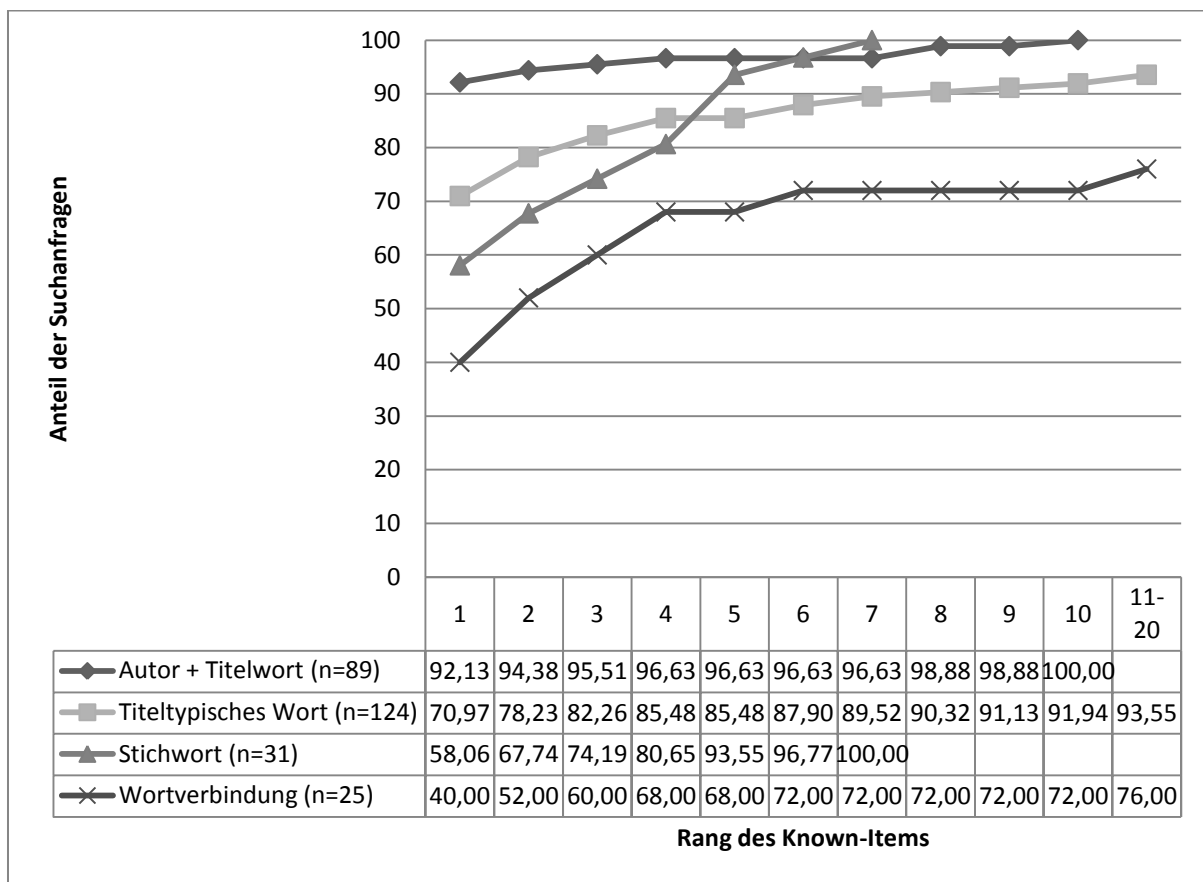


Abbildung 7: Success @n für Known-Item-Suchanfragen mit bestimmten Merkmalen

## 8 Algorithmus zur Erkennung der Known-Item-Suchanfragen

Eine Möglichkeit, die Retrievaleffektivität für Known-Items zu verbessern, ist es, das Relevanzranking für diesen Suchfragentyp anzupassen. Da Known-Items hauptsächlich mit Informationen, die in den Feldern der formalen Beschreibung eines Titels stehen, gesucht werden, könnte man Übereinstimmungen der Suchwörter mit den inhaltlichen Feldern

entweder gar nicht oder weniger stark gewichten. Dies hätte jedoch auch große Auswirkungen auf die Retrievaleffektivität von thematischen oder sonstigen Suchanfragen.

Daher wurde ein Algorithmus entwickelt, der die Known-Item-Suchanfragen an ein Discoverysystem allein anhand ihrer Merkmale erkennen soll. Dieser Algorithmus soll direkt nach Eingabe der Suchanfrage darüber entscheiden, ob eine Known-Item-Suchanfrage vorliegt oder nicht. Wenn es sich um eine Known-Item-Suchanfrage handelt, könnte das alternative Relevanzranking greifen, das die Known-Items in der Relevanzsortierung stark bevorzugt. Es wird vorausgesetzt, dass es möglich ist, ein ideales Relevanzranking für Known-Item-Suchen zu entwickeln.

Bei diesem Ansatz handelt es sich um ein heuristisches Verfahren. Dieser Ansatz unterscheidet sich erheblich von dem Verfahren des maschinellen Lernens, wie es von Kan und Poo (siehe Kapitel 5.2) eingesetzt wurde.

Ein Ansatz hätte sein können, vor allem die Known-Item-Suchen mit einer schlechteren Retrievaleffektivität erkennen zu wollen, da bei diesen am meisten Verbesserungspotential vorhanden ist. In dieser Arbeit wurde stattdessen versucht, einen Algorithmus zu entwickeln, der möglichst alle Known-Item-Suchanfragen erkennt. Dadurch ist eine bessere Vergleichbarkeit zwischen den Ergebnissen der beiden Datensets gegeben. Außerdem kann es wegen der begrenzten Anzahl an untersuchten Suchanfragen und des begrenzten Zeitraums der Datenerhebung sein, dass bei den Suchanfragen des ersten Datensets Merkmale überwiegen, die zu einem anderen Zeitpunkt ganz anders verteilt sein mögen.

## **8.1 Geeignete Merkmale auf Grundlage der intellektuell klassifizierten Suchanfragen**

Um in den Algorithmus aufgenommen zu werden, musste ein Merkmal genügend viele Fallzahlen haben sowie auf möglichst viele Known-Item-Suchanfragen und zugleich auf wenig sonstige Suchanfragen zutreffen. Außerdem mussten die Merkmale für den Algorithmus erkennbar sein. Folgende Merkmale wurden aufgrund der Ergebnisse der intellektuellen Klassifizierung als geeignet erachtet: Anzahl der Wörter, Zahlen, Signalwörter, titeltypische Wörter und Personennamen. Unabhängig von diesen wurden ISBN, ISSN und Signatur getrennt identifiziert. Titeltypische Wörter wurden auf deutsche und englische

Funktionswörter reduziert, d.h. Verben wurden nicht berücksichtigt, ebenfalls nicht „und“, „oder“, „and“ und „or“, da diese auch als Boole'sche Operatoren verwendet werden können.

Um die Eignung eines Merkmals einschätzen zu können, wurde für jedes Merkmal geprüft, wie viel Prozent der mit diesem Merkmal erkennbaren Suchanfragen welchem Suchanfragetyp angehören. Mit Ausnahme von Signalwörtern sowie ISBN, ISSN und Signatur war für alle sonstigen Suchanfragen Bedingung, dass nicht nur das Merkmal, sondern auch mindestens zwei Wörter enthalten waren. Die Ergebnisse sind in Tabelle 3 zu sehen. Am zuverlässigsten mit einem Merkmal ließen sich Known-Item-Suchanfragen mit Signalwörtern erkennen (knapp 90%). Etwa 20% der Suchanfragen, die eine Zahl enthalten, sind keine Known-Item-Suchen.

	<b>Signalwort (n=79)</b>	<b>Zahl und &gt; 1 Wort (n=65)</b>	<b>Funktionswort und &gt; 1 Wort (n=310)</b>	<b>Personenname und &gt; 1 Wort (n=188)</b>	<b>ISBN, ISSN, Signatur (n=23)</b>
<b>Anteil an Suchanfragen mit Merkmal</b>	11,93	9,37	46,83	28,40	3,47
Davon sind:					
<b>Known-Item</b>	89,87	80,65	81,94	85,64	100
<b>Unsicher</b>	6,33	3,23	6,13	4,97	
<b>Sonstige</b>	3,80	16,13	11,94	10,11	

Tabelle 3: Anteil der Suchanfragentypen an den Merkmalen in %

In Tabelle 3 werden Suchanfragen, die mehr als eines dieser Merkmale enthalten, mehrfach gezählt. Mehrere Merkmale treten bei 124 Suchanfragen auf. All diese Merkmale schienen für den Algorithmus geeignet.

## 8.2 Ablauf und Funktionen des Algorithmus

Der Ablauf der Hauptfunktion des Algorithmus wird in Abbildung 8 gezeigt. Dabei werden Wörter im Sinne des Algorithmus wie folgt erkannt:

- **Wort:** Der Suchstring wird anhand vom zusammenhängenden Ketten von Leerzeichen und/oder Satzzeichen getrennt. Ein durch Leerzeichen-/Satzzeichenketten getrenntes Element ist ein Wort. Bindestriche zählen nicht als Satzzeichen.

In den Verzweigungen wird geprüft, ob die folgenden Funktionen wahr oder falsch sind.

- Ist Signatur: Der gesamte Suchstring kommt inklusive Leerzeichen in der Datenbasis im Feld Signatur vor.
- Enthält ISBN: Mindestens ein Wort besteht nach der Entfernung aller Bindestriche aus 10 oder 13 Ziffern, bzw. aus 9 Ziffern und einem x, wobei das jeweils letzte Zeichen dem Ergebnis des entsprechenden ISBN-Prüfsummenverfahren angewandt auf die vorangehenden Zeichen entspricht.
- Enthält ISSN: Mindestens ein Wort besteht aus vier Ziffern, einem Bindestrich und vier Ziffern bestehen. Die letzte Ziffer kann auch ein „x“ sein.
- Enthält Signalwort + drei Zeichen oder mehr: Eine Zeichenkette steht auf der Liste „Signalwort“ und die Anzahl der weiteren Zeichen ist größer als zwei.
- Anzahl Wort=1: Die Anzahl der Wörter wird gezählt. Diese Zahl ist gleich eins.
- Enthält Zahl: Ein Wort besteht nur aus Ziffern.
- Enthält Funktionswort: Ein Wort steht auf der Liste „Funktionswort“.
- Enthält Autor: Führe für die Wörter folgende Suchanfrage durch: ALL Wort\_1 Wort\_n (PER Wort\_1 OR PER Wort\_n)<sup>19</sup>. In mindestens einem der Treffer für alle Wörter wird eines der Wörter im Feld Verfasser gefunden<sup>20</sup>.
- Enthält Vorname: Ein Wort steht auf der Liste „Vorname“.
- Enthält weiteres Wort: Die Anzahl der Wörter ist größer als zwei.

---

<sup>19</sup> ALL=Freitext, PER=Person, Autor, OR=Boole'sches ODER. Diese Notation entspricht einer Suchanfrage in einem Pica-System. Für andere Retrievalsysteme muss sie demensprechend angepasst werden.

<sup>20</sup> Diese Funktion muss noch auf Verbesserungsmöglichkeiten hin geprüft werden, z.B. ob statt der Freitextsuche nur in bestimmten Feldern gesucht werden sollte. Mit dieser Funktion werden Autoren nur erkannt, wenn es auch einen passenden Treffer gibt.

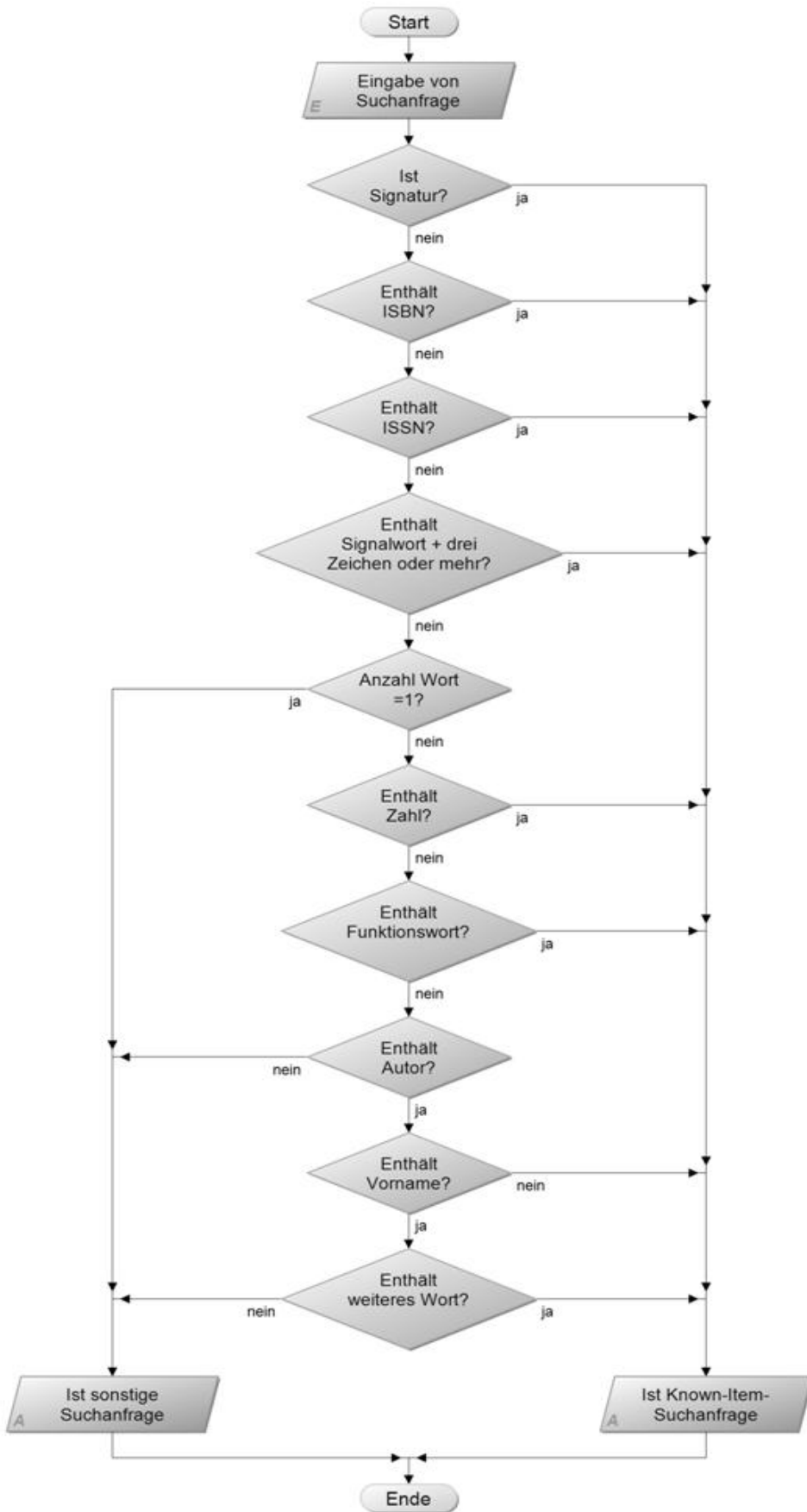


Abbildung 8: Hauptfunktion des Algorithmus zur Erkennung von Known-Item-Suchanfragen



Die Liste „Signalwort“ enthält beispielsweise folgende deutsche und englische Begriffe, die auf eine Publikationsform hinweisen können: *annual, Beiträge, Bericht, Bulletin, dictionary, Encyclopedia, Enzyklopädie, Festschrift, handbook, Handbuch, Heft, Jahrbuch, journal, lexicon, Lexikon, magazin, Mitteilungen, Nachrichten, newspaper, report, Rundschau, Serie, Wörterbuch, Zeitschrift, Zeitung*. Meistens ist eine rechts- und/oder linkstrunkierte Übereinstimmung eines Begriffs ausreichend, daher wird nur die Übereinstimmung einer Zeichenkette geprüft. Die Liste besteht bislang größtenteils aus Begriffen, die in einem der beiden Datensets gefunden wurden, wo sie eindeutig auf Known-Items hinwiesen. So wurde eine vollständige Liste simuliert. Die Begriffe wurden allerdings noch nicht systematisch auf ihre Eignung überprüft. So kommt es durchaus vor, dass mit manchen Signalwörtern gesucht wird, ohne dass sie Teil des Titels sind. Dafür ist eine genauere Untersuchung mit größeren Datenmengen noch nötig.

Die Liste „Funktionswort“ enthält vorwiegend deutsche und englische Präpositionen, Pronomen, Konjunktionen und (meist Hilfs-) Verben mit ihren flektierten Wortformen. Für die Zwecke dieser Untersuchung wurden frei verfügbare deutsche und englische Stoppwortlisten (Bougé 2011) verwendet, von denen Substantive und Zahlwörter entfernt wurden. Insgesamt umfasste die Liste 1004 Einträge. Die einzelnen Wörter müssen noch auf einer größeren Menge an Suchanfragen auf ihre Eignung geprüft werden. Deutschsprachige Beispiele von dieser Liste sind: *auf, aus, bezüglich, binnen, bis, dabei, damit, dann, dass, deiner, der, des, dich, die, du, durch, eine, hinab, im, mit, obwohl, sie, sind, weil, wenn*.

Die Liste „Vorname“ enthält die häufigsten männlichen und weiblichen Vornamen in Deutschland und in den USA sowie Vornamensabkürzungen, insgesamt 2162 Einträge, um das Merkmal über Excel automatisch abfragen zu können. Eine vollständige Liste könnte über einen Abzug des Indexes von beluga erstellt werden.

### **8.3 Klassifikation der Testmenge mit dem Algorithmus**

Da der Algorithmus auf die Merkmale der intellektuell klassifizierten Suchanfragen angepasst war, bestand die Gefahr, dass er nur für diese Suchanfragen zufriedenstellende Ergebnisse liefert. Daher wurde der vorgeschlagene Algorithmus mit einer zufälligen Testmenge evaluiert. Wie bereits in der ersten Datenmenge zur intellektuellen Klassifikation wurden Duplikate entfernt. Die Testmenge bestand aus 1154 Suchanfragen vom 14. bis 16. April 2014.

Merkmale und Algorithmus konnten fast vollständig über Formeln in Excel erhoben bzw. abgebildet werden. Wo es Abweichungen zu dem oben vorgestellten Algorithmus gab, wurde intellektuell der Ablauf des Algorithmus nachvollzogen. Dies war bei folgenden Punkten der Fall:

- Ist/enthält Signatur/ISBN/ISSN: Zeichenketten wurden intellektuell darauf geprüft, ob es sich um eine Signatur, eine ISBN oder ISSN handelte.
- Enthält Signalwort + drei Zeichen oder mehr: Es wurde das Vorhandensein von mindestens drei weiteren Zeichen geprüft.
- Enthält Zahl: Die Excel-Formel gibt auch das Vorhandensein einer Zahl in einem Wort an. Es wurde daher geprüft, dass Kombinationen von Ziffern und Buchstaben in einem Wort nicht erfasst wurden.

Außerdem wurde für die Funktion *Autor* die Autorenerkennung über ein Perl-Skript automatisch im GVK durchgeführt<sup>21</sup>. Prinzipiell ist diese Art der Autorenerkennung in beluga möglich, dafür hätten aber erst technische Anpassungen vorgenommen werden müssen. Allerdings kann es sein, dass dadurch leicht abweichende Ergebnisse ermittelt wurden, da die Indizes von GVK und beluga nicht exakt übereinstimmen.

Mit dem Algorithmus wurden 498 von 1154 Suchanfragen als Known-Item-Suchen erfasst. Dies entspricht 43,15% der Suchanfragen.

### 8.3.1 Evaluierung des Algorithmus

Der Algorithmus teilte die Suchanfragen in die Klassen Known-Item-Suchanfrage und sonstige Suchanfragen ein. Dabei konnten vier Kombinationen von Testergebnissen und tatsächlichen Ergebnissen auftreten:

- Richtig positiv: Eine Known-Item-Suchanfrage wird als Known-Item-Suchanfrage erkannt.
- Falsch negativ: Eine Known-Item-Suchanfrage wird nicht als Known-Item-Suchanfrage erkannt.
- Falsch positiv: Eine sonstige Suchanfrage wird als Known-Item-Suchanfrage erkannt.
- Richtig negativ: Eine sonstige Suchanfrage wird als sonstige Suchanfrage erkannt.

---

<sup>21</sup> Die Autorenprüfung wurde von Dr. Jan Frederik Maas implementiert.

Dieser Algorithmus lässt sich mit den Kennzahlen Precision und Recall evaluieren. Precision bezeichnet den Anteil der richtig erkannten Known-Item-Suchanfragen an allen vom Algorithmus ausgegebenen Known-Item-Suchanfragen. Es wird also ermittelt, wie viele richtig positive und wie viele falsch positive Ergebnisse in der Ausgabemenge vorhanden sind. Recall bezeichnet den Anteil der richtig erkannten Known-Item-Suchanfragen an allen tatsächlichen Known-Item-Suchanfragen. Es wird also ermittelt, wie viele richtig positive und wie viele falsch negative Ergebnisse in der Gesamtmenge aller klassifizierten Suchanfragen vorhanden sind.

Um die Werte berechnen zu können, muss bekannt sein, welche Suchanfragen tatsächlich Known-Item-Suchanfragen sind. Nachdem der Algorithmus die Suchanfragen klassifiziert hatte, wurden daher die Suchanfragen der Testmenge wiederum intellektuell klassifiziert. Für die intellektuelle Klassifikation wurden aus Zeitgründen über eine Excel-Formel 50% der Suchanfragen zufällig ausgewählt. Von diesen 577 Suchanfragen wurden 42,11% intellektuell als Known-Item-Suchanfragen identifiziert.

In Tabelle 4 werden die Anzahl der tatsächlichen, intellektuell klassifizierten Suchanfragentypen dargestellt und wie viele Suchanfragen von dem Algorithmus jeweils erfasst wurden.

	<b>Known-Item-Suchanfrage</b>	<b>Sonstige Suchanfrage</b>	<b>Summe</b>
<b>mit Algorithmus erfasst</b>	207	43	250
<b>mit Algorithmus nicht erfasst</b>	36	291	327
<b>Summe</b>	243	334	577

Tabelle 4: Anzahl der mit dem Algorithmus erfassten Suchanfragen und der intellektuell identifizierten Suchanfragentypen

Es wurden 43,33% aller Suchanfragen von dem Algorithmus als Known-Item-Suchanfragen klassifiziert. Davon sind 82,8% der Suchanfragen richtig positiv und 17,2% falsch positiv erkannt worden. Außerdem wurden 85,2% aller intellektuell klassifizierten Known-Item-Suchanfragen von dem Algorithmus erfasst. Dies entspricht einer Precision von 0,83 und einem Recall von 0,85.

	<b>Signalwort (n=23)</b>	<b>Zahl und &gt; 1 Wort (n=41)</b>	<b>Funktionswort und &gt; 1 Wort (n=157)</b>	<b>Personenname und &gt; 1 Wort (n=86)</b>
<b>Anteil an Suchanfragen mit Merkmal</b>	9,20	16,40	62,80	34,40
Davon sind:				
<b>Known-Item</b>	100	95,12	85,35	75,58
<b>Unsicher</b>			8,28	6,98
<b>Sonstige</b>		4,88	6,37	17,44

**Tabelle 5: Anteil der Suchanfragentypen an den Merkmalen, die mit dem Algorithmus erfasst wurden in %**

In Tabelle 5 wurden Suchanfragen, die mehr als eines dieser Merkmale enthalten, wieder mehrfach gezählt. Im Vergleich zur erwarteten Verteilung aufgrund der Merkmale der intellektuell klassifizierten Suchanfragen (siehe Kapitel 8.1) ist hier am auffälligsten der deutlich höhere Anteil an erkannten Funktionswörtern. Außerdem ist bei den Personennamen der Anteil an sonstigen Suchanfragen sehr groß, was dafür spricht, dass das Verfahren der Autorenerkennung noch überarbeitet werden müsste.

### **8.3.2 Retrievaleffektivität für die erkannten Known-Item-Suchanfragen**

Um einschätzen zu können, wie hoch das Verbesserungspotential der Retrievaleffektivität mit Einsatz des Algorithmus ist, wurden für die erkannten Known-Item-Suchanfragen die Kennzahlen Success @n und Mean Reciprocal Rank erhoben. Von den 207 richtig positiv erfassten Known-Item-Suchanfragen wurden 36% aufgrund nicht vorhandener Known-Items oder Rechtschreibfehler nicht berücksichtigt. Abbildung 9 zeigt Success @n für die auswertbaren Known-Item-Suchanfragen.

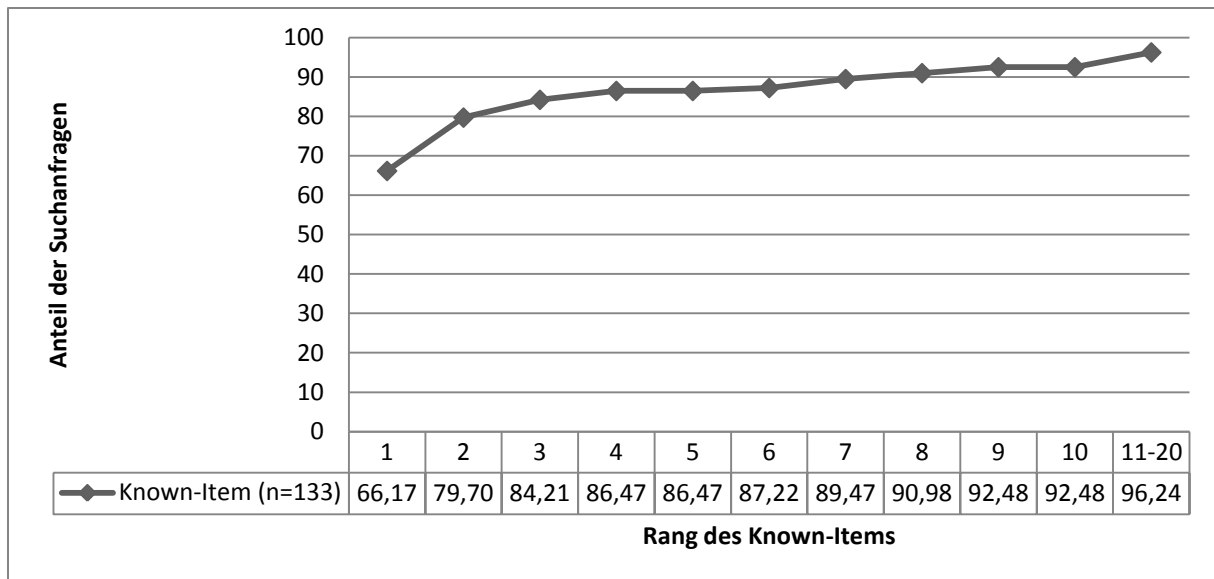


Abbildung 9: Success @n für die Known-Item-Suchanfragen, die mit dem Algorithmus erkannt wurden

In zwei Drittel der Suchanfragen wurde das Known-Item auf Rang 1 gefunden, in 84,2% bis Rang 5 und in 96,2% auf der ersten Trefferseite. Der Mean Reciprocal Rank beträgt 0,74.

Die Retrievaleffektivität ist für diese Suchanfragen also auf den ersten 10 Rängen leicht schlechter als für die intellektuell klassifizierten Known-Item-Suchanfragen aus dem ersten Datenset.

## 9 Diskussion der Ergebnisse

Bevor nun Empfehlungen für die Verbesserung der Retrievaleffektivität von Known-Item-Suchanfragen gegeben werden, soll nun geprüft werden, inwieweit die Hypothesen bestätigt werden konnten.

1. Etwa 50% der Suchanfragen in beluga lassen sich als Suche nach Known-Items identifizieren.

Diese Hypothese konnte bestätigt werden. So ließ sich ein Großteil der Suchanfragen als Known-Item-Suchanfragen identifizieren, nämlich 51,11% unter Berücksichtigung der unsicheren Suchanfragen. Jedoch kann es durchaus sein, dass auch unter den sonstigen Suchen noch weitere Known-Item-Suchanfragen sind, die jedoch nur von der recherchierenden Person als solche benannt werden können.

2. Die Suchanfragen sind einfach formuliert: Am häufigsten werden Titelstichwörter gesucht, Metadaten werden selten kombiniert.

Diese Hypothese trifft auf alle Suchanfragen insgesamt gesehen zu. Suchanfragen sind mit einem Durchschnitt von 3,2 Wörtern etwas länger als in der Web-Suche. Die Stichwortsuche war das am häufigsten ermittelte Merkmal. Die Known-Item-Suchanfragen sind zwar im Schnitt länger als sonstige Suchanfragen, enthalten aber in über der Hälfte der Suchanfragen nur Angaben zum Titel des gesuchten Items. Kombinationen von Metadaten treten nur bei einem guten Drittel der Suchanfragen auf. Zudem werden oft titeltypische Wörter verwendet, was ein Hinweis darauf sein kann, dass die Known-Items häufig mit dem vollständigen Titel und nicht nur mit Titelstichwörtern gesucht werden. Bei drei Viertel der Suchanfragen mit Titelwörtern wurde mit dem vollständigen Titel gesucht.

3. Die Suchanfragen führen in 50% der Fälle auch bei Known-Item-Suchen zu mehr als 10 Treffern.

Die dritte Hypothese konnte nur bedingt bestätigt werden: In etwas weniger als der Hälfte der Known-Item-Suchanfragen werden mehr als 10 Treffer gefunden (44,7%) und nur auf 35,7% der Known-Item-Suchanfragen wird eine überschaubare Treffermenge, d.h. 1 bis 10 Treffer, ausgegeben. Demzufolge sind Trefferanzahlen über 10 nur am häufigsten, wenn null Treffer nicht berücksichtigt werden.

4. Ein Drittel der Known-Items steht nicht an erster Position.

Diese Hypothese kann unter zwei Gesichtspunkten geprüft werden. Wenn nur die Suchen ohne Rechtschreibfehler nach in beluga nachgewiesenen Known-Items berücksichtigt werden, so standen 74% der Known-Items auf Rang 1. Bei unsicheren Suchanfragen war es nur ein Drittel der Known-Items. Nimmt man jedoch alle identifizierten Known-Item-Suchanfragen als Bezugsgröße, so stand nur bei 45% der Suchanfragen das gesuchte Known-Item auf dem ersten Rang. Zudem wurden Bücher sehr viel besser gerankt als Zeitschriften. Selbst wenn nur die ersten drei Treffer berücksichtigt werden, da diese bei den meisten Bildschirmauflösungen ohne scrollen sichtbar sind, hat beluga für das meistgesuchte Medium, nämlich Bücher, immer noch eine Erfolgsrate von knapp 90%. Dies deckt sich damit, dass mit einer Kombination aus Autor und Titelwort 95% der Known-Items bis Rang 3 gefunden werden. Schwierig scheint also nur die Suche nach Zeitschriften zu sein, die nur zu 56% bis Rang 3 und zu 77% bis Rang 10 gefunden werden. Dies ist zum Teil auch der Erhebungsmethode

geschuldet, da bei Suchanfragen ohne Band- oder Heftangabe nur Gesamtaufnahmen von Zeitschriften als Known-Item gewertet wurden. Außerdem wurde festgestellt, dass eine Suche mit titeltypischen Wörtern etwas weniger erfolgreich ist (82% Erfolgsrate bis Rang 3) und Suchen nur mit Stichwörtern oder Wortverbindungen noch seltener dazu führen, dass Known-Items auf oben in der Trefferliste gerankt werden. Die Hypothese lässt sich also in Bezug auf alle Known-Item-Suchanfragen bestätigen sowie für unsichere Suchanfragen und Suchen nach Zeitschriften nur dann, wenn keine Rechtschreibfehler gemacht wurden und das Known-Item im Bestand vorhanden war.

Wie erfolgreich sind also Known-Item-Suchanfragen in beluga? Obwohl zwei der Hypothesen zutreffen, die vermuten lassen, dass Known-Item-Suchanfragen schwierig sein könnten – es werden einfache Suchanfragen gestellt und oft mehr Treffer erzielt als der normale Nutzer gewillt ist durchzugucken –, werden Known-Items gut gefunden, wenn sie vorhanden sind und keine Fehler in der Suchanfrage gemacht werden. Dann kann in 93% der Fälle das gesuchte Known-Item bis Rang 10 gefunden werden, obwohl in 44,7% dieser Suchen mehr als 10 Treffer erzielt wurden. Dies beruht jedoch auf der Annahme, dass Nutzer tatsächlich die ersten 10 Treffer auch betrachten. Eine Möglichkeit, um festzustellen, ob ein Nutzer ein Known-Item auch auf niedrigeren Rängen wahrgenommen hat, war, die Suchanfragen zu zählen, bei denen der Nutzer das Known-Item nicht angeklickt hat. Dies war bei nur 64 von 321 Known-Item-Suchanfragen, d.h. 20% der Fall. Dennoch kann nicht sicher gesagt werden, dass der Nutzer das Known-Item nicht gefunden hat. Möglicherweise ließ sich sein Informationsbedürfnis schon in der Trefferliste befriedigen. Unterstützt wird diese Vermutung dadurch, dass von diesen nicht angeklickten Known-Items 60% auf Rang 1 standen. Im Umkehrschluss heißt das auch, dass 80% der gesuchten Known-Items tatsächlich angeklickt wurden und 60% der nicht angeklickten Known-Items in ihrer Retrievaleffektivität nicht verbessert werden konnten, da sie schon auf Rang 1 stehen. Dies spräche dafür, dass das Relevanzranking von beluga in seiner jetzigen Form ausreichend gute Ergebnisse liefert – mit einigen Ausnahmen.

Misserfolg bei der Known-Item-Suche beruht also meist darauf, dass das gesuchte Objekt nicht in beluga nachgewiesen ist, dass Rechtschreibfehler gemacht werden oder dass ein gesuchter Zeitschriftentitel niedrig gerankt wird.

Es lässt sich also feststellen, dass beluga zumindest für die eindeutig identifizierbaren Known-Item-Suchanfragen eine gute Retrievaleffektivität hat. Geht man jedoch davon aus, dass auch die unsicheren Suchanfragen richtig als Known-Item-Suchanfrage identifiziert wurden, gibt es

einen geringen Anteil an nicht gut gerankten Known-Items. Zudem kann es natürlich auch bei den sonstigen Suchanfragen Suchen nach Known-Items gegeben haben, die allerdings ohne Kenntnis des jeweiligen Informationsbedürfnisses des Nutzers nicht verlässlich eingeordnet werden können. Ein Problem ist auch, dass die Erwartungen der Nutzer und ihr Suchverhalten sich mit fortlaufender Zeit ändern können, u.a. weil beluga als Alternativkatalog zum OPAC zum Zeitpunkt der Untersuchung noch nicht von der großen Masse der potentiellen Nutzer genutzt wurde. Es ist vorstellbar, dass die derzeitigen Nutzer von beluga bei einem Bibliothekskatalog noch kein Relevanzranking erwarten oder nicht daran gewöhnt sind und daher eher bereit sind, das Known-Item auch weiter unten auf der Trefferliste herauszusuchen. Daher ist eine Optimierung des Relevanzrankings immer sinnvoll.

Wie könnte nun die Retrievaleffektivität von Known-Item-Suchanfragen in beluga verbessert werden? Dafür wurde die letzte Hypothese geprüft.

5. Eine Mehrzahl der Known-Item-Suchanfragen lässt sich mittels eines regelbasierten Algorithmus identifizieren.

Auch diese Hypothese konnte bestätigt werden. Mit dem Algorithmus ließen sich 85% der tatsächlichen Known-Item-Suchanfragen der Testmenge mit einem richtig positiven Anteil von 83% Known-Item-Suchanfragen erkennen. Das Ranking von diesen knapp 83% richtig positiv erkannten Suchanfragen ließe sich mithilfe des Algorithmus optimieren, wenn ein perfektes Relevanzranking für Known-Item-Suchanfragen eingesetzt werden könnte. Jedoch ist in der Testmenge nur für etwa 34% der Known-Items eine Verbesserung des Ranges auf Rang 1 möglich. Außerdem muss berücksichtigt werden, dass auch etwa 17% der Suchanfragen falsch positiv klassifiziert wurden. Für diese würde das alternative Relevanzranking nicht relevante Ergebnisse liefern, denn es würde unter Umständen nicht nur die Sortierung der Ergebnisse, sondern auch die Menge der ausgegebenen Treffer beeinträchtigt werden.



## 10 Empfehlungen und Fazit

Wenn der Algorithmus in seiner jetzigen Form eingesetzt werden würde, ließe sich für 30%<sup>22</sup> der auswertbaren Suchanfragen nach Known-Items das Ranking verbessern. Sieht man nur eine Optimierung auf die ersten drei Ränge als absolut notwendig an, so sind dies nur noch 14%. Allerdings gibt es mit 17% falsch positiv klassifizierten Suchanfragen auch einen nicht geringen Anteil, der dadurch vermutlich ein schlechteres Relevanzranking bekommen würde. Mit dem Algorithmus sollte jedoch das Ziel verfolgt werden, das Relevanzranking für einen Suchanfragetyp verbessern zu können, ohne das Relevanzranking für sonstige Suchanfragen negativ zu beeinflussen. Der positive Effekt des besseren Rankings für Known-Items scheint nicht groß genug zu sein, um die negative Auswirkung auf sonstige Suchanfragen aufwiegen zu können. Daher müsste der Algorithmus vor einer Umsetzung noch umfangreich darauf getestet werden, wie die falsch positiv klassifizierten Suchanfragen minimiert werden können oder in welchem Maße das alternative Relevanzranking das Ranking von sonstigen Suchanfragen beeinträchtigt.

Außerdem sollte gezielt geprüft werden, wie das Relevanzranking von Zeitschriften verbessert werden kann. Ein Grund für das schlechte Abschneiden von Zeitschriften in der Auswertung ist der Erhebungsmethode geschuldet, die nur die Gesamtaufnahmen als Known-Item gewertet hat. Zeitschriftenbände werden jedoch vor der Gesamtaufnahme gerankt, da die Bandaufnahmen ein aktuelleres Erscheinungsjahr haben und die Aktualität von Titeln im Relevanzranking berücksichtigt wird. Bei laufenden Zeitschriften wird in der Kurzanzeige der Gesamtaufnahmen das erste Erscheinungsjahr als „seit YYYY“ angezeigt. Stücktitelaufnahmen zu einer Zeitschrift werden u.a. auch deshalb angezeigt, da Nutzer, die nach einem bestimmten Jahrgang suchen, bei Gesamtaufnahmen aufgrund der fehlenden gesuchten Jahreszahl denken, dass dies nicht der gewünschte Titel ist (vgl. DENTON & COYSH 2011, S. 19). Dennoch sollte überprüft werden, ob die Anzeige von Stücktitelaufnahmen tatsächlich den Mehrwert bringt, wie vermutet wird. Denn Trefferseiten werden bei der Suche nach einer Zeitschrift oft mit vielen Stücktitelaufnahmen zu dieser Zeitschrift „überschwemmt“, zum einen weil etwa doppelt so viel Stücktitelaufnahmen wie Gesamtaufnahmen verzeichnet sind, zum anderen weil Zeitschriften als Dokumenttyp geboostet werden. Problematisch ist dies, wenn der Nutzer tatsächlich nicht diese Zeitschrift

---

<sup>22</sup> Das ist der Anteil der mit dem Algorithmus erfassten auswertbaren Known-Item-Suchanfragen, die nicht auf Rang 1 standen (n=45), an allen intellektuell klassifizierten auswertbaren Known-Item-Suchanfragen in der Testmenge (n=150).

gesucht hat. Stattdessen könnten die Einzelaufnahmen der Zeitschriftenbände nur innerhalb der Gesamtaufnahme verlinkt werden. Dafür sollte mithilfe von Usabilitytests geprüft werden, welche Lösung dem Nutzer mehr Sucherfolg bringt, d.h. ob Nutzer auch über Stücktitelaufnahmen die gewünschten Informationen zu einem Zeitschriftentitel finden, oder ob sie durch die „Überschwemmung“ mit Stücktitelaufnahmen irritiert werden und die Suche nach anderen Medien als Zeitschriften erschwert wird<sup>23</sup>. Wenn letzteres der Fall ist, wäre zu überlegen, ob die Anzeige von Einzelaufnahmen für jeden Zeitschriftenband in der Trefferliste überhaupt sinnvoll ist. Stattdessen könnte in der Anzeige des Erscheinungsverlaufs von laufenden Zeitschriften das aktuelle Jahr oder ein ergänzender Text wie „bis heute“ automatisch ergänzt werden. Die Ergänzung des aktuellen Jahres hätte auch positive Auswirkungen auf das Ranking der Gesamtaufnahme.

Ein Problem bei der Suche nach Zeitschriften ist außerdem, dass sie oftmals nur kurze Titel haben. Daher wurden relativ viele Suchanfragen nach Zeitschriften als unsichere Suchanfragen klassifiziert oder wurden möglicherweise oft nicht als solche erkannt. In Solr/Lucene lassen sich mit dem QueryElevationComponent (siehe Kapitel 3.1) für Suchanfragen bevorzugte Treffer anlegen, die immer auf diese Suchanfrage an erster Stelle gerankt werden. Der bevorzugte Treffer kann auch in der Trefferliste hervorgehoben werden (vgl. SOLR 2014). Es wäre empfehlenswert, zu den häufigsten Suchanfragen nach Zeitschriften sowie zu oft gesuchten Zeitschriften mit kurzem oder sehr allgemeinem Titel solche bevorzugten Treffer anzulegen. Auch für oft gesuchte Zeitschriftenabkürzungen wäre dies sinnvoll. Dieses Verfahren ist jedoch mit einigem Aufwand verbunden.

Ergänzend dazu könnte ein Aspekt zur Erkennung von Suchen nach Zeitschriftentiteln aus dem vorgestellten Algorithmus zur Erkennung von Known-Item-Suchanfragen übernommen werden. Dies wäre die Liste mit Signalwörtern, die auf zeitschriftentypische Begriffe beschränkt werden könnte. Die Eignung der Begriffe müsste genau überprüft werden, unter Umständen ist auch die Aufnahme weiterer Publikationsformate sinnvoll. Außerdem ließe sich möglicherweise die Liste mit Funktionswörtern soweit verbessern, dass weniger Suchanfragen als derzeit 15% fehlklassifiziert werden. Die Ergebnisse der intellektuellen Klassifikation zeigten, dass titeltypische Wörter am häufigsten Merkmal von Known-Item-Suchanfragen sind und dass bei diesen nur 70% der Known-Items auf Rang 1 standen. Der negative Effekt der Fehlklassifizierung ließe sich auch dadurch verringern, dass beim

---

<sup>23</sup> Der ideale Suchweg wäre in diesem Fall, die gewünschte Medienart über die Facetten auszuwählen. In dieser Arbeit wurde die Nutzung der Facetten nicht berücksichtigt, außerdem ist die Nutzung eines Filters immer eine zusätzliche Hürde, die nicht alle Nutzer auf sich nehmen.

alternativen Relevanzranking nicht ausschließlich das perfekte Ranken von Known-Items angestrebt wird – dann würden vermutlich Felder mit Inhaltsbezug ignoriert werden. Stattdessen könnten Titelfelder mit nebeneinanderstehenden Wörtern stärker geboostet werden. Für das jetzige Ranking wäre solch eine Maßnahme nicht unbedingt sinnvoll, da mehr sonstige Suchen Wortverbindungen enthalten als Known-Item-Suchen.

In dieser Arbeit wurde festgestellt, dass die Retrievaleffektivität für den Großteil der Known-Item-Suchanfragen in beluga gut ist. Dieses Ergebnis gilt jedoch nur für die eindeutigen Known-Item-Suchanfragen. Für Suchanfragen mit wenigen Wörtern, wie sie vermutlich bei Suchen nach Zeitschriften oder Büchern mit kurzem Titel vorkommen, war der Unsicherheitsfaktor bei der intellektuellen Klassifikation zu hoch. Solche kurzen Suchanfragen haben außerdem selten solche Merkmale, mit denen der Algorithmus sie erfassen könnte. Für diese Suchanfragen müssten Anpassungen im allgemeinen Relevanzranking vorgenommen werden, so dass z.B. Übereinstimmungen mit dem Titel eines Mediums höher gerankt werden als Übereinstimmungen in anderen Feldern.

Falls das derzeitige Relevanzranking von beluga zukünftig noch verändert werden sollte, kann mit den hier gewonnen Ergebnissen überwacht werden, wie diese Änderungen sich auf die Retrievaleffektivität von Known-Item-Suchanfragen auswirken. Ein möglicher nächster Schritt wäre die Klassifikation mithilfe von überwachten selbstlernenden Verfahren, z. B. mit einer Support Vector Machine oder einem Bayes-Klassifikator. Die hier gewonnenen Erfahrungen mit dem heuristischen Verfahren könnten Hinweise auf gute Vorverarbeitungsschritte für solche lernenden Verfahren ergeben.

## 11 Literaturverzeichnis

### **Allen 1989**

ALLEN, Bryce: Recall cues in known-item retrieval. In: *Journal of the American Society for Information Science* Bd. 40 (1989), Nr. 4, S. 246–252

### **Apache 2014**

THE APACHE SOFTWARE FOUNDATION: *Log Files - Apache HTTP Server Version 2.2*.  
<http://httpd.apache.org/docs/2.2/logs.html>. – Abruf: 2014-11-30

### **Bade 2007**

BADE, David: Relevance ranking is not relevance ranking or, when the user is not the user, the search results are not search results. In: *Online Information Review* Bd. 31 (2007), Nr. 6, S. 831–844

### **Baeza-Yates 2011**

BAEZA-YATES, Ricardo: *Modern information retrieval : the concepts and technology behind search*. 2. Aufl. Harlow u.a. : Addison-Wesley/Pearson 2011. — ISBN 0321416910

### **Barton und Mak 2012**

BARTON, Joshua ; MAK, Lucas: Old Hopes, New Possibilities: Next-Generation Catalogues and the Centralization of Access. In: *Library Trends* Bd. 61 (2012), Nr. 1, S. 83–106

### **beluga 2013**

STAATS- UND UNIVERSITÄTSBIBLIOTHEK HAMBURG CARL VON OSSIETZKY (Hrsg.): *beluga » Blog Archiv » Usability und Geschwindigkeit*. – Stand: 2013-10-10. <http://beluga-blog.sub.uni-hamburg.de/blog/2013/10/10/usability-und-geschwindigkeit/>. – Abruf: 2014-11-30

### **Bougé 2011**

BOUGÉ, Kevin: *Download stop words*. – Stand: 2011-11-29  
<https://sites.google.com/site/kevinbouge/stopwords-lists>. – Abruf: 2014-11-27

### **Brenes et al. 2009**

BRENES, David J. ; GAYO-AVELLO, Daniel ; PÉREZ-GONZÁLEZ, Kilian: Survey and Evaluation of Query Intent Detection Methods. In: *Proceedings of the 2009 Workshop on Web Search Click Data, WSCD '09*. New York, NY : ACM 2009. — ISBN 978-1-60558-434-8, S. 1–7

### **Broder 2002**

BRODER, Andrei: A taxonomy of web search. In: *ACM Sigir forum*. Bd. 36 2002, S. 3–10

### **Chapman et al. 2013**

CHAPMAN, Suzanne ; DESAI, Shevon ; HAGEDORN, Kat ; VARNUM, Ken ; MISHRA, Sonali ; PIACENTINE, Julie: Manually Classifying User Search Queries on an Academic Library Web Site. In: *Journal of Web Librarianship* Bd. 7 (2013), Nr. 4, S. 401–421

### **Chowdhury et al. 2008**

CHOWDHURY, Gobinda G. ; BURTON, Paul F. ; MCMENEMY, David ; POULTER, Alan: *Librarianship: an introduction*. London : Facet Publishing 2008. — ISBN 9781856046176

**Christensen 2009**

CHRISTENSEN, Anne: Partizipative Entwicklung von Diensten in der Bibliothek 2.0: Methoden und Ergebnisse aus Katalog 2.0-Projekten. In: *Bibliotheksdienst* Bd. 43 (2009), Nr. 5, S. 527–537

**Christensen 2010**

CHRISTENSEN, Anne: Katalog 2.0 im Eigenbau. Das beluga-Projekt der Hamburger Bibliotheken. In: BERGMANN, Juliane ; DANOWSKI, Patrick (Hrsg.): *Handbuch Bibliothek 2.0*. Berlin [u.a.] : De Gruyter Saur 2010. — ISBN 978-3-11-023210-3, S. 317–332

**Christensen 2014**

CHRISTENSEN, Anne: *A growing organism : Über demografischen Wandel in Bibliotheken - und, natürlich, auch ein bisschen über Discovery*. – Stand: 2014-01-11.  
<http://xenzen.wordpress.com/2014/01/11/eine-art-neujahrs-post/>. – Abruf: 2014-11-30

**Cmor & Litwin 2014**

CMOR, Dianne ; LITWIN, Rory: Should We Retire the Catalog? In: *Reference & User Services Quarterly* Bd. 53 (2014), Nr. 3, S. 213–216

**Craswell et al. 2004**

CRASWELL, Nick ; HAWKING, David ; WILKINSON, Ross ; WU, Mingfang: Overview of the TREC 2003 Web Track. In: *TREC*. Bd. 3 2004, S. 15

**Dellit & Boston 2007**

DELLIT, Alison ; BOSTON, Tony: *Relevance ranking of results from MARC-based catalogues: from guidelines to implementation exploiting structured metadata*. National Library of Australia Staff Papers 2007. – Online verfügbar unter: <http://www-prod.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1052>. – Abruf: 2014-12-01

**Denton & Coysh 2011**

DENTON, William ; COYSH, Sarah J.: Usability testing of VuFind at an academic library. In: *Library Hi Tech* Bd. 29 (2011), Nr. 2, S. 301–319

**Dwyer et al. 1991**

DWYER, Catherine ; GOSSEN, Eleanor A. ; MARTIN, Lynne M.: Known-Item Search Failure in an OPAC. In: *RQ* Bd. 31 (1991), Nr. 2, S. 228–36

**Fry 2011**

FRY, Amy: *ACRLog : Embracing Discovery*. – Stand: 2011-05-26.  
<http://acrlog.org/2011/05/26/embracing-discovery/>. - Abruf 2014-11-30

**Gantert & Hacker 2008**

GANTERT, Klaus ; HACKER, Rupert: *Bibliothekarisches Grundwissen*. 8., vollst. neu bearb. und erw. Aufl. München : Saur 2008. — ISBN 978-359-81177-1-8

**Grant 2014**

GRANT, Carl: *Thoughts from Carl Grant : Do-they-or-don't-they? EBSCO Information Services; Part 2*. – Stand: 2014-04-24. <http://thoughts.care-affiliates.com/2014/04/do-they-or-dont-they-ebSCO-information.html>. – Abruf: 2014-11-30

**Hamilton & Thurlow 2005**

HAMILTON, Scott ; THURLOW, Helen: Transaction log analysis@ State Library of Queensland. In: *3rd International Evidence Based Librarianship Conference*. Brisbane, Queensland. Bd. 17 2005

### **Hänger 2013**

HÄNGER, Christian: *Discovery-System und Organisation – was ändert sich? Was bleibt?* DINI-AG E-Framework 2013. <http://dini.de/fileadmin/workshops/informationsinfrastruktur-gestalten/08-Haenger-Primo.pdf>. – Abruf: 2014-11-30

### **Hennies & Hochstein 2006**

HENNIES, Markus ; HOCHSTEIN, Juliane: Clients information seeking behaviour. In: . [S.l.] : *Australian Library and Information Association* 2006 — ISBN 978-0-86804-564-1, S. 1–10

### **Hessel & Fransen 2012**

HESSEL, Heather ; FRANSEN, Janet: Resource Discovery: Comparative Survey Results on Two Catalog Interfaces. In: *Information Technology & Libraries* Bd. 31 (2012), Nr. 2, S. 21–44

### **Hjørland 2010**

HJØRLAND, Birger: The foundation of the concept of relevance. In: *Journal of the American Society for Information Science and Technology* Bd. 61 (2010), Nr. 2, S. 217–237

### **Höchstötter 2007**

HÖCHSTÖTTER, Nadine: Suchverhalten im Web. Erhebung, Analyse und Möglichkeiten. In: *Information Wissenschaft und Praxis* Bd. 58 (2007), Nr. 3, S. 135–140

### **Hofmann & Yang 2012**

HOFMANN, Melissa A. ; YANG, Sharon Q.: “Discovering” what’s changed: a revisit of the OPACs of 260 academic libraries. In: *Library Hi Tech* Bd. 30 (2012), Nr. 2, S. 253–274

### **Jansen & Pooch 2001**

JANSEN, Bernard J. ; POOCH, Udo: A review of Web searching studies and a framework for future research. In: *Journal of the American Society for Information Science and Technology* Bd. 52 (2001), Nr. 3, S. 235–246

### **Jansen & Spink 2003**

JANSEN, Bernard J. ; SPINK, Amanda: An Analysis of Web Documents Retrieved and Viewed. In: *The 4th International Conference on Internet Computing*. Las Vegas, Nevada 2003, S. 65–69

### **Jansen et al. 2010**

JANSEN, Heiko ; KEMNER-HEEK, Kirstin ; SCHWEITZER, Roswitha: *Konkurrenzanalyse ausgewählter kommerzieller Suchindizes*. Köln : Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen [u.a.] 2010

### **Jones et al. 2000**

JONES, Steve ; CUNNINGHAM, Sally Jo ; MCNAB, Rodger ; BODDIE, Stefan: A transaction log analysis of a digital library. In: *International Journal on Digital Libraries* Bd. 3 (2000), Nr. 2, S. 152–169

### **Kan & Poo 2005**

KAN, Min-Yen ; POO, Danny C.C.: Detecting and supporting known item queries in online public access catalogs. In: *Digital Libraries 2005. JCDL '05. Proceedings of the 5th ACM/IEEE-CS Joint Conference*. Denver : IEEE 2005, S. 91–99

#### **Kilgour 1995**

KILGOUR, Frederick G.: Effectiveness of surname-title-words searches by scholars. In: *Journal of the American Society for Information Science* Bd. 46 (1995), Nr. 2, S. 146–151

#### **Kilgour 2001**

KILGOUR, Frederick G.: Known-item online searches employed by scholars using surname plus first, or last, or first and last title words. In: *Journal of the American Society for Information Science and Technology* Bd. 52 (2001), Nr. 14, S. 1203–1209

#### **Kilgour 2004**

KILGOUR, Frederick G.: An experiment using coordinate title word searches. In: *Journal of the American Society for Information Science and Technology* Bd. 55 (2004), Nr. 1, S. 74–80

#### **Kilgour & Moran 2000**

KILGOUR, Frederick G. ; MORAN, Barbara B.: Surname plus recallable title word searches for known items by scholars. In: *Journal of the American Society for Information Science* Bd. 51 (2000), Nr. 1, S. 83–89

#### **Kilgour et al. 1999**

KILGOUR, Frederick G. ; MORAN, Barbara B ; BARDEN, John R.: Retrieval effectiveness of surname-title-word searches for known items by Academic Library users. In: *Journal of the American Society for Information Science* Bd. 50 (1999), Nr. 3, S. 265–270

#### **Kneifel 2010**

KNEIFEL, Fabienne: Der Katalog 2.0: Mit Web 2.0 zum Online-Katalog der nächsten Generation. In: BERGMANN, Juliane ; DANOWSKI, Patrick (Hrsg.): *Handbuch Bibliothek 2.0*. Berlin [u.a.] : De Gruyter Saur 2010. — ISBN 978-3-11-023210-3, S. 37–62

#### **Krebs 2013**

KREBS, Matthias: Die known-item search in Bibliothekskatalogen – Nutzerverhalten, Probleme, Lösungsmöglichkeiten. In: *Bibliothek, Forschung und Praxis* Bd. 37 (2013), Nr. 1, S. 91–102

#### **Langenstein & Maylein 2009**

LANGENSTEIN, Annette ; MAYLEIN, Leonhard: Relevanz-Ranking im OPAC der Universitätsbibliothek Heidelberg. In: *BIT online* Bd. 12 (2009), Nr. 4, S. 408–413

#### **Lee et al. 2006**

LEE, Jin Ha ; RENEAR, Allen ; SMITH, Linda C.: Known-Item Search: Variations on a Concept. In: *Proceedings of the American Society for Information Science and Technology* Bd. 43 (2006), Nr. 1, S. 1–17

#### **Lewandowski 2010**

LEWANDOWSKI, Dirk: Der OPAC als Suchmaschine. In: BERGMANN, Juliane ; DANOWSKI, Patrick (Hrsg.): *Handbuch Bibliothek 2.0*. Berlin [u.a.] : De Gruyter Saur 2010. — ISBN 978-3-11-023210-3, S. 87–107

**Lewandowski 2011a**

LEWANDOWSKI, Dirk: Evaluierung von Suchmaschinen. In: LEWANDOWSKI, Dirk (Hrsg.): Handbuch Internet-Suchmaschinen: Neue Entwicklungen in der Web-Suche. Bd. 2. Heidelberg : AKA, Akad. Verl.-Ges. 2011. — ISBN 978-3-89838-651-7, S. 203–228

**Lewandowski 2011b**

LEWANDOWSKI, Dirk: The retrieval effectiveness of search engines on navigational queries. In: *Aslib Proceedings* Bd. 63 (2011), Nr. 4, S. 354–363

**Lewandowski 2014**

LEWANDOWSKI, Dirk: Wie lässt sich die Zufriedenheit der Suchmaschinennutzer mit ihren Suchergebnisse erklären? In: KRAH, Hans; MÜLLER-TERPITZ, Ralf: *Suchmaschinen*. Münster : LIT 2014 (Passauer Schriften zur interdisziplinären Medienforschung 4), S. 35–52

**Lewandowski et al. 2012**

LEWANDOWSKI, Dirk; DRECHSLER, Jessica ; VON MACH, Sonja: Deriving query intents from web search engine queries. In: *Journal of the American Society for Information Science and Technology* Bd. 63 (2012), Nr. 9, S. 1773–1788

**MacFarlane 2007**

MACFARLANE, A.: Evaluation of web search for the information practitioner. In: *Aslib Proceedings* Bd. 59 (2007), Nr. 4/5, S. 352–366

**Malliari & Kyriaki-Manessi 2007**

MALLIARI, Aphrodite ; KYRIAKI-MANESSI, Daphne: Users' behaviour patterns in academic libraries' OPACs: a multivariate statistical analysis. In: *New Library World* Bd. 108 (2007), Nr. 3/4, S. 107–122

**Maylein & Langenstein 2013**

MAYLEIN, Leonhard ; LANGENSTEIN, Annette: Neues vom Relevanz-Ranking im HEIDI-Katalog der Universitätsbibliothek Heidelberg. In: *BIT online* Bd. 16 (2013), Nr. 3, S. 190–200

**McKay & Buchanan 2011**

MCKAY, Dana ; BUCHANAN, George: One of these things is not like the others: how users search different information resources. In: *Research and Advanced Technology for Digital Libraries* (2011), S. 260–271

**Meadow & Meadow 2012**

MEADOW, Kelly ; MEADOW, James: Search Query Quality and Web-Scale Discovery: A Qualitative and Quantitative Analysis. In: *College & Undergraduate Libraries* Bd. 19 (2012), Nr. 2-4, S. 163–175

**Mizzaro 1997**

MIZZARO, Stefano: Relevance: The whole history. In: *Journal of the American Society for Information Science* Bd. 48 (1997), Nr. 9, S. 810–832

**Petras 2013**

PETRAS, Vivien: Methoden für die Evaluation von Informationssystemen. In: UMLAUF, Konrad (Hrsg.): *Handbuch Methoden der Bibliotheks- und Informationswissenschaft*



*Bibliotheks-, Benutzerforschung, Informationsanalyse*. Berlin [u.a.] : De Gruyter 2013. — ISBN 9783110255546

#### **Piwik 2014**

PIWIK.ORG: *How is a 'visit' defined in Piwik?* [http://piwik.org/faq/general/faq\\_36/](http://piwik.org/faq/general/faq_36/). – Abruf: 2014-12-01

#### **Roscher 2014**

ROSCHER, Mieke: *Fachdisziplinäre Bedürfnisse in der Gestaltung von Discovery-Lösungen: Wirklich ein Katalog für alle?* Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin 2014 (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft 356)

#### **Sadeh 2013**

SADEH, Tamar: *From search to discovery*. Paper presented at: IFLA WLIC 2013 - Singapore - Future Libraries: Infinite Possibilities in Session 98 - Knowledge Management with Academic and Research Libraries.

#### **Saracevic 2007**

SARACEVIC, Tefko: Relevance Part II: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. In: *Journal of the American Society for Information Science and Technology* Bd. 58 (2007), Nr. 13, S. 1915–1933

#### **Schneider 2009**

SCHNEIDER, René: OPACs, Benutzer und das Web. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Bd. 56 (2009), Nr. 1, S. 9–19

#### **Schneider & Bugnon 2008**

SCHNEIDER, René ; BUGNON, Nicolas: *OPACs and Users - The ACUEIL Project Explains Online Search Behaviour and Outlines Simple and Flexible Search Methodologies* (SSRN Scholarly Paper Nr. ID 1128055). Rochester, NY : Social Science Research Network 2008

#### **Slaven et al. 2011**

SLAVEN, Cathy ; EWERS, Barb ; VOLLMERHAUSE, Kurt: From “ I hate it ” to “ it’s my new best friend! ”: Making heads or tails of client feedback to improve our new Quick Find discovery service. In: *ALIA Information Online* Bd. 2 (2011), Nr. 2, S. 16

#### **Slone 2000**

SLONE, Debra J.: Encounters with the OPAC: On-line searching in public libraries. In: *Journal of the American Society for Information Science* Bd. 51 (2000), Nr. 8, S. 757–773

#### **Solr 2014**

*Solr Wiki : QueryElevationComponent*. – Stand: 2014-02-19  
<https://wiki.apache.org/solr/QueryElevationComponent>. – Abruf: 2014-09-13

#### **Steilen 2012**

STEILEN, Gerald: *Discovery-Systeme - die OPACs der Zukunft?* Vortrag auf dem Bibliothekartag 2012-05-22. <http://www.opus-bayern.de/bib-info/volltexte/2012/1188/>. – Abruf: 2014-11-30

**Wildemuth & O'Neill 1995**

WILDEMUTH, Barbara M. ; O'NEILL, Ann L.: Research Notes The „Known“ in Known-Item Searches: Empirical Support for User-Centered Design. In: *College & Research Libraries* Bd. 56 (1995), Nr. 3, S. 265–281

**Wolfram 2008**

WOLFRAM, Dietmar: Search characteristics in different types of Web-based IR environments: Are they the same? In: *Information Processing & Management* Bd. 44 (2008), Nr. 3, S. 1279–1292

**Zamora et al. 2014**

ZAMORA, Juan ; MENDOZA, Marcelo ; ALLENDE, Héctor: Query intent detection based on query log mining. In: *Journal of web engineering* Bd. 13 (2014), Nr. 1-2, S. 24–52

## Anhang

### Legende für die folgenden Tabellen

Suchanfragentyp	Merkmale	Rang des Known-Items
0: Sonstige 1: Known-Item 2: Unsicher	A: Personennamen B: Zahl mit 1-3 Ziffern CP: Copy+Paste D: Zahl mit 4 Ziffern I: ISBN, ISSN K: Stichwort N: Wortverbindung O: Ort P: Phrase S: Signatur T: Titeltypische Wörter: Funktionswörter und Verben V: Verlag X: Sonstige/Suchschlüssel	999: nicht im Bestand 1111: vorhanden Medien, aber fehlerhafte Suchanfrage

## A. Auszug aus den intellektuell klassifizierten Suchanfragen und Suchanfragentyp (zufällige Auswahl)

Aufgrund des Umfangs der gesammelten Daten werden die vollständigen Daten auf einem externen Datenträger (CD) dieser Arbeit beigelegt.

Suchanfrage	Suchanfragentyp
urban diy	0
wirtschaft	0
Abwicklungsplan	0
Die Formale Darstellung der Gegenwart und die Aufdeckung der Politischen Wirklichkeit in la Muerte de Artemio Cruz	1
Strafverteidiger StV	1
Entdeckend üben – ühend entdecken	1
pflegelexikon	2
StraFo	1
perspektiven des sachunterrichts	1
doaj	1
Juristische Rundschau	1
Bankenregulierung	0
Rechnungslegung pellend	1
"Industrie 4.0"	0
investitionsgüter b2b	0
ollegium (Helsinki: Helsinki Collegium for Advanced Studies) 5	1
Methoden der empirischen Sozialforschung	1
Ebert Schifferer	0
Die sozialgeschichtliche Entwicklung der Straße	1
kurunmaki	0
werte statt grenzen	1
Moderation von Arbeitsgruppen und Qualitätszirkeln : Ein Handbuch	1
kliebisch, udo, 2013	1
grundschule	0
hamburger sportverein	0
strategisches Managemenet Haas	1
he Systematic Identification of Flavonoids	1
Integriertes Reporting going public	1
9783531172385	1
stormaner hefte	1
kusukawa	0
berliner bogen	0
schulautonomie in europa	1
Japan guide	0
merianstich	0
Goldstein, Paul copyright	1
Clemens wöllner	0
Bankenabwicklungsmechanismus mit Mängeln	1
A Cultural History of Food in the Medieval Age	1
claudia brasse	0
svetlana lebedeva	0
neruda	0
Protein Purification: Principles, High Resolution Methods, and Applications	1

## B. Häufigkeiten der Merkmale bei den intellektuell klassifizierten Suchanfragen

Merkmale	Known-Item-Suchen	% der Known-Item-Suchen	Unsichere Suchen	% der unsicheren Suchen	Sonstige Suchen	% der sonstigen Suchen
A	12	2,29	3	3,90	155	27,00
AD	4	0,76	0	0,00	0	0,00
ADK	4	0,76	0	0,00	0	0,00
ADT	1	0,19	0	0,00	0	0,00
AK	78	14,91	5	6,49	2	0,35
AKB	1	0,19	0	0,00	0	0,00
AN	25	4,78	2	2,60	1	0,17
AO	1	0,19	0	0,00	0	0,00
AT	39	7,46	1	1,30	0	0,00
Ax	1	0,19	0	0,00	0	0,00
BDK	2	0,38	0	0,00	0	0,00
BK	4	0,76	1	1,30	2	0,35
BN	1	0,19	0	0,00	0	0,00
BT	4	0,76	0	0,00	0	0,00
CP	14	2,68	0	0,00	1	0,17
DK	1	0,19	0	0,00	0	0,00
DN	0	0,00	0	0,00	1	0,17
DT	9	1,72	0	0,00	0	0,00
I	17	3,25	0	0,00	1	0,17
K	44	8,41	27	35,06	292	50,87
KO	1	0,19	0	0,00	0	0,00
KV	1	0,19	0	0,00	0	0,00
Kx	3	0,57	1	1,30	8	1,39
N	36	6,88	10	12,99	64	11,15
Nx	1	0,19	0	0,00	0	0,00
OV	0	0,00	0	0,00	1	0,17
P	4	0,76	1	1,30	6	1,05
S	5	0,96	0	0,00	0	0,00
T	204	39,01	24	31,17	36	6,27
TV	1	0,19	0	0,00	1	0,17
Tx	5	0,96	0	0,00	1	0,17
V	0	0,00	1	1,30	1	0,17
x	0	0,00	1	1,30	1	0,17
<b>Summe</b>	<b>523</b>		<b>77</b>		<b>574</b>	

### C. Success @n für Known-Item-Suchanfragen nach bestimmten Medientypen

Rang des Known-Items	AK	AK %	AN	AN %	AT	AT %	T	T%	K	K%	N	N%	(B oder D) + K	(B oder D) + K%	Signalwort	Signalwort%
1	42	91,30	16	94,12	24	92,31	88	70,97	18	58,06	10	40,00	7	58,33	22	52,38
2	1	2,17	1	5,88	0	0,00	9	7,26	3	9,68	3	12,00	0	0,00	6	14,29
3	1	2,17	0		0	0,00	5	4,03	2	6,45	2	8,00	1	8,33	1	2,38
	0	0,00	0		1	3,85	4	3,23	2	6,45	2	8,00	1	8,33	0	0,00
5	0	0,00	0		0	0,00	0	0,00	4	12,90	0	0,00	1	8,33	1	2,38
6	0	0,00	0		0	0,00	3	2,42	1	3,23	1	4,00	0	0,00	2	4,76
7	0	0,00	0		0	0,00	2	1,61	1	3,23	0	0,00	0	0,00	1	2,38
8	1	2,17	0		1	3,85	1	0,81	0		0	0,00	0	0,00	0	0,00
9	0	0,00	0		0		1	0,81	0		0	0,00	0	0,00	1	2,38
10	1	2,17	0		0		1	0,81	0		0	0,00	0	0,00	0	0,00
11-20	0	0,00	0		0		2	1,61	0		1	4,00	1	8,33	2	4,76
>20	0	0,00	0		0		8	6,45	0		6	24,00	1	8,33	6	14,29

## D. Auszug aus den Suchanfragen, die mit dem Algorithmus klassifiziert wurden

Aufgrund des Umfangs der gesammelten Daten werden die vollständigen Daten auf einem externen Datenträger (CD) dieser Arbeit beigelegt.

Suchanfrage	Autor (C)	Vorname (D)	Funktionswort (E)	Signalwort (F)	Zahl (G)	Wörteranzahl (H)	Suchanfragentyp (Algorithmus <sup>24</sup> )	Suchanfragentyp (manuell)	Rang des Known-Items (manuell)
varaition plagiat china	0	0	0	0	0	3	0	1	999
geriatric nursing	0	0	0	0	0	2	0	2	1
Klimaveränderung	0	0	0	0	0	1	0	0	
test	1	0	0	0	0	1	0	0	
Zeitschrift Das Aquarium	0	0	1	1	0	3	1	1	999
medienmanagement gläser	1	0	0	0	0	2	1	1	1
Herbert Weisskamp	1	1	0	0	0	2	0	0	
roberto bolano infernalismo	0	1	0	0	0	3	0	2	1111
Die Jugendkultur der HipHop	0	0	1	0	0	4	1	1	999
Verpackungsdesign und Preiskommunikation	0	0	0	0	0	3	0	1	1
schlaich	1	0	0	0	0	1	0	0	
3896692151	0	0	0	0	1	1	1	1	1
Transsexuellengesetz	0	0	0	0	0	1	0	0	
Grundlagen des Controlling lorson	1	0	1	0	0	4	1	1	1
clothing technology	0	0	0	0	0	2	0	0	

<sup>24</sup> Funktion der Excel-Formel für die zweite Zeile: =WENN(ODER(F2=1;UND(H2>1;(ODER(G2=1;E2=1;(ODER(UND(C2=1;ODER(D2=0;UND(D2=1;H2>2)))))))));1;0)

### **Eidesstattliche Versicherung**

Ich versichere, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangabe kenntlich gemacht.

Ort, Datum Unterschrift