Masterarbeit

Prototypische Entwicklung einer Software für
die Erfassung und Analyse explorativer Suchen
in Verbindung mit Tests zur Retrievaleffektivität

vorgelegt von

Sebastian Sünkler

Studiengang Informationswissenschaft und -management

Hochschule für Angewandte Wissenschaften
Fakultät Design, Medien & Information
Department Information
Finkenau 35
22081 Hamburg

Verfasser: Sebastian Sünkler
Abgabedatum: 29.02.2012

1. Prüfer: Prof. Dr. Dirk Lewandowski
2. Prüfer: Prof. Dr. Ulrike Spree

Abstract. Gegenstand dieser Arbeit ist die Entwicklung eines funktionalen Prototyps einer Webanwendung für die Verknüpfung der Evaluierung von explorativen Suchen in Verbindung mit der Durchführung klassisches Retrievaltests. Als Grundlage für die Programmierung des Prototyps werden benutzerorientierte und systemorientierte Evalulierungsmethoden für Suchmaschinen analysiert und in einem theoretischen Modell zur Untersuchung von Informationssysteme und Suchmaschinen kombiniert.

Bei der Gestaltung des Modells und des Prototyps wird gezeigt, wie sich aufgezeichnete Aktionsdaten praktisch für die Suchmaschinenevaluierung verwenden lassen, um auf der einen Seite eine Datengrundlage für Retrievaltests zu gewinnen und andererseits, um für die Auswertung von Relevanzbewertungen auch das implizierte Feedback durch Handlungen der Anwender zu berücksichtigen.

Retrievaltests sind das gängige und erprobte Mittel zur Messung der Retrievaleffektiviät von Informationssystemen und Suchmaschinen, verzichten aber auf eine Berücksichtigung des tatsächlichen Nutzerverhaltens. Eine Methode für die Erfassung der Interaktionen von Suchmaschinennutzern sind protokollbasierte Tests, mit denen sich Logdateien über Benutzer einer Anwendung generieren lassen. Die im Rahmen der Arbeit umgesetzte Software bietet einen Ansatz, Retrievaltests auf Basis protokollierter Nutzerdaten in Verbindung mit kontrollierten Suchaufgaben, durchzuführen.

Das Ergebnis dieser Arbeit ist ein fertiger funktionaler Prototyp, der in seinem Umfang bereits innerhalb von Suchmaschinenstudien nutzbar ist.

Schlagwörter. Information-Retrieval-Evaluierung, explorative Suche, Interactive Information Retrieval, Nutzerverhalten, Retrievaleffektivität, Retrievaltests, Suchmaschine, Suchsessions, Suchmaschinenvergleich, Websuchmaschine

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abb.	Abbildung
CLEF	Cross Language Evaluation Forum
etc.	et cetera (und so weiter)
et al.	und andere
ff.	folgende
IR	Information Retrieval
IIR	Interactive Information Retrieval
Kap.	Kapitel
RAT	Relevance Assessment Tool
S.	Seite
SERP	Search Engine Result Page
TREC	Text Retrieval Conference
u. a.	unter anderem
usw.	und so weiter
vgl.	vergleiche
z. B.	zum Beispiel

1. Einleitung

Suchmaschinen sind in der heutigen Zeit eine der wichtigsten Quellen für die Recherche und Beschaffung von Informationen. Die Suche im Internet ist, hinter dem Verfassen und Versenden von E-Mails, die meistgenutzte Anwendung im Internet (vgl. VAN EIMEREN / FREES 2011, S. 340). Täglich werden alleine in Deutschland mehrere Milliarden Suchanfragen von Nutzern an Websuchmaschinen abgeschickt (vgl. COMSCORE 2010), um Informationen zu den vielfältigsten Themen zu recherchieren. Suchmaschinen werden dabei sowohl von Laien als auch von Informationsprofis zur Informationsbeschaffung eingesetzt (vgl. SCHMIDT-MÄNZ 2007). Vor diesem Hintergrund stellt sich die Frage nach der Qualität der genutzten Suchdienste und den jeweils angebotenen Suchergebnissen, aber auch die Frage nach dem Nutzungsverhalten der Suchmaschinenbenutzer (vgl. LEWANDOWSKI 2011b, S. 203 – 204 & KUMAR ET AL. 2005).

Dabei ist auf der einen Seite die Relevanz und Güte der ausgelieferten Suchmaschinentreffer von hoher Bedeutung, da eine Suchmaschine ohne die Fähigkeit relevante Treffer auszugeben, nur wenig Zuspruch finden wird. Auf der anderen Seite ist es aber auch notwendig, Wissen über die Suchmaschinennutzer und ihren Interaktionen zu erfassen. Diese Daten liefern wichtige Hinweise für die Gestaltung des Suchdienstes, den angebotenen Suchfunktionen und der Usability, Ohne Wissen über den Nutzer und einer ausschließlichen Fokussierung auf die Suchergebnisse, lässt sich ein interaktiver Suchdienst nur wenig gebrauchstauglich designen (vgl. LEWANDOWSKI 2011b, S. 203 – 204 & LEWANDOWSKI / HÖCHSTÖTTER 2007).

Bei der Frage nach der Qualität von Suchergebnissen werden in Studien entweder Treffer von allgemeinen Suchmaschinen mit spezialisierten Datenbanken verglichen (vgl. LEWANDOWSKI 2011b, 203 - 204) oder Suchergebnisse gleichartiger Suchmaschinen gegenübergestellt (z. B. u. a. in BAR-ILAN / LEVENE 2011, LEIGHTON / SRIVASTAVA 1999, VÉRONIS 2006, GRIESBAUM 2004, MACHILL U. A. 2003). Oft wird dabei auch untersucht, ob die Dominanz des Marktführers Google durch die Güte der Suchergebnisse gerechtfertigt ist (vgl. BAILEY ET AL. 2007). Google ist mit Abstand die meistgenutzte Websuchmaschine vor allem in westlichen Ländern (vgl. LUNA-PARK 2011A, LUNA-PARK 2011B, NET APPLICATIONS 2012 & WEBHITS 2011). Dadurch stellt sich auch häufig die Frage zu der Neutralität der Ergebnisse von Google, welches in der Literatur als Suchmaschinen-Bias bezeichnet wird (vgl. WEBER 2011).

Insgesamt haben Untersuchungen zur Güte und Retrievaleffektivität von Informationssystemen eine lange Tradition in der Geschichte des Information Retrieval , die zur Herausbildung verschiedenen Standards bei der Evaluierung von Suchtreffern geführt hat (vgl. ROBERTSON 2008). Die Ergebnisse dieser Tests werden dabei unter anderem auch genutzt, um Indexierungslgorithmen und Suchalgorithmen zu vergleichen. Die Messung der Retrievaleffektivität ist der Versuch einer Messung der objektiven Qualität der Suchergebnisse eines Suchdienstes bei häufiger Ausblendung tatsächlicher Nutzungsszenarien (vgl. GÖDERT ET AL. 2012, S .329ff. & LEWANDOWSKI 2011b, S. 204).

Mit Tests zur Güte der Suchergebnisse eines Dienstes lassen sich alleine keine Aussagen über die Gesamtqualität treffen, da diese nur einen Aspekt des Systems betrachten. Wie bereits erwähnt sind Informationen über die Interaktionen des Nutzers mit der Suchmaschine sehr relevant. Ein Nutzer führt Recherchen oft in sogenannten Suchsessions durch. Er reformuliert Suchanfragen, wenn er durch die erste Anfrage nicht sofort die gewünschten Informationen findet (vgl. JÄRVELIN 2009, S. 291 – 292). Er nutzt eventuell mehrere Suchdienste, um an die Informationen zu kommen. Diese Herangehensweise des Nutzers zur Lösung eines Informationsproblems wird als explorative Suche bezeichnet (vgl. MARCHIONINI 2006) und unterscheidet sich von dem einfachen Anfrage-Ergebnise-Paradigma, das oft als Annahme bei der Durchführung traditioneller Retrievaltests dient. (vgl. JÄRVELIN 2009, S. 290ff. & LEWANDOWSKI 2011b, S. 205)

Die beiden genannten Faktoren sind sehr wichtig bei der Bestimmung der Qualität von Suchmaschinen. Darüber hinaus gibt es aber auch noch weitere Aspekte, die zur Qualitätsmessung herangezogen werden müssen. Lewandowski und Höchstötter benennen insgesamt vier Bereiche, die dabei Helfen ein Gesamtbild über die Güte einer Suchmaschine zu erhalten (vgl. LEWANDOWSKI / HÖCHSTÖTTER 2007):

1. Qualität des Index (Größe und Vollständigkeit des Suchmaschinenindexes, Aktualität des Indexes, länderspezifische Unterschiede)
2. Qualität der Suchresultate (Relevanz der Ergebnisse, Einzigartigkeit der Ergebnisse)
3. Qualität der Suchfunktionen (Umfang der angebotenen Funktionen; Funktionsfähigkeit der angebotenen Funktionen)
4. Nutzerfreundlichkeit von Suchmaschinen (Usability der Suchdienste, Umgang des Suchenden mit der Suchmaschine, Design des Interfaces, Akzeptanz der Suchfunktionen und Operatoren, Verarbeitung von Anfrage, Benutzerführung)

Das Modell von Lewandowski und Höchstötter zeigt die Komplexität bei der Qualitätsbestimmung von Suchmaschinen.

Für jeden Bereich müssen vielfältige Methoden und Kennzahlen genutzt werden, um aussagekräftige Ergebnisse zu erhalten. In der Praxis lassen sich die Aspekte auch nicht immer trennscharf untersuchen. Beispielsweise hängt die Qualität der Suchergebnisse immer von der Qualität des Indexes ab und der Umfang der angebotenen Funktionen sowie die Funktionsfähigkeit fallen auch in den Bereich der Usability der Suchdienste.

Ferner stellt sich insgesamt die Frage nach dem jeweiligen Aufwand, der betrieben werden muss, um Untersuchungen und Studien zur Bestimmung der Qualität von Suchmaschinen durchzuführen (vgl. dazu Kapitel 2.2. und Kapitel 2.3 zu den Methoden der Information Retrieval-Evaluierung). Auch ist die Aussagekraft der Ergebnisse solcher Studien nicht immer eindeutig, da vielfältige Faktoren Einfluss auf die durchgeführten Studien haben.

Im Laufe der Geschichte zur Evaluierung von Information Retrieval-Systemen haben sich verschiedene Verfahren, Methoden und Kennzahlen herausgebildet, um die oben genannten Aspekte zu untersuchen (vgl. GÖDERT ET AL. 2012, S 329ff., GRIESBAUM 2000, S. 27ff. & HAWKING ET AL. 2001) . In der Tradition des Information Retrieval Evaluation wird dabei zwischen systemorientierten und benutzerorientierten Verfahren unterschieden, die innerhalb dieser Arbeit vorgestellt und diskutiert werden. Innerhalb der letzten Jahre hat sich dabei der Begriff Interactive Information Retrieval herausgebildet, der versucht die verschiedenen Verfahren zusammenzuführen, da sich Informationssysteme nur bedingt von der System- oder Nutzerseite untersuchen lassen (vgl. BORLUND 2003 & KELLY 2009, S. 9ff.).

Ein weitere Frage ergibt sich dabei auch über die technischen Hilfsmittel, die in der Praxis eingesetzt werden, um Daten zu erheben und auszuwerten. Viele Studien zur Retrievaleffektiviät oder zur Nutzung der Suchmaschinen geben keine oder ungenaue Hinweise zu den eingesetzten Tools zur Gewinnung und Analyse der erhobenen Daten. Eine Standard-Anwendung, die zur Durchführung von Suchmaschinenstudien eingesetzt wird, existiert soweit es die Untersuchungen zeigen, nicht. Es stehen aber beispielsweise Anwendungen zur Datenerhebung der Nutzerinteraktionen in Bezug auf Suchsessions zu Verfügung sowie ein Relevance Assessment Tool, das zur Gestaltung und Durchführung von Retrievaltests eingesetzt werden kann. Bisher gibt es aber keine frei nutzbaren Tools, die beide Testarten zusammenführen (vgl. dazu Kapitel 5).

1.1. Zweck und Abgrenzung der Arbeit

Im Rahmen dieser Arbeit soll zunächst aufgezeigt werden, welche Standarverfahren sich bei der Evaluierung von Suchmaschinen etabliert haben und welche Grenzen und Probleme bei diesen Ansätzen bestehen. Dabei werden die Aspekte der Qualität von Suchmaschinentreffer und Interaktionen des Nutzers näher betrachtet und innerhalb eines Modells zur Kombination systemorientierter und benutzerorientierter Tests zusammengeführt, weil eine isolierte Betrachtung der Retrievaleffektivität in einigen Fragestellungen nur noch bedingt ausreicht, um die Qualität einer Suchmaschine zu messen. Auch wenn andere genannte Bereiche zur Qualitätsbestimmung (z. B. zur Usability einer Suchmaschine) durch diese Zusammenführung nicht abgedeckt werden können, ist eine Verbindung der genannten Aspekte eine sinnvolle Ergänzung traditioneller Retrievaltests durch das Einbeziehen des realen Suchverhaltens eines Nutzers miteinbezogen. Dem Nutzer können beispielsweise seine tatsächlich gesichteten Suchergebnisse aus seiner Suchsession zur Bewertung vorgelegt werden. Damit kann zusätzlich zu den erfassten Daten der Verweildauer und Zeitpunkts des Abrufs, noch konkret eine Relevanzbewertung des Suchenden zum Treffer erhoben werden. In dem Modell wird dabei der Ansatz verfolgt die Daten aus geloggten Nutzerinteraktionen und mit der Durchführung klassischer Retrieval-Tests zu verbinden.

Für die Erarbeitung des hier vorgestellten theoretischen Modells werden auf die Erkenntnisse, Frameworks und Modelle aus dem Infomation Retrieval und dem sogenannten Interactive Information Retrieval zurückgegriffen. Innerhalb der Forschungen zum Interaction Information Retrieval wurden bereits Versuche unternommen Nutzerinteraktionen mit systemorientierten Evaluierungsmethoden zu verbinden, auch mit Hilfe neuer Messindikatoren, die die Zufriedenheit der Nutzer mit den Suchergebnissen mehr in den Vordergrund stellen (vgl. KELLY 2009, S. 35ff. & ROBINS 2000, S. 57 – 58). Es werden die Vor- und Nachteile der möglichen Untersuchungsmethoden klassischer Retrievaltests, Nutzerinteraktionen im Information Retrieval sowie bisher genutzte Kennzahlen diskutiert.

Weiter wird in dieser Arbeit ein funktionaler Prototyp entwickelt. Der Prototyp basiert einerseits auf dem oben genannten theoretischen Modell und andererseits auf bereits existierenden technischen Hilfsmitteln. Ein Tool ist dabei der Search Logger, der zur Evaluierung sessionbasierter Suchen eingesetzt wird[1]. Das andere Tool ist das Relevance Assessment Tool, dass zur Durchführung von Retrievaltests genutzt werden kann[2]. Beide Anwendungen sind webbasiert und lassen sich durch ihre Beschaffenheit kombinieren.

1 Aktuelle Entwicklungen und Informationen zum Search Logger lassen sich unter http://searchtaskrepo.blogspot.com/ abrufen.

Das Ziel ist am Ende eine fließende Integration des Seach Loggers in das Relevance Assessment Tools, um Möglichkeiten zu bieten Projekte anzulegen und zu verwalten, die sowohl die Untersuchung explorativer Suchprozesse als auch Tests zur Retrievaleffektivität miteinander verbinden.

1.2 Aufbau der Arbeit

Die folgende Arbeit setzt sich insgesamt aus zwei Teilen zusammen, die in verschiedenen Kapiteln gegliedert sind. Im ersten Teil wird zunächst die Forschung im Bereich der IR-Evaluation und der Evaluierung von Suchmaschinen näher erläutert. Es werden Methoden, Kennzahlen, Vor- und Nachteile aus der langen Tradition der Untersuchungen zu Information Retrieval Systemen diskutiert und Besonderheiten bei der Evaluation von Websuchmaschinen herausgestellt. Die Erkenntnisse aus diesen Forschungsbereichen bilden anschließend die Grundlage für das anschließend vorgestellte Modell zur Verbindung klassischer Retrievaltests mit Nutzerdaten zu explorativen Suchen.

Die Entwicklung des Modells soll dabei durch folgende Fragen geleitet werden:

1. Wie können Daten der Nutzerbeobachtung sinnvoll für die Gestaltung von Retrievaltests genutzt werden?
2. Welche Grenzen werden gesetzt, wenn zwei sehr unterschiedliche Ansätze in einem Modell zusammengefasst werden?
3. Wie kann eine Vergleichbarkeit von Suchsystemen geschaffen werden, wenn der tatsächliche individuelle Benutzer Ausgangspunkt der Evaluierung ist?
4. Welche Faktoren aus der traditionellen Information-Retrieval-Evaluation müssen bei dieser Form von Tests berücksichtigt werden?
5. Wie lässt sich der Aufwand bei der Gestaltung und Datenerhebung bei solchen Untersuchungen minimieren?

Der zweite Teil der Arbeit thematisiert die technische Umsetzung des Modells innerhalb eines funktionalen Prototyps, der auf zwei bereits vorhandenen Webanwendungen basiert.

---

2 Das Frontend desRelevance Assessment Tools kann unter http://search-studies.org/rat/ aufgerufen werden. Der Access-Code für ein Testszenario ist „RAT-Test“ ohne Anführungszeichen.

2. Evaluierung von Information-Retrieval-Systemen

Information-Retrieval-Systeme sind computergestützte Anwendungen, die dabei Helfen Informationen wieder aufzufinden, die für die Bewältigung eines Problems benötigt werden. Sie bieten Nutzern über angebotene Suchmechanismen Zugang zu Dokumenten, zu den Informationen in den Dokumenten, zu Metadaten über Dokumenten die in spezialisierten Fachdatenbanken gespeichert oder frei im World Wide Web verfügbar sind. (vgl. GRIESBAUM 2000, S. 12 ff. & MANDL 2008, S. 27 – 28).

Allgemein ist Information Retrieval die Schlüsseltechnologie im Wissensmanagement, da sie den Zugang zu großen Mengen von unstrukturierten und strukturierten Daten ermöglicht. Im Information Retrieval geht es um die Speicherung und Repräsentation der erfassten Wissens und das Wiederauffinden des gespeicherten Wissens relevant zu den jeweiligen Zielen oder Aufgaben eines Nutzers, für die er die angeforderten Informationen benötigt (vgl. MANDL 2008, S. 27 – 28 & AKHIGBE ET AL. 2011a, S. 35) . IR-Systeme antworten typischerweise auf Suchanfragen von Nutzern, die aus wenigen Wörtern der natürlichen Sprache bestehen. Es folgt ein Abgleich mit den Suchanfrage und dem Dokumentenrepräsentation, die durch Indexierung der Dokumente gebildet wird. Als Ergebnis erhält der Nutzer eine Aufbereitung der Dokumente und Inhalte innerhalb einer vom System abhängigen Ergebnispräsentation.Der Nutzer hat durch diese Präsentation der Ergebnisse nun die Möglichkeit, die für sein Problem benötigten Dokumente und Inhalte zu sichten und auszuwerten (vgl. MANNING ET AL. 2009, S. 1ff. & WEBBER 2010, S. 4ff.). Typische Anwendungen, die auf Methoden aus dem Information Retrieval zurückgreifen sind Websuchmaschinen, kostenpflichtige Fachdatenbanken, Webshops aber auch Suchdienste für gespeicherte Daten auf einem Computer (vgl. AKHIGBE ET AL. 2011b).

Zusammenfassend betrachtet besteht ein IR-System aus verschiedenen Regeln und Prozeduren um einen Teil oder alle der unten aufgeführten Operationen auszuführen: (vgl. ROBERTSON 1981, S. 9):

- Indexing (Erstellen einer inhaltlichen und/oder formalen Dokumentrepräsentation)
- Suchanfrage(-formulierung) (Repräsentation des Informationsbedürfnisses)
- Suche (Abgleich zwischen Suchanfrage und Dokumentrepräsentation)
- Feedback (iterative Wiederholung und/oder Modifikation der oben genannten Prozesse in Abhängigkeit von der Einschätzung der vorhergegangen Prozessergebnisse)
- Erstellung einer Indexierungssprache (bzw. Aufstellung von Dokumentrepräsentationsregeln)

Das Ziel von guten Information-Retrieval-Systemen und Suchmaschinen ist dem Nutzer die relevanten Dokumente und damit im weiteren Sinne, die Informationen zu liefern, die dieser für sein jeweiliges Informationsbedürfnis benötigt und durch seine Suchanfrage ausdrückt (vgl. AKHIGBE ET AL. 2011b, S. 6). Aussagen darüber inwieweit ein IR-System diesem Ziel gerecht wird, lässt sich nur mit Hilfe von Evaluationen bestimmen. Bei den Untersuchungen der Systeme können dabei die Aspekte der Retrievaleffektivität und Retrievaleffizienz betrachtet werden. Retrievaleffizienz beschreibt Faktoren wie die Antwortzeit, die Lernzeit, Kosten und auch andere wirtschaftliche Faktoren die im Verhältnis zur Ergebnismenge betrachtet werden. Die Retrievaleffektivität hingegen beschreibt wie gut gesuchte Informationen geliefert oder referenziert werden. Die Effektivität gibt wieder inwieweit ein System dazu in der Lage ist, passende Antworten, Informationen und Dokumente auf die Anfragen der Nutzer zu liefern, wohingegen die Effizienz eines Systems zeigt, welcher Aufwand dafür betrieben von Nutzern betrieben werden muss. (vgl. GÖDERT ET AL. 2012 S. 329 – 330 & WORMS-HACKER 2004, S. 227)

Im Folgenden wird zunächst anhand der Funktionsweise einer Websuchmaschine gezeigt, wie ein Information-Retrieval-System aussehen kann. Darauf folgt eine Darstellung der Methoden zur Untersuchung von Informationssystemen und Suchmaschinen. Dabei wird nach systemorientierten und benutzerorientierten Ansätze in der IR-Evaluierung differenziert, die sich jeweils auf verschiedene Schwerpunkte beziehen. Während in der systemorientierten Evaluierung vor allem die Funktionsfähigkeit eines Systems in Bezug auf das Wiederauffinden von Dokumenten im Vordergrund steht, werden bei der benutzerorientieren Evaluierung die Benutzer stärker in den Fokus der Untersuchungen gerückt. (vgl. TURPIN/SCHOLER 2006, S. 12ff.)

Studien aus der Praxis, in denen sowohl Methoden aus der systemzentrierten Evaluierung und benutzerorientierten Untersuchung verbunden haben, werden in Kapitel 3 diskutiert.

2.1 Beispielaufbau eines Information-Retrieval-Systemens anhand der Typologie einer Websuchmaschine

Im Folgenden zeigt die Typologie einer Websuchmaschine, wie eine alltägliche Information-Retrieval-Anwendung gestaltet sein kann, um den Zugang zu Informationen zu ermöglichen. Eine Websuchmaschine bietet alle Komponenten, die nötig sind, damit alle oben genannten Aufgaben eines IR-Systems zu bewältigt werden können.

Eine genannte Aufgabe ist die Speicherung von Dokumenten und Inhalten und die Aufbereitung der Informationen innerhalb einer Dokumentenrepräsentation. Internetsuchmaschinen greifen auf zwei Möglichkeiten zurück, um Inhalte in ihren Index aufzunehmen und diese damit wieder auffindbar zu machen. Einerseits werden sogenannte Webcrawler eingesetzt, die das World Wide Web automatisch nach Inhalten für die Suchmaschine durchsuchen, eine Schlüsseltechnologie für Internetsuchmaschinen und andererseits können Betreiber von Webseiten auch eine manuelle Anmeldung an der Suchmaschine vornehmen. In beiden Fällen wird bei der Extraktion der Inhalte aus den Dokumenten und der Indexierung auf Methoden aus dem Information Retrieval zurückgegriffen. Es kommen unter anderem statistische und linguistische Verfahren zum Einsatz. Beispielsweise wird bei der Indexierung von Volltexten die exakte Schreibweise erfasst und die Groß- und Kleinschreibung normalisiert. Die so ermittelten Dokumenteninhalte werden bei der Indexierung noch durch weitere spezifische Faktoren ergänzt, die unter anderem für das Ranking der Suchergebnisse eingesetzt werden. Diese Faktoren spielen eine große bei der Aufgabe, dem Nutzer anhand seiner Ziele Absichten, die er durch die Suchanfrage ausdrückt, relevante Suchergebnisse auszuliefern (vgl. (vgl. GRIESBAUM 2009 S. 28ff.). In Abb. 1 sind alle elementaren Komponenten einer Suchmaschine dargestellt.

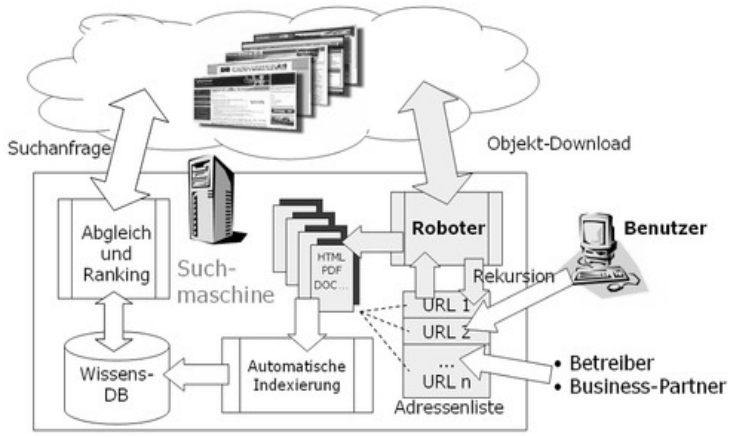


Abbildung 1: Aufbau einer Websuchmaschine (Quelle: GRIESBAUM ET AL. 2009, S. 28)

Abhängig von der Suchmaschine werden die gefunden Treffer zu der Suchanfrage in einer bestimmten Art und Weise bereitgestellt und ausgegeben. Es findet ein Ranking der Inhalte und Informationen statt, die das Informationsbedürfnis des Benutzers befriedigen sollen.

Neben den genannten Komponenten einer Suchmaschine, sind aber auch Benutzerführung, Recherchesprache und angebotene Suchfunktionen relevant, damit der Nutzer die Dokumente und Inhalte findet, die er für seine Aufgabe und Ziele benötigt.

2.2 Systemorientierte Evaluation

Die systemorientierte Bewertung von IR-Systemen hat eine lange Geschichte und Tradition und bildet auch den Ausgangspunkt der meisten Evaluierungsinitiativen, von denen in Kapitel 2.5 einige exemplarisch zusammenfassend vorgestellt werden.

Bereits 1953 wurden mit de sogenannten ASTIA und Cranfield Tests die ersten systemorientierten Evaluierungen in der Forschung zum Information Retrieval durchgeführt . Die in den Cranfield Tests eingesetzten Methoden für die Untersuchung verschiedener Systeme bilden auch heute, wenn auch angepasst an moderne Verhältnisse und Standards, die Grundlage in der systemzentrierten Evaluation von Suchdiensten. (vgl. GÖDERT ET AL. 2012, S. 333ff. & ROBERTSON 2008)

Eine klassische Methode der systemorientierten Evaluierung von IR-Systememn sind sogenannte Retrievaltests. Im folgenden wird diese Testart näher mit methodischen Fragen, Planung und Durchführung der Tests sowie entwickelten Maßzahlen für die Auswertung der Tests näher vorgestellt. Dabei werden vor allem auch die Vor- und Nachteile dieser Methode allgemein und in Bezug auf die Evaluierung von Websuchmaschinen hervorgehoben.

2.2.1 Retrievaltests

In der systemorientierten Evaluierung von Suchsystemen wird mit Retrievaltests gearbeitet, die in der Regel einem Standardaufbau folgen, bei dem eine bestimmte Menge an Suchanfragen an verschiedene Suchmaschinen geschickt wird. Die zurückgegebenen Ergebnisse werden dann durchmischt und anonymisiert und anschließend Juroren zur Bewertung vorgelegt. Nach der Bewertung werden die Suchtreffer wieder den Systemen und den ursprünglichen Trefferplätzen, auf denen sie gefunden wurden, zugeordnet (vgl. HAWKING ET AL. 2001 & VOORHEES / HARMAN 2006). Die Auswertung erfolgt mit Hilfe verschiedener Effektivitätsmaße, die Aussagen über die Qualität des Systems bzw. der Trefferdokumente ermöglichen.

Bei diesen Tests wird meistens nur eine bestimmte Anzahl von Treffern bewertet, da besonders bei Web-Suchmaschinen die Anzahl der gefundenen Suchergebnisse von einem Juror kaum prüfbar wäre (vgl. LEWANDOWSKI 2011b, S. 206).

Deshalb wird in der Praxis auf die Poolingmethode zurückgegriffen, um die Dokumente der zu prüfenden Systeme zu gewinnen. Mit dieser Methode soll eine möglichst präzise Annäherung an die Gesamtzahl alles relevanten Dokumente zu eines Systems zu einer Suchanfrage zu erreicht werden.

Es wird eine feste Anzahl der Dokumente bewertet, die durch die Systeme zurückgegeben werden. Alle nicht gefundenen Dokumente werden als irrelevant eingestuft (vgl. BUCKLEY ET AL. 2007, S. 498ff., WORMS-HACKER 2004, S. 229).

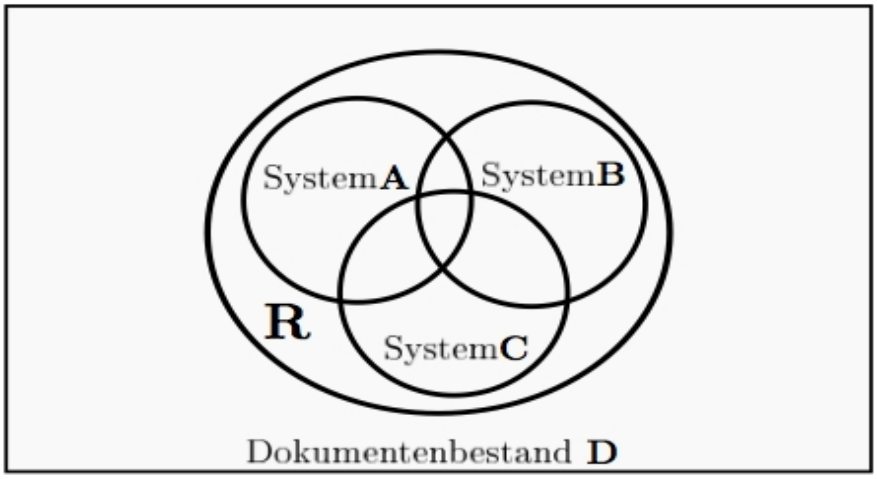


Abbildung 2: Prinzip der Pooling-Methode (Quelle: LAMM 2008, S. 8)

Abbildung 2 zeigt das Prinzip dieser Methode. R stellt hier die gefundenen relevanten Dokumente zu einer Suchanfrage dar und wird aus dem gesamten Dokumentenbestand D gewonnen. Der Schätzwert R wird durch die Gesamtzahl der zurückgelieferten relevanten Dokumente aus den Systemen A, B und C gebildet. In Bezug auf Websuchmaschinen würde das bedeuten, dass D = Gesamtzahl aller im Web vorhandenen Dokumente ist. Die einzelnen Systeme repräsentieren dabei jeweils eine andere Suchmaschine. Als Gesamtzahl der relevanten Treffer werden also durch die ausgegebenen relevanten Ergebnisse aller zu untersuchenden Systeme gebildet.

Juroren bei Retrieval-Tests werden je nach Zweck und einem vorher festgelegten Nutzerprofil ausgesucht. So geht TREC (vgl. Kapitel 2.5) von der Prämisse aus, dass es sich bei dem Nutzer des Systems um einen sogenannten „dedicated searcher“ handelt. Einem Benutzer der sich vom Laien abgrenzt, der sich einen hohen Recall und eine hohe Precision wünscht und dafür auch dazu bereit ist eine Vielzahl von Dokumenten durchzusehen (vgl. LEWANDOWSKI 2011b, S. 206 & MANDL 2008, S. 28). Dementsprechend werden bei Retrieval-Tests durch TREC unabhängige Experten eingesetzt, die die Dokumente auf ihre Relevanz hin überprüfen, damit eine einheitliche und objektive Bewertung stattfindet.

Das Profil eines Websuchmaschinennutzers unterscheidet sich von der Prämisse durch TREC. So werden Suchmaschinen von allen Nutzergruppen eingesetzt, sowohl der Informationslaie als auch der Profi nutzt Websuchmaschinen, wobei Laiennutzung der Standardfall ist. Dies zeigt sich durch die allgemein formulierten und sehr kurzen Suchanfragen, die Suchmaschinennutzer eingeben.

Daneben sieht sich ein Suchmaschinennutzer in der Regel auch nur wenige Dokumente während einer Suchsession an. (vgl. SCHMIDT-MÄNZ 2007, HÖCHSTÖTTER / KOCH 2009, S. 54ff., KUMAR ET AL. 2005, S. 5ff. & JANSEN / SPINK 2006, S. 254ff.)

In den klassischen Retrievaltests werden daher die zu untersuchenden Treffer bis zu einem bestimmten Cuttoff-Wert zu einer Suchanfrage durch den Nutzer bewertet. Der Cutoff-Wert steht für eine vorher festlegte Anzahl an maixmalen Treffern eines Sytems zu einer Anfrage (vgl. LEWANDOWSKI 2011b, S. 210 – 211 & MANNING ET AL. 2009, S. 166). Also beispielsweise die ersten 10 Treffer jeder Suchmaschine. Retrievaltests bieten dieses Nutzerverhalten nicht ab, da in der Regel alle Dokumente die mit Hilfe der Pooling-Methode zusammengestellt wurden, durch Juroren bewertet werden, Benutzer aber in ihren Suchsessions nicht zwingend alle Dokumente einer Search Engine Result Page (SERP) anklicken (vgl. BUSCHER ET AL. 2004, S.,46 – 47, JANSEN / SPINK 2006, S. 257 – 258, JOACHIMS ET AL. 2005 & SCHMIDT-MÄNZ 2007)

Werden Retrievaltests in Rahmen von Evaluierungsinitiativen durchgeführt. Wird in der Regel eine Testkollektion von Dokumenten bereitgestellt, um eine optimale Vergleichbarkeit der IR-Systeme zu gewährleisten (vgl. KELLY 2009, S. 196 – 197, TAGUE-SUTCLIFFE 1992, S. 469 – 470). Alle zu überprüfenden Systeme erhalten durch die Testkollektion den gleichen Index und damit die gleiche Ausgangslage. Die Ausgangslage aller Systeme ist damit gleich. Dieses Verfahren lässt sich aber nicht auf größere Datenbestände beziehen und kann auch nur angewendet werden, wenn Systeme freiwillig an den Evaluierungsinitiativen teilnehmen und ihre Systeme offen legen. Websuchmaschinen nehmen in der Regel nicht an solchen Initiativen teil. Daher gibt es bei vergleichenden Retrievaltests von Websuchmaschinen nicht die gleichen Ausgangsbedingungen in Bezug auf die indexierten Dokumente. Websuchmaschinen erheben zwar einen Anspruch darauf das gesamte Web abzudecken, aber einigen Studien konnten zeigen, dass sich die Indexe von Websuchmaschinen durchaus sehr unterscheiden können. Darüber hinaus ergibt sich noch das Problem der Skalierbarkeit einer beschränkten Testkollektion auf das gesamte Web. Da die Untersuchung der Güte von Suchergebnissen der Websuchmaschinen vom großen Interesse ist, müssen die aufgezeigten Nachteile des Index-Einflusses in Kauf genommen werden. (vgl. LEWANDOWSKI 2011b, S. 206ff.)

2.2.1.1 Methodik von Retrievaltests

Der Aufbau von Retrievaltests orientiert sich meistens an die durch TAGUE-SUTCLIFFE 1992 aufgestellten Schritten. TAGUE-SUTCLIFFE 1992 bieten eine Zusammenstellung der wichtigsten methodischen Entscheidungen, die bei der Konzipierung von Retrievaltests getroffen werden sollten:

1. Testen oder nicht testen? Ein Test ist nur sinnvoll, wenn von ihm neue Erkenntnisse zu erwarten sind.
2. Welche Art von Test soll durchgeführt werden?
3. Wie sollen die Variablen operationalisiert werden?
4. Welche Informationsysteme sollen untersucht werden?
5. Wie werden die Suchanfragen und Informationsbedürfnisse erschlossen? Die Auswahl der Suchanfragen bestimmt darüber, ob in den Tests tatsächliche Informationsbedürfnisse von Nutzern Ausgangspunkt der Untersuchung sind oder ob es eine künstliche Verzerrung bei den Ergebnissen gibt.
6. Wie werden die Suchanfragen durchgeführt?
7. Wie soll die Testanordnung aussehen?
8. Wie werden die Daten im Test erhoben?
9. Wie werden die erhobenen Daten ausgewertet?
10. Wie sollen die Ergebnisse präsentiert werden?

Diese Übersicht zeigt bereits, dass eine Vielzahl von Entscheidungen bei der Konzeption von Retrievaltests zu fällen sind, die einen großen Einfluss auf die Ergebnisse haben. Jeder dieser Punkte muss gut durchdacht werden, damit optimale Testbedingungen gewährleistet werden.

Neben den von TAGUE-SUTCLIFFE 1992 aufgestellten methodischen Entscheidungen, die vor jeder Durchführung eines Retrievaltests gefällt werden müssen stellten GORDON / PATHAK 1999 und darauf aufbauend HAWKING ET AL. 2001 noch weitere Kriterien auf, die bei Retrievaltests mit Websuchmaschinen wichtig sind:

1. Bei den Untersuchungen sollen reale Informationsbedürfnisse von Nutzern abgebildet werden.
2. Werden Informationsvermittler eingebunden, muss das originäre Informationsbedürfnis sorgfältig mitgeteilt werden.
3. Es soll eine genügend große Anzahl an Suchanfragen genutzt werden.
4. Alle wichtigen Suchmaschinen sollten im Test evaluiert werden.
5. Die Untersuchung muss sorgfältig und gut aufgebaut und durchgeführt werden.

Die Übersicht der genannten Kriterien und methodischen Entscheidungen bilden eine solide Grundlage für die Durchführung von Retrievaltests. Die genannten

Voraussetzungen, die bei der Untersuchung von Websuchmaschinen berücksichtigt werden sollten, werden in den meisten neueren Studien auch erfüllt.

2.2.1.2 Aufbau von Retrievaltests

LEWANDOWSKI 2011b gibt einen vollständigen Überblick für den Aufbau eines Retrievaltests mit den zu berücksichtigenden Faktoren und Entscheidungen. Die folgende Schritte zeigen, wie die Konzeption und Durchführung eines Retrievaltests im Einzelnen aussieht:

1. Auswahl der Suchmaschinen / Information-Retrieval-Systeme: Die Auswahl ist abhängig vom Zweck der Untersuchung.
2. Anzahl der Suchanfragen: Festlegung der Anzahl von Suchanfragen, umso mehr umso besser. In der Praxis hat sich eine Anzahl von mindestens 50 Suchanfragen durchgesetzt, auch wenn es für diese Zahl keine empirischen Belege gibt.
3. Art der Suchanfragen: Die Auswahl der Art hängt vom Untersuchungsziel ab.
4. Herkunft der Suchanfragen: Suchanfragen lassen sich auf verschiedene Arten gewinnen, z. B. könnten Anfragen aus Suchmaschinen-Logfiles extrahiert werden oder von realen Nutzern mit Nennung des Suchbedürfnisses kommen.
5. Darstellung der Suchanfragen: Hier wird entschiedene wie die Suchfragen den Juroren vorgelegt werden. Dabei kann auch das Informationsbedürfnis und das Relevanzkriterium für Ergebnisse genannt werden.
6. Anzahl der Suchergebnisse zu einer Anfrage: Die Anzahl hängt von der Nutzergruppe ab, die im Test abgebildet werden soll. Für den durchschnittlichen Suchmaschinennutzer ist ein Cutoff-Wert von 10 sinnvoll, da dieser sich in der Regel nur wenige Ergebnisse ansieht.
7. Anzahl der Juroren: Die Anzahl der Juroren in einem Test wird durch die Anzahl der zu untersuchenden Suchanfragen bestimmt. Ein Juror bewertet alle Dokumente, die zu einer jeweiligen Suchanfrage von den zu untersuchenden Systemen zurückgegeben wurden.
8. Auswahl der Juroren: Die Auswahl der Juroren richtet sich nach Fragestellung und den zu Systemen, die getestet werden.
9. Anzahl der Juroren pro Suchanfrage: Meistens werden alle Treffer zu einer Anfrage nur von einer Person bewertet. Sollen verschiedenen Nutzergruppen für die Untersuchung von Systemen getestet werden, ist es notwendig, dass die Ergebnisse einer Suchanfrage jeweils einem Vertreter aus der jeweiligen Gruppe vorgelegt wird.

10. Bewertung der Ergebnisse: Für die Bewertung der Suchergebnisse durch die Juroren müssen innerhalb des Tests Skalen festgelegt werden. Häufig werden die Dokumente binär bewertet. Dabei wird abgefragt ob ein Dokument relevant oder nicht relevant ist. Es empfiehlt sich dabei zusätzlich eine 5-Punkte Skale für die Abstufung der Relevanz anzubieten, da Suchmaschinennutzer dazu tendieren Ergebnisse mit wenig Informationsgehalt trotzdem als relevant einzustufen.
11. Bewertung der Trefferbeschreibungen: Neben den Ergebnissen sollten auch die Trefferbeschreibungen erfasst und zur Beurteilung freigeben werden, da im Kontext der Websuche davon auszugehen ist, dass ein Suchergebniss ohne relevante Trefferbeschreibung in der Ergebnisliste nicht angeklickt wird.
12. Sonstige Hinweise für den Aufbau von Retrievaltests: Wie bereits erwähnt findet in der Regel eine Anonymisierung der Suchergebnisse innerhalb von Tests statt. Dieser Vorgang ist notwendig, damit Juroren nicht von der Herkunft des Treffers beeinflusst werden. So hat sich beispielsweise in einigen Studien gezeigt, dass Suchmaschinennutzer Suchergebnisse vom weltweiten Marktführer Google in der Regel besser bewerten als die Ergebnisse anderer Suchmaschine, selbst wenn diese bessere oder gleichwertige Treffer liefert. Daneben muss auch die Reihenfolge der Treffer durchmischt werden, damit Lerneffekte ausgeschlossen sind. Sonst könnte sich ein Juror an dem ursprünglichen Ranking der Suchmaschinen orientieren und dementsprechend über die Güte der gezeigten Ergebnisse urteilen. Dubletten sollten ebenfalls gefiltert werden, damit ein Ergebnis nicht mehrfach bewertet werden muss. Bei jedem Testdesign sollte auch immer der Aufwand bedacht werden, den ein Juror für die Bewertung der Ergebnisse aufbringen muss.:

    Anzahl Suchmaschinen x Anzahl der Suchanfragen
    x Anzahl der Suchergebnisse x Juroren pro Suchanfrage

Sollen beispielsweise 3 Suchmaschinen mit insgesamt 50 Suchanfragen und 10 Suchergebnisse pro Anfrage von einem Juror bewertet werden, müsste dieser insgesamt 1.500 Items bewerten.

Der dargestellte Überblick für den Standardaufbau eines Retrievaltests macht deutlich, welcher Aufwand betrieben werden muss, damit valide und reliable Ergebnisse durch Relevanztests generiert werden. Ein wichtiges Problem bei der Messung der Retrievaleffektivtät ist dieRelevanz.

2.2.1.3 Das Problem der Relevanz

Relevanz ist eines der zentralen Konzepte in der Informationswissenschaft und bildet auch wie bereits gezeigt den Ausgangspunkt für die Messung der Retrievaleffektivität von Informationssystemen und die Durchführung von Retrievaltests. Es wurde bereits mehrfach erwähnt, dass die Fähigkeit einer Suchmaschine relevante Ergebnisse zu Suchanfragen auszugeben ist ein wichtiges Qualitätsmerkmal darstellt. Allerdings herrscht in der Fachliteratur keine Einigkeit darüber, wie Relevanz definiert werden kann, auch nicht wie sich Relevanz messen lässt. FERBER 2003 kritisiert beispielsweise die Abhängigkeit der Bewertungsmaße von menschlichen Einflüssen.

MÖHR 1980 spricht vm Problem der Stellvertreterentscheidung,
da ein Juror die Suchergebnisse zu einer Suchanfrage beurteilt, ohne das tatsächliche Informationsbedürfnis hinter der Anfrage zu kennen. MIZZARO 1997 wertete 160 Veröffentlichungen zum Begriff der Relevanz aus und zeigt, dass Relevanz sich aus verschiedenen Faktoren aus zwei Gruppen zusammensetzt.

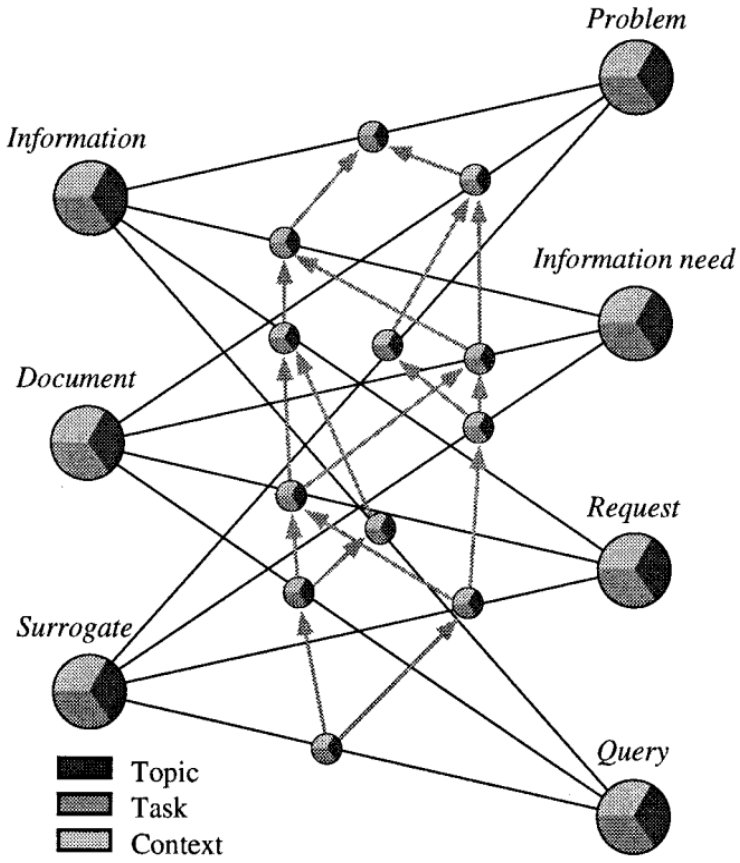


Abbildung 3: Relevanz-Modell (Quelle: MIZZARO 1997, S. 812)

Abb. 3 zeigt beide Gruppen mit den jeweiligen Elementen und macht deutlich, dass zwischen allen Elementen Beziehungen bestehen. Die Gruppe auf der linken Seite enthält dabei die Information, die bei dem Nutzer ankommt, das Ergebnisdokument zu der Anfrage sowie die Dokumenteneinheit für das Dokument. Die Gruppe auf der bildet den Nutzer ab. Der Nutzer hat ein Problem, das dieser mit Hilfe des Suchsystems lösen möchte. Dabei ergibt sich ein konkretes Informationsbedürfnis, das sich im Bewusstsein des Nutzers bildet, sich dabei aber vom konkreten Problem unterscheiden kann, wenn der Benutzer nicht dazu in der Lage ist sein Informationsproblem nicht richtig oder nicht vollständig wahrnimmt. Das Anliegen ist die Repräsentation des Informationsbedürfnisses in natürlicher Sprache und die Suchanfrage schließlich die Formulierung des Informationsbedürfnisses in der Sprache des Information-Retrieval-Systems. (vgl. MIZZARO 1997, S. 811ff.)

Das Modell zeigt bereits, dass es möglich ist von verschiedenen Arten der Relevanz zu sprechen. Ist ein Dokument bereits relevant, wenn die Suchanfrage passende Ergebnisse zu einer Anfrage liefert? Oder ist es notwendig, dass die gefundenen Ergebnisse die Informationen bereitstellen, die der Nutzer benötigt, um sein Informationsbedürfnis zu befriedigen. Diese Arten von Relevanz lassen sich als Systemrelevanz und Benutzerrelevanz bezeichnen. Wobei die Systemrelevanz aussagt inwieweit das System dazu in der Lage ist Dokumente auszugeben, die zu der Eingabe des Benutzers passen und die Benutzerrelevanz aussagt wie nützlich die gefundenen Dokumente für den Nutzer sind. Daneben wird in den Informationswissenschaften noch zwischen den Begriffen der Relevanz und Pertinenz unterschieden. Von Relevanz wird dann gesprochen, wenn ein Dokumenten dem objektiven Informationsbedarf erfüllt, also wenn zu einer Suchanfrage passende Dokumente gefunden werden. Der Begriff Pertinenz wird dagegen bei einem Vergleich der Übereinstimmung des Informationsbedürfnisses mit den Suchtreffern gebraucht (vgl. STOCK 2007, S. 68ff.).

Ein großes Problem des Begriffs der Relevanz ist, dass dieser als objektiv operationalisiertes Bewertungsmaß bei Retrievaltests eingesetzt wird, aber im Grunde genommen nicht objektivierbar ist (vgl. GRIESBAUM 2000, S. 24 – 25 & LEWANDOWSKI 2005, S. 97). Entscheidungen über die Relevanz eines Dokuments oder von Informationen werden immer durch subjektive Eigenschaften des Jurors beeinflusst. Daher kann die Relevanz von Dokumenten nur im Zusammenhang mit der Handlungsrelevanz einer Person verstanden werden, die versucht ein konkretes Informationsproblem mit der Verwendung von Suchmaschinen zu lösen. Bei der Durchführung klassischer Retrievaltests muss daher von der konkreten persönlichen Situation abstrahiert werden. Relevanzurteile sind immer abhängig von interpersonellen und intertemporalen Unterschieden, selbst wenn der personenbezogene Handlungskontext ignoriert wird.

Die Unterschiede ergeben sich durch die Voraussetzungen des jeweiligen Jurors dazu zählen u. a. Die kognitiven Fähigkeiten und der aktuelle Wissensstand (vgl. GRIESBAUM 2000, S. 24 ff. & SARODNICK / BRAU 2006, S. 45ff.)

Die genannten Probleme beim Konzept der Relevanz zeigen noch einmal die Dringlichkeit Retrievaltests sorgfältig zu konzipieren und durchzuführen. In der Aufstellung der Schritte zum Aufbau von Retrievaltests wurden bereits einige Punkte genannte, die bei dem Problem der Relevanzbewertung Unterstützung bieten sollen. So ist die Angabe des tatsächlichen Informationsbedürfnisses und eine Beschreibung zu der jeweiligen Suchanfrage hilfreich, um den Juror eine Hilfestellung bei der Bewertung zu geben. Auch die zusätzliche Bewertung von Trefferbeschreibungen ermöglicht eine genauere Relevanzbeurteilung der Treffer. Daneben gibt es auch noch einige Besonderheiten bei der Bewertung der Ergebnisse von Suchergebnissen aus Websuchmaschinen. Im Web gibt es zu nachgefragten Themen in der Regel eine Vielzahl grundsätzlich relevanter Dokumente und Juroren neigen generell auch dazu ein Dokument als relevant einzustufen. Nur eine differenzierte Beurteilung der Dokumente mit Hilfe abgestufter Skalen ermöglicht einen Vergleich von Websuchmaschinen, da bei reinen binären Bewertungen zu viele Dokumente als relevant eingestuft werden (vgl. LEWANDOWSKI 2011b, S. 211 – 212).

2.2.1.4 Die Standardmaße Recall und Precision

Im Laufe der Geschichte der Information-Retrieval-Evaluierung und vor allem im Zusammenhang mit der Durchführung von Retrievaltests, haben sich einige Messzahlen herausgebildet, die für die Bewertung der Retrievaleffektivät von Informationssystemen genutzt werden.

Dabei haben sich vor allem zwei Standardmaße durchgesetzt: die Precision und der Recall (vgl. GÖDERT ET AL. 2012, S. 329ff. & TAGUE-SUTCLIFFE 1992, S. 472 – 473) . Beide Werte werden ermittelt, indem aus der vorhandenen Dokumentenkollektion vier Teilmengen gebildet werden, die zu einander ins Verhältnis gesetzt werden. Die Standardmaße wurden mit der Annahme gebildet, dass eine vollständige und vollständig relevante Treffermenge vorliegt.

Die Abb. 4 zeigt die vier Teilmengen, die für die Berechnung von Recall und Precision benötigt werden.

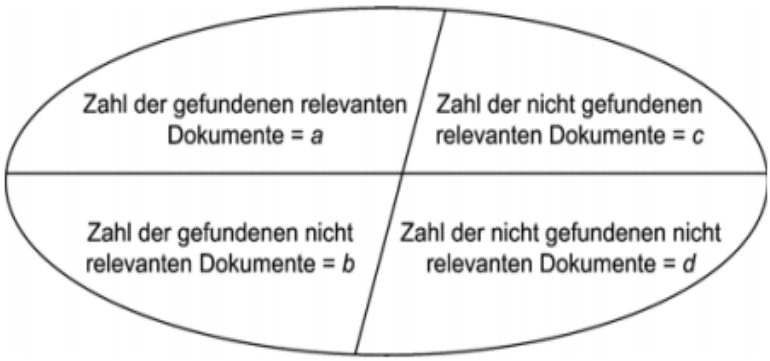


Abbildung 4: Teilmengen einer Dokumentenkollektion zur Bestimmung von Recall und Precision (Quelle: GÖDERT ET AL. 2012, S. 330)

Precision:
Die erste klassische Maßzahl ist die Precision (auch: Präzision oder Relevanzrate) ist die am häufigsten in Relevanztests genutzte Maßzahl. Sie misst die Anzahl der gefunden relevanten Treffer im Verhältnis zu der Zahl aller Dokumente in der Treffermenge.

$$p = \frac{a}{a+b}$$

Für die Berechnung werden dem Juror oder mehreren Juroren alle Ergebnisse zu einer Suchanfrage vorgelegt und danach wird der Anteil der relevanten Treffer ausgezählt. Bei großen Treffermengen, wie sie bei Websuchmaschinen zu erwarten sind, wird die Anzahl der gesamten Treffer durch einen Cutoff-Wert (z. B. Die ersten 10 Suchergebnisse) beschränkt. In modernen Untersuchungen werden häufig Ableitungen aus diesem Standardmaß für die Messung genutzt (siehe Seite 20).

Recall:
Die zweite klassische Maßzahl ist der Recall (auch: Rückruf- oder Vollständigkeitsrate). Der Recall wird durch das Verhältnis der gefundenen relevanten Treffer zu allen relevanten Treffern im Datenbestand ermittelt.

$$r = \frac{a}{a+c}$$

Der Datenbestand setzt sich dabei aus der Gesamtzahl der Dokumente innerhalb einer Datenbank oder in Bezug auf das Web aller im Internet vorhandenen relevevanten Seiten. Dieser Umstand zeigt, dass es sehr schwierig ist den Recall zu messen bzw. bei Tests mit Websuchmaschinen unmöglich ist. In diesem Kontext kann der Recall nur mit Hilfsmethoden, wie der Pooling-Methode erfolgen.

Die Werte der Precision und des Recalls können zwischen 0 und 1 annehmen. Wobei 1 der beste und 0 der schlechteste Wert ist. Ist die Precision 1 sind alle gefundenen Dokumente in der Treffermenge relevant. Ist der Wert hingegen 0, dann wurde kein einziges relevantes Dokument durch das Informationssystem zurückgegeben in der Treffermenge zurückgegeben. Beim Recall bedeutet der Maximalwert 1, dass alle relevanten Dokumente aus der Dokumentenkollektion gefunden wurden, während der Minimalwert 0 zeigt, dass kein einziges relevantes Dokument gefunden wurde. Ein ideales Information-Retrieval-System sollte demnach einen maximalen Recall bei größtmöglicher Precision aufweisen.

Bei der Messung der beiden Werte muss darauf geachtet welcher Zweck mit dem Informationssystem verfolgt wird. So ist es bei einer Patentrecherche essenziell alle relevanten Dokumente zu finden während bei einer Websuche eher eine hoher Precision erwünscht ist. Aus diesem Grund lassen sich auch Suchsysteme die verschiedene Absichten verfolgen nicht mit Hilfe der Standardmaße Precision und Recall vergleichen. Darüber hinaus kann aber auch abhängig von der momentanen Aufgabe und des Informationsproblems des Nutzer bei gleichartigen Systemen entweder mehr Wert auf die Präzision oder Vollständigkeit gelegt werden. (vgl. LEWANDOWSKI 2011b, S. 213 – 214)

Recall und Precision mathematisch nicht voneinander abhängig, es hat sich aber gezeigt, um so höher die Precision um so niedriger der Recall und umgekehrt. Gibt das Information-Retrieval-System ein relevantes Dokument zurück liegt die Precision bei 1, aber der Recall ist sehr gering, weil alle anderen relevanten Dokumente im Datenbestand nicht ausgegeben werden. Wenn das System die gesamte Datenbank zu einer Suchanfrage ausgibt (tausende oder millionen von Dokumenten) wird ein sehr hoher Recall erreicht, der aber gleichzeitig eine sehr niedrige Precision bedeutet. (vgl. LEWANDOWSKI / HÖCHSTÖTTER 2008, S. 316 – 317)

Weiterhin weißt LEWANDOWSKI 2011b noch darauf hin, dass sich die Standardmaße in Bezug auf die von BRODER 2002 entwickelte Unterteilung von Suchanfragetypen nur informationsorientierte Suchanfrage mit Recall und Precision messen lassen. Bei informationsorientierten Anfragen versucht der Nutzer eines Systems sein Informationsbedürfnis mit einer bestimmten Anzahl relevanter Dokumente zu befriedigen.

Weitere Anfragetypen nach BRODER 2002 sind navigationorientierte und transaktionsorientie Suchanfragen. Bei navigationsorientierten Anfragen soll eine ganz bestimmte Webseite wiedergefunden werden. Es gibt also nur ein relevantes Dokument. Für die Messung solcher Anfragetypen müssen andere Kennzahlen genutzt werden, z. B. success@n, welche anzeigt, auf welcher Position das Dokument im Durchschnitt bei einer bestimmten Suchmaschine angezeigt wird (vgl. LEWANDOWSKI 2011A, S. 354). Führt der Nutzer eine transaktionsorientierte Suchanfrage durch, verfolgt er das Ziel eine Transaktion auf einer bestimmten Webseite auszuführen. Das kann der Kauf eines Produkts oder die Nutzung einer Anwendung wie ein Spiel sein.

2.2.1.5 Weitere Kennzahlen

Neben den genannten Standardmaße zur Messung der Retrievaleffektivität haben sich in den Jahren auch noch weitere Kennzahlen herausgebildet. Dabei sind Maße für die allgemeinen Evaluierungen, aber auch für die Evaluation Websuchmaschinen entwickelt worden. Im Bereich der Websuche konnte sich bisher aber noch kein Maß durchsetzen Es hat sich nur gezeigt, dass besondere Kennzahlen nötig sind, um die Retrievaleffektivität von Websuchmaschinen zu messen, die einerseits auf Dokumentebene ansetzen und welche die auch die Bewertung der Trefferliste bis zu einem bestimmten Cuttoff-Wert beziehen. Die folgenden Tabellen bietet eine Übersicht zu weiteren Beispielen für klassische Retrievalmaße, die hier der Vollständigkeit halber genannt werden. Häufig handelt es sich bei den weiteren Retrievalmaßen um Ableitungen des Standardmaß Precision. (vgl. VAN RIJSBERGEN 1979, KELLY 2009, S. 109, LAMM 2008, S. 15 & RADLINSKI / CRASWELL)

Maßzahl	Beschreibung
F-Measure	Das F-Measure ist eine Maßzahl, die Recall und Precision miteinander kombiniert und eine Möglichkeit die Gewichtung des Verhätlnisses zwischen Recall und Precision zu steuern.
P@*n*	Bei dieser Kennzahl wird der Precision Wert in den ersten *n*-Treffern einer Ergebnismenge berechnet. Häufige Werte für *n* sind dabei 10 oder 20, passend zu dem typischen Verhalten einen Suchmaschinennutzers, der sich in der Regel nur die ersten Suchergebnisse ansieht.
Average Precision (AP)	Bei der Average Precision wird die Relevanz für jedes zurückgegeben Dokument berechnet, wobei die relevanten Dokumente, die nicht gefunden wurden den Wert 0 erhalten. Die einzelnen Werte der Dokumente werden anschließend zusammengezählt und durch die Anzahl aller relevanten Dokumente in der Dokumentenkollektion dividiert. Die Average Precision berücksichtigt dabei auch die Position der Treffer.
Mean Average Precision (MAP)	MAP ist der Mittelwert der durchschnittlichen Precision für die jeweilige Suchanfrage.

R-Precision	Bei der R-Precision wird die Precision an der R-ten Position im Ranking berechnet. R = die Anzahl der relevanten Dokumente zu der aktuellen Anfrage. Der Parameter eignet sich dazu das Verhalten des Algorhitmus eines Systems für jede einzelne Anfrage zu betrachten.

Tabelle 1: Übersicht klassischer Retrievalmaße

Die gezeigten Retrievalmaße sind nur ein Ausschnitt von Kennzahlen die in der Information-Retrieval-Forschung zur Bestimmung der Qualität von IR-Systemen eingesetzt werden. Über die genannten Retrievalmaße hinaus, hat sich gezeigt, dass weitere Kennzahlen nötig sind, die auch mehr die Funktionsweise und Anforderungen von Websuchmaschinen und ihre Nutzer berücksichtigen. Die unten aufgeführte Tabelle gibt einige Beispiele zu neueren Ansätzen für die Messung der Güte von Suchergebnissen, die durch LEWANDOWSKI 2012 zusammengetragen wurden.

Maßzahl	Beschreibung
Median Measure	Der Median Mesure bezieht sich auf alle Dokumente, die gefunden werden. Dabei wird nicht nur gemessen wie relevant die Ergebnisse sind, sondern auch wie die Relevanz in Bezug auf die nicht-relevanten Ergebnisse aussieht.
Importance of completeness of search results and Importance of precision of the search to the user	Mit diesen beiden Kennzahlen wird versucht das typische Nutzerverhalten in den Evaluationsprozess zu integrieren. Dabei wird gemessen, ob der Nutzer nur wenige Dokumente zur Lösung einer Aufgabe oder eine komplette Treffermenge benötigt und dabei eine niedrige Precision akzeptiert wird.
Salience	Mit der Salience wird die Precision einer Suchmaschine in Relation zum Gesamtabschneiden aller Suchmaschinen bewerten. Dabei wird auch berücksichtigt, dass die Ergebnisqualität allgemein je nach Anfrage schwanken kann und nicht nur zwischen den Suchmaschinen.
Relevance concentration	Misst die Anzahl der relevanten Dokumente mit einem Wert von 4 oder 5 auf einer 5-Punkt-Skala.
CBC ratio	Die CBC ratio misst den Anteil der inhaltsorientierten Klicks an der Gesamtzahl der Klicks, die in einem Suchverlauf. Insgesamt zeigt die Maßzahl wie viele oder wie wenige Klicks notwendig sind, um an die relevante Informationen zu kommen.
Quality of result ranking	Dient dem Vergleich des automatischen Rankings durch den Suchdients und dem Ranking, dass durch menschliche Gutachter erfolgt. Gemessen wird dabei die Übereinstimmung der Rankings.

Ability to retrieve top ranked pages	Wenn die Ability to retrieve top ranked pages bestimmt werden soll, wird eine bestimmte Anzahl der Topergebnisse verschiedener Suchmaschinen zusammengeführt und von Juroren bewertet. Anschließend wird eine bestimmte Menge, z. B. 75% der bestbewerteten herausgefiltert und der jeweilige Anteil der Suchmaschinen an diesen Dokumenten berechnet.

Tabelle 2: Übersicht Retrievalmaße für Websuchmaschinen

Die aufgeführten Beispiele für neuartige Kennzahlen haben gemeinsam, dass sie zwar in vereinzelten Studien und Experimenten eingesetzt wurden, aber bisher nur selten im größeren Rahmen, z. B. innerhalb von Evaluierungsinitiativen zur Messung der Qualität von Suchdiensten genutzt wurden (vgl. LEWANDOWSKI 2007, 7). Ferner haben die meisten dargestellten Kennzahlen auch den Nachteil, dass der tatsächliche Nutzer kaum bei den Messungen Berücksichtigungen findet. Trotz der genannten Einschränkungen bieten Retrievaltests aber eine Grundlage zur Messung der Qualität von Suchergebnissen von Information-Retrieval-Systemen und Suchmaschinen. Im folgenden werden sowohl die Vor- als auch die Nachteile zusammenfassend dargestellt.

2.2.1.6 Vorteile von Retrievaltests

Trotz aller bisher genannten Kritikpunkte an Retrievaltests, die als Hauptmethode für die systemorientierte Evaluierung von Informationsdiensten dienen, ergeben sich doch einige Vorteile, die die Berechtigung dieses Verfahrens aufzeigen:

1. Retrievaltests sind eine kostengünstige Methode zur Überprüfung von Suchsystemen. Diese Art der Evaluierung lässt sich mit relativ wenig Aufwand durchführen, da beispielsweise aufwendige Laborexperimente nicht für die Durchführung der Tests notwendig sind.
2. Durch die lange Geschichte und Tradition von Retrievaltests haben sich eine Vielzahl von Standards und Kennzahlen etabliert, die für vergleichende und wiederholende Tests angewendet werden können.
3. Der systemorientierte Ansatz eignet sich für die Überprüfung eines eigenen Systems, den Vergleich des eigenen Systems mit anderen Systemen oder auch für den Vergleich fremder Systeme untereinander (beispielsweise den bekanntesten Suchmaschinen).

2.2.1.7 Nachteile von Retrievaltests

Neben den genannten Vorteilen bestehen aber auch einige Nachteile, die bereits innerhalb der dargestellten Methodik und den Maßzahlen angesprochen wurden:

1. Retrievaltests blenden das tatsächliche interaktive Verhalten des Nutzers aus. Daher ist es fragwürdig inwieweit die Ergebnisse solcher Untersuchungen auf den realen Benutzer übertragbar sind. Ihnen liegt das Anfrage-Ergebnis Paradigma zu Grunde, das nur noch teilweise für den Web-Kontext gültig ist, das aber nicht zeigt, welche vielfältigen Interaktionen zwischen Nutzer und Suchdienst während den Suchsessions entstehen.
2. Die Präsentation von Suchergebnissen erfolgt heutzutage besonders bei den Websuchmaschinen nicht mehr nur in einer einfachen Listenform. Es findet eine Vermischung verschiedener Ergebnistypen und Medientypen auf der Ergebnisseiten statt. Dieser Umstand führt dazu, dass nicht mehr alle Treffer in den Augen des Nutzers gleich sind.
3. Die Bewertung der Relevanz ist eine Grundlage für die Bewertung von Ergebnissen in den Retrievaltests. Wie bereits gezeigt, ergeben sich bei dem Konzept der Relevanz aber einige Schwierigkeiten, da eine „objektive“ Relevanz nicht definiert werden kann.
4. Trotz des genannten Vorteils über den relativ geringen Aufwand, der betrieben werden muss, damit Relevanztests möglich sind, kann aber auch ein großer Aufwand nötig sein, um für eine große Menge an Suchanfragen eine entsprechende Anzahl an Juroren zu gewinnen. Wie in Kapitel 2.2.1.3 gezeigt muss eine entsprechende Anzahl an Gutachtern gewonnen werden, damit eine real überprüfbare Menge an Suchergebnissen pro Person gewährleistet werden kann.

2.2.2 Systemorientierte Evaluierung mittels Klickdaten

Wie oben noch einmal deutlich geworden, entstehen eine Reihe von Nachteilen bei der Durchführung von Retrievaltests, zum einen wird nur wenig von einem realen Nutzer durch die Tests abgedeckt und zum anderen ist die gesamte Konzeption und auch die Akquirierung von Juroren für die Tests mit einem gewissen Aufwand verbunden. Daher wurde in der Geschichte der Information-Retrieval-Evaluierung nach Methoden gesucht, die einerseits weniger aufwendig sind und andererseits auch eine Vielzahl tatsächlich gestellter Suchanfragen und Suchmaschinennutzer miteinbeziehen. Ein Ansatz ist die Analyse von Logfiles, die durch die Suchsysteme generiert werden. Dabei lässt sich dann mit Hilfe dieser Daten auswerten, welcher Nutzer (Identifikation über IP-Adresse, User-Login, durch Cookies auch Möglichkeit der Wiedererkennung eines Benutzers), welche Suchanfragen gestellt und welche Suchergebnisse gesichtet hat (vgl. HÖCHSTÖTTER 2009, S. 181 – 182).

Daneben kann ggf. auch noch erfasst werden, wie viel Zeit vergeht bevor ein Nutzer von einem Suchergebnis zurück auf die Trefferliste zurückgeht. Bei diesem Verfahren werden dabei die Anzahl der Klicks und die Verweildauer auf einem Suchergebnis gemessen. Suchmaschinen selber können dieses Verfahren nutzen, um das Ranking zu verbessern, da die Anzahl der Klicks sowie die Verweildauer Indikatoren für relevante Treffer sind. Meistens wird ein vorher abgegrenzter Ausschnitt der Logfiles analysiert, z. B. könnten die Suchanfragen eines Monats untersucht werden. (vgl. LEWANDOWSKI 2011b, S. 221 – 222)

Da bei solchen Tests mit Nutzerdaten gearbeitet wird und keine Juroren für die Relevanzbeurteilung benötigt werden, lässt sich eine Vielzahl von Bewertungen zu einem Treffer sammeln. Weiterhin muss berücksichtigt werden, dass ein Klick auf einen Suchergebnis nicht unbedingt bedeutet, dass es sich bei dem Ergebnis, um einen relevanten Treffer handelt. Lediglich die Trefferbeschreibung in der Liste weißt daraufhin, dass das Suchergebnis relevant sein könnte, auch das Ranking selber kann den Nutzer immer beeinflussen, wenn dieser eventuell einfach nur die Top 3 Treffer anklickt, weil ihm die Suchmaschine die Ergebnisse in einer bestimmten Reihenfolge anbietet. Ferner muss auch angemerkt werden, dass keine wirkliche systematische Evaluation der Suchergebnisse stattfindet, da keine direkte Bewertung von Treffer stattfindet. Alle Relevanzurteile werden implizit erfasst. Hierbei ist anzumerken, dass FOX ET AL. 2005 in ihren Studien zeigen konnten, dass die Anzahl der Klicks auf Dokumente ein wichtiger Indikator für die Relevanz des Ergebnisses ist. (vgl. EBD., S. 221 – 222)

In Kapitel 3 wird bei der Zusammenfassung einer Studie von JOACHIMS ET AL. 2005 gezeigt inwieweit Klickdaten in der Praxis für die Identifikation von interaktiven Prozessen durch Nutzer und für die Relevanzbewertung nutzbar sind.

Die folgende Übersicht stellt noch einmal die Vor- und Nachteile von Klicktests zusammenfassend dar.

2.2.2.1 Vorteile von Klicktests

Die Nutzung von Logfiles zur Relevanzbeurteilung bietet einige Vorteile, die bei der Beschreibung zu der Methode schon teilweise aufgeführt wurden:

1. Die Nutzung von Logfiles zur Messung der Retrievaleffektivität ist eine kostengünstige von Juroren unabhängige Methode zur Gewinnung von Daten in der systemorientierten Evaulierung von Suchsystemen.
2. Es fällt eine große Menge an Daten an, die für die Analyse des Systems nutzbar ist. Eigene Erhebungen, wie sie im Rahmen von Retrievaltests durchgeführt werden müssen, sind hier nicht notwendig.

3. Die Basis für die Messung der Retrievaleffektivität sind echte Suchanfragen von echten Suchmaschinennutzern in Verbindung mit ihren Suchmaschinenverhalten in einem bestimmten Zeitraum. Es finden bei Klicktests keine Einschränkungen bei der Anzahl und Auswahl der Suchanfragen und der Anzahl der Juroren statt.
4. Ergebnisse solcher Untersuchungen kann der Betreiber eines Suchdienstes zur Verbesserung seines Rankings nutzen.

2.2.2.2 Nachteile von Klicktests

Auch wenn der Einsatz von Klicktests einige Vorteile und Chancen bietet, gibt es doch einige Nachteile, die hier zusammenfassend dargestellt werden:

1. Die Ausgangslage zur Durchführung solcher Tests ist der Zugriff auf die Logfiles der Suchmaschinen. In der Regel geben Suchmaschinenanbieter diese Daten nicht heraus und damit ist können nur die Betreiber selber und kooperierende Forschungseinrichtungen die Daten für Testdurchführungen nutzen. Vergleichende Tests mit anderen Suchmaschinen und Suchdiensten sind damit nur theoretisch möglich.
2. Anfallende Logfiles für Tests zu benutzen ist ethisch fragwürdig, da die Nutzer der Schmaschinen keine Zustimmung zur Auswertung ihrer Daten geben und durch Cookies oder auch Suchanfragen nach eigenen Personen eine Anonymität nicht immer gewährleistet ist.
3. Relevanzurteile werden nicht direkt gefällt. Es empfiehlt sich für die Bewertung, auch wenn nur das eigene System getestet werden kann, eine Kombination aus Klicktest und Juroren-Bewertung innerhalb von Retrievaltests durchzuführen. Eine alleinige Analyse der Klickdaten sagt zu wenig über die Relevanz der Suchergebnisse aus.
4. Daten aus Logfiles liefern keine Erklärungen zum Verhalten und zu den Absichten des Suchmaschinennutzers. (vgl. MANDL 2006)

Die bisher vorgestellten Methoden der Information-Retrieval-Evaluierung fanden bisher auf der Ebene der Suchtreffer selber oder der Trefferliste statt. Das Verhalten des Nutzers wird bei diesen Verfahren nur wenig betrachtet. Innerhalb klassischer Retrievaltests findet der Benutzer Beachtung durch eine Definition der Nutzergruppe, die innerhalb der Tests abgebildet werden soll. Bei der Analyse von Logfiles wird der Nutzer mit seinen tatsächlichen Suchanfragen und seinem Klickverhalten etwas stärker in den Fokus gerückt. Dabei können aber keine direkten Beobachtungen zu seinem Verhalten erfolgen. Es besteht nur die Möglichkeit Vermutungen zu seinen Intentionen zu äußern und auf Grundlage der Daten Versuche zu unternehmen sein Verhalten nachzuvollziehen und zu bewerten.

Darüber hinausgehende Informationen erhalten die Testleiter nicht durch Hilfe der Klickdaten. Da aber der Nutzer eine zentral Rolle bei der Evaluierung von Informations-Retrieval-Systemen und Suchmaschinen spielt, haben sich in der Geschichte auch benutzerorientierte Ansätze entwickelt, um Systeme zu testen.

Der folgende Abschnitt bietet eine kurze Übersicht zur benutzerorientierten Evaluierung von Informationssystemen und geht dabei auch stärker auf die Rolle des Nutzers innerhalb des Information-Retrieval ein.

2.3 Benutzerorientierte Evaluierung

Im Gegensatz zu der systemorientierten Evaluierung verfolgt die benutzerorientierte Evaluierung das Ziel Anwendungssituationen in Informationssystemen realistisch zu simulieren. Der Benutzer steht bei benutzerorientierten Methoden im Fokus der Untersuchung. Daher werden für Tests in diesem Bereich Versuchspersonen benötigt, die innerhalb vorgegebener Anwendungsszenarien mit einem oder mehreren Systemen interagieren. Je nach zu untersuchender Fragestellung können dabei die Suchprozesse selbst, die Qualität der Lösungen, die Beobachtungen durch den Testleiter oder das persönliche Erleben der Probanden stehen. Diese Ziele zeigen bereits, dass eine benutzerorientierte Evaluierung aufwendiger ist als systemorientierte Methoden. Weiterhin steht dabei auch nicht immer unbedingt die Qualität der Informationssysteme bzw. bestimmte Aspekte und Funktionen des Systems im Vordergrund. Bei Untersuchungen zum Nutzerverhalten werden dann auch Methoden eingesetzt, die aus dem Bereich der Psychologie stammen (vgl. KELLY 2009, S. 12).

Ein Ausgangspunkt für benutzerorientierte Evaluierungsmethoden ist die kognitive Sichtweise auf Information-Retrieval-Systeme und Suchmaschinen. Innerhalb dieser Betrachtung, werden die komplexen kognitiven Handlungen der Nutzer miteinbezogen, die während aller Interaktionen mit dem System anfallen. Der cognitive Viewpoint wurde das erste Mal im Workshop on the cognitive Viewpoint von DE MEY 1977 formuliert. DE MEY 1977 formulierte, dass hinter jedem Informationsprozess Systeme von Kategorien und Konzepten stehen, die als Model innerhalb der Systeme als eine eigene Welt abgebildet werden. Diese Definition trifft dabei sowohl auf Computersystem als auch auf den Nutzer selber zu. Der Nutzer interagiert immer mit seiner eigenen konzeptionellen Welt und der abgebildetem Wirklichkeit innerhalb des Systems.

INGWERSEN / JÄRVELIN 2005 haben fünf zentrale und zusammenhängende Aspekte herausarbeitet, die sich auf die kognitiven Strukturen von Nutzern, aber auch auf die Abbildung im Information-Retrieval-System, z. B. repräsentiert durch die Systementwickler und den Redakteuren, beziehen.:

1. Die Verarbeitung von Informationen findet auf Seite des Senders und des Empfängers statt. Dabei können Nutzer und Systeme beide Rollen einnehmen.
2. Die Verarbeitung findet auf verschiedenen Ebenen statt. Während die Systeme auf die Verarbeitung von Daten mit Hilfe von Indexierungssprachen beschränkt sind, stehen dem Nutzer weitere Ebenen zur Verfügung. Er ist in der Lage Informationen zu Interpretieren und in dem Kontext seiner eigenen Erfahrungen, Erwartungen und Ziele zu sehen und zu verarbeiten.
3. Während der Kommunikation von Informationen ist jeder Akteur im System durch seine vergangenen, seine gegenwärtigen Erlebnisse und Erfahrungen (Zeitpunkt der Nutzung) und durch sein soziales, organisatorisches und kulturelles Umfeld beeinflusst.
4. Jeder Akteur beeinflusst das System, die Umgebung und Domäne.
5. Informationen sind situations- und kontextbezogen.

Durch die Darstellung der aufgeführten Aspekte wird deutlich, welche Komplexität die Evaluierung von Informationssystemen sich bei der Untersuchung unter kognitiven Gesichtspunkten ergibt. Jeder Nutzer erlebt Prozesse des Information Retrieval unterschiedlich und es ist auch davon auszugehen, dass während der Nutzung verschiedene kognitive Veränderungen stattfinden, die nicht beobachtet werden können, aber höchstwahrscheinlich das Nutzerverhalten beeinflussen. Daher sind, wie bei den systemorientierten Methoden, auch vereinfachende Schlussfolgerungen und Abstraktionen notwendig.

2.3.1 Erhebungsmethoden zur benutzerorientierten Evaluierung

In der Evaluierung von Softwaresystemen wird zwischen zwei Arten von Erhebungsverfahren unterschieden. Auf der einen Seite gibt es die objektiven Methoden, die messbare Ergebnisse in Zahlen liefern sollen. Auf der anderen Seite gibt es die subjektiven Methoden, die verbal geäußerte Ergebnisse von Probanden liefern sollen. Deshalb werden die Verfahren auch als quantitativ und qualitativ bezeichnet. (vgl. OPPERMANN / REITERER 1994, S. 342ff. & DAHM 2006, 319)

Weiter unterscheidet HÖCHSTÖTTER 2008 auch noch zwischen überwachten und nicht überwachten Methoden bei der Datenerhebung in der Suchmaschinenforschung. Bei den überwachten Methoden ist ein Versuchsleiter anwesend oder die Probanden haben Kenntnis darüber, dass sie getestet werden.

Werden unüberwachte Methoden angewendet, werden Testpersonen ohne deren Wissen und häufig auch ohne deren Einverständnis beobachtet. Die Datenerhebung erfolgt dabei automatisiert und erzeugt häufig eine große Menge von Daten.

2.3.1.1 Objektive Methoden

Die Beobachtung ist das gängigste Verfahren der objektiven Evaluierung (vgl. OPPERMANN / REITERER 1994, S. 343). Die Beobachtungen werden dabei häufig in Laborexperimenten durchgeführt, in denen Videoaufzeichnungen, die Protokollierung mit Hilfe von Logfiles oder auch Eye-Tracking Geräte eingesetzt werden. Alle genannten Hilfsmittel ergänzen die Beobachtungen des Testleiters. Durch Logfiles, lassen sich einzelne Benutzereingaben und Klicks genau nachvollziehen. Da Logfiles aber nur wenig über Motive und tatsächliche Handlungen der Testpersonen aussagen können, wird häufig mit Hilfe von Videos das Verhalten des Benutzers und dessen Äußerungen während der Beobachtung aufgezeichnet. Zusätzlich könnten Eye-Tracking Geräte noch dabei helfen, zu erfassen welche Bereiche auf dem Bildschirm vom Probanden angesehen wurden. (vgl. HÖCHSTÖTTER 2008, S. 177ff.) Eye-Tracking Studien wurden bereits oft im Kontext von Websuchmaschinenstudien eingesetzt (z. B. in BUSCHER ET AL. 2004, GRANKA ET AL. 2004, JOACHIMS ET AL. 2005, LORIGO ET AL. 2008).

Laborexperimente sind häufig Teil von Usability-Untersuchungen bestehender Information-Retrieval-Systeme und Suchmaschinen (vgl. SARODNICK / BRAU 2006, S. 219ff.). Sie führen auch zu Ergebnissen dazu, wie Benutzer ihre Informationsbedürfnisse mit dem zu untersuchenden System bearbeiten, sind für die Prüfung von Funktionen in der Praxis geeignet und liefern Daten über die Performanz. Nachteile ergeben sich bei Laborexperimenten aber durch eine künstliche Umgebung, die ein Verhalten auf die Probanden und damit auch auf die Ergebnisse hat (vgl. HÖCHSTÖTTER 2008, S. 177 – 178 & KELLY 2009, S. 27 – 28).

Weiterhin können Zeit- und Fehlermessungen vorgenommen werden, um Aussagen über die Qualität des Systems zu erhalten. SARODNICK / BRAU 2006 nennen als Messgrößen dafür beispielsweise die Zeit, die eine Testperson für die Bearbeitung eines Suchaufgabe benötigt oder auch die Anzahl der Aufgaben, die innerhalb einer definierten Zeitspanne mit Hilfe des Systems gelöst werden. Daneben kann auch in Bezug auf Bedienungsfehler gemessen werden, wie das Verhältnis zwischen fehlerhaften und erfolgreichen Benutzereingaben aussieht oder die Zeit, die benötigt wird, um eigenständig Fehler zu lösen. Solche Messungen sind aber immer nur im Zusammenhang mit den gestellten Testaufgaben valide und bei anderen Situationen und Aufgaben durchaus stark abweichende Ergebnisse denkbar sind.

Zu weiteren objektiven Methoden zählen Liveticker und sogenannte Zählpixel. Liveticker werden von ein paar Suchmaschinen angeboten. Sie eignen sich dazu Suchanfragen zu erfassen, die von Nutzern in Echtzeit gestellt werden. Weitere Informationen beispielsweise zum Nutzer oder auch zu den Suchergebnissen lassen sich dadurch aber nicht beziehen. Zählpixel sind nicht sichtbare sehr kleine Grafiken, die auf Webseiten platziert werden. Sie helfen dabei das Navigationsverhalten von Benutzern zu erheben Jeder Aufruf eines Zählpixel generiert einen Eintrag in einer Logdatei über den Pfad, den Nutzer auf einer Webseite verfolgen. Diese Art der Erhebung bietet aber keine weiteren Informationen zu den Benutzern. (vgl. HÖCHSTÖTTER 2008, S. 179ff.)

2.3.1.2 Subjektive Methoden

Zu den subjektiven Methoden in der Evaluierung zählen beispielsweise Befragungen und die Think-Aloud-Methode, bei der Testpersonen sämtliche Interaktionen mit einem System verbal kommentieren sollen. Bei subjektiven Methoden steht der Benutzer noch mehr im Vordergrund. Das persönliche Erleben des Nutzers ist dabei die Grundlage der Untersuchung.

Befragungen sind eine der weitestverbreiteten Methoden in der empirischen Nutzerforschung und können zu objektiv erhobenen Daten noch Aussagen der Testpersonen gesammelt durch Fragebögen und Interviews gesammelt werden. So bieten Befragungen beispielsweise die Möglichkeit persönliche Einschätzungen und Eindrücke, wie Stärken und Schwächen zu einem zu evaluierenden System durch die Probanden zu erheben. Schriftliche Befragungen bieten dabei den Vorteil, dass sie sich einfacher auswerten lassen. Ein Problem besteht hierbei aber in der Validität der Daten, da Probanden das Gefühl haben könnten selber getestet werden und daher nicht immer wahrheitsgemäße Antworten auf die Fragen geben. (vgl. HÖCHSTÖTTER 2008, S. 177 – 178). Diese Gefahr lässt sich durch die angemessene Gestaltung des Fragebogens mindern.

Die Erhebungsmethode des lauten Denkens bietet den Vorteil, dass die Motive der Benutzer festgehalten werden. Nachteile ergeben sich nach SARODNICK / BRAU 2006 durch die Doppelbelastung der Probanden, die einerseits Aufgaben lösen müssen und dabei auch noch jeden Schritt kommentieren sollen. Ferner mach diese Methode die Testperson auch auf Inkonsistenzen im Denken aufmerksam und kann dadurch Probleme generieren, die bei einer Benutzung unter realen Bedingungen nicht auftauchen. Durch Äußerungen während der Interaktionsprozesse sinkt auch die Bearbeitungsgeschwindigkeit, was bedeutet das Performanzmessungen nicht mit der Think-Aloud-Methode kombinierbar sind.

In der Praxis empfiehlt es sich objektive und subjektive Methoden zu kombinieren, um eine ganzheitliche Betrachtung des Systems zu erhalten.). Dabei müssen die Methoden je nach Evaluierungsschwerpunkte gewählt werden. Es ist beispielsweise üblich innerhalb von Laborexperimenten auch subjektive Eindrücke von Probanden zu erfassen, z. B. durch die Videoaufzeichnungen der Versuchspersonen.

2.3.1.2 Explorative Suche

Eine bestimmte Art der Benutzerinteraktion innerhalb von Suchmaschinen oder Information Retrieval-Systemen ist die explorative Suche. Explorative Suchen werden oft als offene, abstrakte und wage und mehrdeutige Informationsbedürfnisse beschrieben, bei denen die Suchvorgänge opportunistisch, iterativ und mit wechselnden Suchstrategien erfolgen (vgl. WHITE ET AL. 2008, S. 433). MARCHIONINI 2006 beschreibt die explorative Suche sogar als eine grundlegende Lebensaktivität und zeigt in einer seiner Darstellungen drei Arten von Suchaktivitäten, wobei die explorative Suche sich aus Learn und Investigate zusammensetzt. Führt ein Nutzer eine explorative Suche durch, fehlt es ihm an Wissen über das Thema, zu dem er recherchiert. Daher ist ein einfaches Fakten-Retrieval der gewünschten Information nicht möglich. Der Nutzer wird in diesem Fall eine Reihe von Aktivitäten innerhalb komplexer mehrfacher Suchen und weiterer Informationsprozesse benötigen. Explorative Suchen sind hoch interaktive Vorgänge, in denen der Nutzer innerhalb der Prozesse dazu lernt und die Ergebnisse, die seine Anfragen generieren, untersuchen muss.

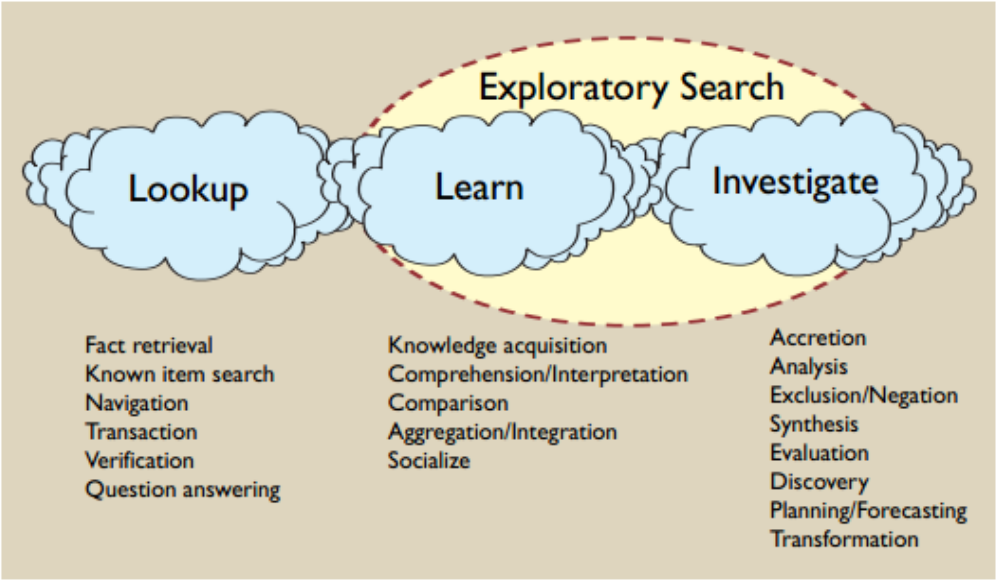


Abbildung 5: Suchaktivitäten in der explorativen Suche (Quelle: Marchionini 2006, S. 42)

Die erste Art von Suche in Marchionis Modell ist die Look-Suche. Sie folgt dem Anfrage-Ergebnis Paradigma, das die Ausgangslage für klassische Retrievaltests bildet. Der Suchende recherchiert dabei Fakten oder sucht eine Antwort auf eine bestimmte Frage. Dabei wird oft nach Wer? Wann? und Wo? gefragt. Sollte für den Nutzer aber auch das Wie? und Warum? wichtig sein, beginnt eine explorative Suche, in der das Searching to learn eine Art von Suche ist. Der Nutzer versucht mit seinen Suchen Ergebnisse zu finden, die ihn tiefgründigere Informationen und Antworten liefern. Er lernt durch seine Suche etwas zu dem Thema hinzu, für das er recherchiert. Bei dieser Suche werden iterative Schritte durchgeführt. Der Nutzer gibt Suchanfragen ein, sichtet die Ergebnisliste, schaut sich Dokumente selektiv an, interpretiert und verarbeitet die Informationen verändert eventuell seine Suchstrategie, um sich weiter zu informieren und ergibt erneut eine Suchanfrage ein, die ein weiteres Set von Ergebnissen erzeugt. Eine weitere Form der Suche, die auch innerhalb explorativer Suchen stattfindet, ist Searching to investigate erforder wieder eine Vielzahl iterativer Schritte und wird oft auch über einen längeren Zeitraum durchgeführt. Sie dienen dazu Vorhersagen und Planungen durchzuführen oder Evaluierungen und Analysen zu bestimmten Themen durchzuführen. MARCHIONINI 2006 deutet darauf hin, dass diese Art von Suche die anspruchsvollsten Ziele einer Suche verfolgt.

Die Darstellung der verfolgten Ziele einer explorativen Suche zeigen, dass solche Prozesse häufig nicht innerhalb weniger Minuten durchzuführen sind, oft benötigen Suchende Stunden oder Tage zur Lösung ihres Problems. In Bezug auf Websuchmaschinen bieten diese auch immer nur den Zugangspunkt für die Durchführung einer explorativen Suche. Der Nutzer entscheidet dabei noch über weiter Suchdienste, die er für die Bearbeitung der Aufgabe benutzen möchte und führt dabei auch Aktivitäten durch, die zur explorativen Suche gehören, in dem er selektiv Suchergebnisse anklickt, Suchanfragen verändert, andere Suchmaschinen hinzuzieht.

Das Verhalten der Nutzer bei der Bearbeitung explorativer Suchaufgaben lässt sich mit oben genannten subjektiven und objektiven Methoden in der benutzerorientierten Evaluierung untersuchen. Dabei wird in der Praxis häufig auf Laboruntersuchungen zurückgegriffen. Bei denen kleine Testgruppen beobachtet werden können, aber durch die Begrenzung auf eine kleine Gruppe von Testpersonen nur wenig repräsentative Ergebnisse erzeugt werden. Ein anderer Ansatz bieten Logdateien, die aber für sich nicht genug Informationen zum tatsächlichen Nutzerverhalten bieten. Eine Lösung dieses Problems können Browserplugins schaffen, da diese das wirkliche Verhalten des Nutzers bei der Nutzung von Suchmaschinen aufzeichnen.

Innerhalb der Forschung wurden auch bereits Ansätze diskutiert, um sogenannte Exploratory Search Systems zu evaluieren (vgl. WHITE ET AL. 2006). Dabei wurden auch diverse Kennzahlen und Maße entwickelt, um die Güte solcher Systeme zu messen. Von diesen Kennzahlen werden auch einige kurz im nächsten Kapitel zu der Übersicht von Kennzahlen in der benutzerorientierter Evaluierung aufgeführt.

2.3.1.3 Kennzahlen und Maße in der benutzerorientierten Evaluierung

Im Gegensatz zur systemorientierten Evaluierung von Informationsystemen, die vor allem über Retrievaltests durchgeführt werden, haben sich innerhalb der benutzerorientierten Evaluierung bisher keine Standardmaße herausgebildet. Für die Messung der Güte von Systemen aus der Benutzerperspektive werden Kennzahlen aus verschiedenen Disziplinen eingesetzt. Findet die Evaluierung unter den Gesichtspunkten der Usability des Systems statt können die Standardmaße der Effizienz und Effektivität aus der DIN EN ISO 9241 Teil 11 angewendet werden. Dabei bezeichnet Effektivität die Genauigkeit und Vollständigkeit, mit der die Benutzer ein bestimmtes Ziel erreichen. Die Effizienz steht für den Aufwand in Relation mit der Genauigkeit und Vollständigkeit der Zielerreichung, also praktisch für das Kosten-Nutzen-Verhältnis und die Zufriedenstellung wird als „Freiheit von Beeinträchtigungen und positive Einstellungen gegenüber der Nutzung des Produkts“ (DIN 9241 1998, S. 4) definiert.

Wie bereits oben erwähnt sind aber auch nach SARODNICK / BRAU 2006 die Zeit oder die Anzahl des gelösten Aufgaben innerhalb einer bestimmten Zeitspanne als Messgrößen denkbar. Das generelle Problem dieses Größen wurde schon benannt. Solche Zahlen lassen sich nur auf den konkreten Anwendungsfall beziehen und machen damit eine Vergleichbarkeit schwer.

Weitere Kennzahlen, die auch für die benutzerorientierte Evaluierung nutzbar sind wurden im „Workshop on Evaluating Exploratory Search Systems“ (vgl. WHITE ET AL. 2006) definiert:

Maßzahl	Beschreibung
Engagement and enjoyment	Misst die Zufriedenheit der Nutzers mit dem System. Dabei soll erfasst werden wie engagiert der Benutzer ist und welche positiven Gefühle er dabei ausdrückt. Als Vorschlag für eine messbare Kennzahl könnte die Anzahl der genutzten interaktiven Möglichkeiten innerhalb des Systems sein.
Task success	Beim task success wird nicht nur gemessen, ob der Benutzer mit seiner Anfrage bereits relevante Dokumente zu seiner Fragestellung gefunden hat, sondern es soll gemessen werden, ob bereits eine zufriedenstellende Anzahl von Informationen zur Lösung der Problems, durch den Benutzer mit Hilfe der Suchergebnisse zusammengetrafen wurden.
Task time	Die Zeit, die benötigt wird um einen bestimmten Grad für die Fertigstellung der Aufgabe wird gemessen, um die Effizienz zu erfassen. Dabei sollen die Gesamtzeit, die Zeit der Sichtung nicht relevanter Dokumente und das Verhältnis von direkter Suche und Erkundung erhoben und ausgewertet werden.
Learning and cognition	Messung der kognitiven Anstrengung des Benutzer, die erbrachte Leistung des Lernprozesses, die Zielerreichung aus der Sicht des Nutzers nach durchgeführter Aufgabe, der Grad der Zielerreichung bezogen auf das Thema und die Anzahl der neuen Erkenntnisse der Testperson.

Tabelle 3: Übersicht möglicher Kennzahlen zur benutzerorientierten Evaluierung

Alle genannten Kennzahlen zeigen, dass sie sich schwer für eine Vergleichbarkeit von Information-Retrieval-Systemen nutzen lassen bzw. eine Messung bestimmter Faktoren und Engagement sich nur schwer in Zahlen ausdrücken lässt. Die Abhängigkeit der kognitiven Fähigkeiten und interpersonellen Voraussetzungen der Testpersonen, machen eine Evaluierung aus der Benutzerperspektive schwieriger als systemzentrierte Methoden. Die beiden nächsten Kapitel fassen noch einmal die Vor- und Nachteile der benutzerorientierten Evaluierung zusammen.

2.3.1.4 Vorteile von benutzerorientierten Evaluierungsmethoden

Trotz der Schwierigkeiten, die sich besonders bei der Standardisierung benutzerzentrierter Methoden und Kennzahlen zeigen, gibt es doch einige Vorteile, die noch einmal zusammenfassend dargestellt werden:

1. Durch benutzerzentrierte Methoden erhalten Betreiber von Informationssystemen viele Informationen über die Retrievaleffektivität ihres Systems hinaus. Die gewonnenen Erkenntnisse können beispielsweise dabei helfen die Usability zu verbessern.
2. In benutzerorientierten Einsätzen wird das tatsächlich Verhalten von Benutzern berücksichtigt. Die Prämisse, dass ein Suchender eine Vielzahl von Dokumenten sichtet, ist hier nicht der Ausgangspunkt.
3. Spezielle Anwendungsfälle wie explorative Suchen oder Suchsessions lassen sich mit dieser Art von Evaluierung abbilden.
4. Die Art der Evaluation bezieht Faktoren wie die Darstellung der Suchergebnisse von Suchmaschinen mit ein.

2.3.1.5 Nachteile von benutzerorientierten Evaluierungsmethoden

Im Folgenden werden auch die Nachteile benannt, die bei Evaluierungen mit Fokus auf die Benutzer bestehen:

1. Es gibt bisher keinen Standardaufbau und keine Standard-Maßzahlen für eine benutzerorientierte Evaluierung von Informationssystemen. Dadurch sind vor allem vergleichende und wiederholende Evaluierungen schwer durchführbar.
2. Die gängige Methode von Laborexperimenten ist mit viel Aufwand und Kosten verbunden. Daneben bestehen Testgruppen im Labor häufig nur aus einer geringen Anzahl von Nutzern, was zu wenig repräsentativen Ergebnissen führt.
3. Die Probanden in solchen Evaluierungen bringen alle verschiedene kognitive Fähigkeiten und Voraussetzungen mit. Dadurch werden Testergebnisse von vornherein verzerrt.

2.4 Interactive Information Retrieval

Die bisher vorgestellten Methoden bezogen sich mehr oder weniger immer auf eine Sichtweise, wie sich Informationssyteme und Suchmaschinen untersuchen lassen. Entweder Stand das System mit den Algorithmen und der Retrievaleffektivität im Vordergrund oder der Benutzer mit seinem Verhalten. Eine Disziplin, die versucht beide Sichtweisen der Evaluierung zu kombinieren ist das sogenannte Interactive Information Retrieval. Was sich vor allem mit der Beschaffenheit interaktiver Information Retrieval Systeme und deren Untersuchung befasst:

„The overall purpose of the evaluation of IIR systems is to evaluate the systems in a way which takes into account the dynamic natures of information needs and relevance as well as reflects the interactive information searching and retrieval processes.„ (BORLUND 2003)

Nach BORLUND 2003 bietet IIR die Möglichkeit beide Ansätze zu kombinieren. Auch Kelly ordnet IIR Studien in der Mitte zwischen systemfokusierten und benutzerfokusierten Studien ein. Während sich Information Retrieval mit der Frage beschäftigt, ob Suchdienste relevante Dokumente wieder auffinden können, fragt Interactive Information Retrieval auch danach, ob die Benutzer das System verwenden können, um relevante Dokumente zu finden. Dabei könnte angenommen werden, dass benutzerzentrierte Ansätze dieser Frage bereits nachgehen. Das Problem, dass sich dabei aber ergibt, besteht darin, dass benutzerorientierte Methoden meist nur auf eine kleine Anzahl von Testpersonen beschränkt ist und auch nur wenige vergleichbare Daten anfallen. Oft werden im Bereich dieser Methoden eher Studien zum Nutzerverhalten allgemein erhoben ohne wirklichen Bezug zu den genutzten Systemen.

LEWANDOWSKI 2011b spricht davon, dass sessionbasierte Evaluierungen immer stärker in den Fokus von Untersuchungen rücken. Er nennt dabei die Möglichkeit Anwendungen zu nutzen, die das tatsächliche Nutzerverhalten protokollieren und dabei beispielsweise Klickdaten und Suchanfragen erheben. Daneben können auch demografische Daten und Suchmotivationen erfasst und für die Auswertung des Nutzers und sein Verhalten verwertet werden. Auch tiefere Nutzerbefragungen sind denkbar. Insgesamt lassen sich qualitative und quantitative Daten zusammen nutzen. Die Protokolle und Klickdaten können dann auch dahin genutzt werden, der Testperson die tatsächlich angeklickten Teffer noch einmal explizit auf die Relevanz beurteilen zu lassen. So ergäbe sich ein Ansatz beide Evaluierungssichtweisen zu kombinieren, in dem reales Nutzerverhalten in Form von Protokollen aufgezeichnet wird und eine nachträgliche Messung der Retrievaleffektivität auf Grundlage der generierten Daten erfolgt. Nachteile ergeben sich dabei allerdings durch die Akquirierung von Testpersonen, da diese entweder an einem Laborexperiment teilnehmen oder eine Anwendung zum protokollieren auf ihrem eigenen System installieren müssten. Weiter wird bei interaktiven Szenarien auch eine Menge von Daten anfallen, die sich nicht systematisch evaluieren lassen und Beeinflussung durch die Marken bestimmter Suchmaschinen lassen sich nicht ausschließen, wenn eine freie Wahl der Suchdienste besteht. Einige Vorteile dieses Verfahrens wurden aber schon benannt, so könnten Retrievaltests basierend auf den generierten Daten erfolgen und damit die echten Informationsbedürfnisse, Suchanfragen und geklickten Treffer berücksichtigen.

2.5 Überblick zu Evaluierungsinitiativen

Die vorangegangen Abschnitte haben gezeigt, dass die erfolgreiche Evaluierung von Informationssystemen einer sorgfältigen Vorbereitung bedarf. Es ergeben sich eine Menge von Kriterien, die sowohl bei systemorientierten als auch bei benutzerorientieren Untersuchungen berücksichtigt werden müssen. Probleme bei den Evaluierungen ergeben sich vor allem bei der Vergleichbarkeit von Systemen, der Nutzung von Testkollektionen und bei der Wiederholung von Tests mit gleichen Standards und Methoden. Aus den genannten Gründen haben sich in den letzten Jahrzehnten eine Reihe von Evaluierungsinitiativen entwickelt, die einerseits die Plattformen, Dokumentenkollektion, Jurorengruppen für die systematische Überprüfung von IR-Systemen anbieten und auf der anderen Seite auch ein Forum für den Austausch von Systementwicklern und Wissenschaftlern bieten.

2.5.1 Text Retrieval Conference (TREC)

Die Text Retrieval Conference wurde das erste Mal im Jahr 1992 durchgeführt und stellt den Defacto-Standard für die Information-Retrieval-Evaluierung dar. Sie wird jährlich vom National Institute for Standards and Technology (NIST). Durch TREC[3] konnte das erste Mal eine Möglichkeit der Vergleichbarkeit von Systemen untereinander geschaffen werden. Der Untersuchungsschwerpunkt lag bei TREC zunächst auf sogenannten Ad-hoc und Routing-Aufgaben. Bei Ad-hoc-Aufgaben wird geprüft, inwieweit System dazu in der Lage sind relevante Dokumente zu unbekannten Testanfragen in einer bekannten Dokumentenkollektion zu finden. Routing-Aufgaben hingegen beziehen darauf wie gut relevante Dokumente zu bekannten Testanfragen in einer unbekannten Testkollektion wieder auffinden. Bei der ersten TREC wurden Überschriften von Nachrichtenartikeln extrahiert, um tatsächliche Informationsbedürfnisse zu generieren. Der Nutzer wird von TREC dabei als „dedicated searcher“ definiert. Diese Art von Benutzer ist bereit, sich ausführlich mit Hilfe hunderter Dokumente zu einem Thema zu beschäftigen und grenzt sich dabei vom durchschnittlichen Suchmaschinennutzer ab. (vgl. MANDL 2008, S. 28 – 29 & KELLY 2009, S. 17ff.)

Neue Ergebnisse und Anregungen aus der Forschungsgemeinschaft haben zu der Entwicklung sogenannter Tracks geführt, die für die Evaluierung in spezielleren Information-Retrieval Bereichen gedacht sind. In Bezug auf die Untersuchung der Nutzerinteraktionen haben wurden dabei der Interactive Track, der HARD Track und ciQA durchgeführt (vgl. KELLY 2009, S. 18ff.).

3 Im Internet erreichbar unter http://trec.nist.gov/

Ziel dieser Tracks war zu den eher technisch ausgerichteten Systemevaluierungen auch den Nutzer und seine Interaktionen stärker zu berücksichtigen (vgl. Lewandowski 2011b, S. 207, MANDL 2008, S. 28ff.). Mit dem HARD Track wurde beispielsweise versucht Informationen über den Suchenden und den Kontext der Suche in Ergebnislisten mit zu berücksichtigen, aber das tatsächliche Nutzerverhalten wird bei der Messung der Retrievealeffektivität bisher nicht berücksichtigt.

Die Teilnahme an den Tracks steht den Anbietern von Informations-Retrieval-Systemen zur Verfügung. TREC bietet diesen die Möglichkeit die Retrievaleffektivität ausführlich mit bewährten Methoden testen zu lassen. Die Voraussetzung ist also immer eine Freiwilligkeit, was bedeutet das im Rahmen von TREC keine der kommerziell erfolgreichen Suchmaschinen untersucht wird, da diese in der Regel ihr System nicht offenlegen. Auch bezogen auf die Inhalte in World Wide Web bietet TREC keine vergleichenden Untersuchungen von Suchdiensten, damit mit vordefinierten Testkollektionen gearbeitet wird. Zwar wurden in der Vergangenheit auch Dokumentenkollektionen mit Web-Inhalten erstellt beispielsweise im Web Track, doch decken diese Zusammenstellungen nur einen kleinen Teil der Webseiten und Dokumente ab und lassen sich nur bedingt auf das gesamte Web übertragen.

2.5.2 Cross-Language Evaluation Forum (CLEF)

Die europäische Initiative CLEF[4] ist aus dem Cross-Language Track (CLIR) von TREC hervorgegangen und wird seit 2000 eigenständig vom Istituto di Scienza e Tecnologie dell'Informazione – Consiglo Nazionale delle Ricerche (ISTI-CNR) koordiniert. Die Organisation übernehmen verschiedene Institutionen aus unterschiedlichen Sprachräumen. Schwerpunkt dieser Initiative ist das multilinguale Information Retrieval und die Besonderheiten, die sich dabei für die einzelnen Systeme ergeben. Mit CLEF sollen multilinguale Systeme eine Plattform erhalten, um die Retrievaleffektivität auch bei Anfragen in anderen Sprachen messen und testen zu lassen. (vgl. vgl. MANDL 2008, S. 30ff.)

Da CLEF als europäische Konferenz startete, wurden zunächst Dokumentenkollektionen in den europäischen Kernsprachen zur Verfügung gestellt. Im Laufe der letzten Jahre sind aber auch viele weitere Sprachen, z. B. Arabisch und Russich mitaufgenommen wurden.

Wie bei TREC passt sich CLEF Neuerungen aus den Informationswissenschaften an. So wurden 2004 auch ein Interactive CLEF (iCLEF) und 2005 ein Web CLEF initialisiert.

4 Aufrufbar über http://www.clef-initiative.eu

Nennenswert in Bezug auf die benutzerorientiere Evalueirung ist hier auch die Entwicklung eines sogenannten Robust Tracks. Dabei wird davon ausgegangen, dass die Qualität der wiedergegeben Ergebnisse von der Schwierigkeit der Topics abhängt. Bei der Evaluierung werden daher schwierigere Aufgaben stärker gewichtet als leichte. Ziel des Tracks ist die Messung der Stabilität des Systems in Bezug auf die Retrievaleffektivität und nicht auf die erbrachte durchschnittliche Leistung. Ergebnisse des Robust Tracks zeigten, dass sich schlechte Retrievalergebnisse in der Regel besonders negativ auf die Zufriedenheit der Benutzer und deren Gesamteindruck des Suchdienstes auswirken. (vgl. LAMM 2008, S. 13)

2.6 Zusammenfassung

Das Ziel des vorgestellten Kapitels, war einen Überblick zu den Methoden, sowie zu den Schwierigkeiten im Bereich der Qualitätsmessung und Evaluierung von Informations-Retrieval-Systemen und Suchmaschinen zu geben. Dabei wurde deutlich, dass der Bereich der systemorientierten Evaluation bereits einige Standards und Methoden bietet, die bei sorgfältiger Planung und Durchführung von Tests eine solide Grundlage bieten. Im Bereich der benutzerzentrierten Evaluierung fehlen hingegen fest definierte Abläufe und Standards, auch mangelt es dabei an passenden Kennzahlen, die Vergleichbarkeitsstudien ermöglichen. Abhilfe schafft hier auf bestimmten Ebenen das Interactive Information Retrieval, das sich vor allem auf den Aspekt der Interaktivität beschränkt.

3. Evaluierungsstudien

In dem Bereich der Evaluierung von Informationssystemem und Suchmaschinen wurden in den letzten Jahrzehnten zahlreiche Studien durchgeführt. Einige konzentrierten sich dabei auf die Retrievaleffektivität, auf Effektivität von Indexierungssprachen, auf Vergleiche zwischen IR-Systememen untereinander, Websuchmaschinen untereinander, Fachdatenbanken im Vergleich zu Websuchmaschinen. Die Messung der Qualität von Information-Retrieval-Systemen ist eine traditioneller Forschungsbereich mit zahlreichen Untersuchungen. Neben der Evaluierung der Systeme erfolgten auch zahlreiche Studien zu Benutzern und ihrem Verhalten sowie zur Usability von Suchdiensten.

In diesem Abschnitt werden Studien aufgeführt, die einen Nutzen für die Entwicklung des Prototyps haben. Es werden Studien aufgezeigt, die sich mit sessionbasierten Nutzerverhalten, Erhebung von Klickdaten zur Relevanzbewertung und zur Analyse vom Nutzerverhalten sowie einer zusätzlichen expliziten Bewertung von Suchergebnissen durch Beurteiler beschäftigten. Bei der Auswahl der Studien, ging es vor allem darum, Untersuchungen zu finden, die Nutzerverhalten protokollierten, um daraus Schlüsse auf das Verhalte und die Relevanzbewertungen von Dokumenten zu ziehen. Die Nutzung vom implizierten Feedback zur Verbesserung des Rankings bei Websuchmaschinen ist eine Methode im Relevance Feedback. KELLY 2009 nennt Relevance Feedback auch eine der ersten Interaktionen, die im Zusammenhang mit Information Retrieval auftauchten. Relevance Feedback ist allgemein eine Methode, die es dem Benutzer ermöglicht, die Größe der gefundenen Treffermenge und ihre Beschaffenheit zu verändern. Dies ist beispielsweise möglich, in dem ähnliche Dokumente auf Basis des gefundenen Treffers ermittelt werden, weitere relevante Dokumente durch Fokussierung der Suchanfrage auf Tehmenblöcke gefunden werden oder dem Nutzer Möglichkeiten zur Justierung des Rankings der Treffer angeboten werden. (vgl LEWANDOWSKI 2005, S. 151)

3.1 FOX ET AL. 2005

FOX ET AL. 2005 gingen 2005 der Frage nach, inwieweit explizites Feedback und implizit gemessenes Feedback übereinstimmen. Dabei wollten sie in einer zweiten Frage auch evaluieren, welche implizierten Daten am meisten über die Nutzerzufriedenheit aussagen.

3.1.1 Methodik der Studie

Für die Erhebung der Daten entwickelten FOX ET AL. 2005 ein Browser Add-In für den Internet Explorer, der das Benutzerverhalten aufzeichnete und die Ergebnisse an eine SQL-Datenbank übermittelte. Neben der Erfassung des implizierten Feedbacks wurden auch explizite Urteile eingeholt. Zum einen bezogen sich diese auf die jeweils geklickten Treffer und zum anderen auf die gesamte Suchsession.

Die Daten wurden von 146 internen Mitarbeitern bei Microsoft erhoben. Dabei wurde kein systematisches Experiment durchgeführt. Es wurden einfach alle Daten der Teilnehmer über 6 Wochen gesammelt. Insgesamt wurden so 2560 Suchsessions und 3659 geklickte Treffer mit implizierten und expliziten Feedback Daten erhoben. Die implizierten Maße wurden dabei auf Trefferebene und Sessionebene unterschieden. Zu den Maße auf Trefferebene gehörten beispielsweise die Zeit, die aufgebracht wurde, um ein Ergebnisdokument zu lesen, die Position in der Trefferliste oder ob Nutzer die Seite zu ihren Favoriten hinzufügten oder ausdruckten. Auf Sessionebene wurden unter anderem die Anzahl der Suchanfragen, die geklickten Treffer oder auch der Durchschnitt gebookmartker oder gedruckter Suchergebnisse ermittelt.

3.1.2 Ergebnisse der Studie

Damit eine Verbindung zwischen den implizierten und expliziten Daten hergestellt werden konnte, nutzen FOX ET AL. 2005 ein Bayesian Model, das alle Abhängigkeiten zwischen den erhobenen Maßen und abhängigen Variablen (in diesem Fall explizite Urteile zur Zufriedenheit und Relevanz mit den Daten zu den einzelnen geklickten Treffern und Suchsessions) darstellen kann. Sie entwickelten daraus zwei Ansätze, um Zufriedenheit auf Trefferebene und auf Sessions ebene voraussagen zu können.Weiterhin wurde noch eine Analyse zur Auswertung von Mustern in einer Suchsession durchgeführt.

Insgesamt zeigte sich bei den Nutzern, eine Zufriedenheit von 39%, wenn diese einen Treffer geklickt haben. Das Bayesian Model konnte in der Studie zu 70% voraussagen, wann ein Nutzer zufrieden mit dem Ergebnis war, zu 47% vorhersagen, wann der Nutzer teilweise zufrieden war und zu 52% wenn dieser nicht zufrieden war. Auf der Ebene der gesamten Suchsession wurde ein Zufriedenheit von 74%, eine Teilzufriedenheit von 57% und eine Vorhesage über eine Unzufriedenheit zu 62% berechnet. Bei der Erhebung der expliziten Zufriedenheit gaben 57% der Beteiligten an, dass sie zufrieden mit der gesamten Session waren.

Auf Trefferebene zeigte sich, dass die Maße der Zeit und der exit type, hier die Art wie der Nutzer das Ergebnis verlässt (Browser schließen, zurück navigieren, neue

Suchanfrage starten, anderes Ergebnis aufrufen) am besten für eine Vorhersage geeignet sind. Bei der Suchsession eigneten sich die Anzahl der gesichteten Treffer und Webseiten sowie die Art der Terminierung der Session (Eingabe einer neuen URL, schließen des Browsers) als Maße am besten für die Vorhersagen.

Bei der Analyse der Suchmuster zeigte sich die höchste Zufriedenheit bei dem Muster: Suchsession starten, Suchanfrage abschicken, Suchergebnisseite sichten, ein Suchergebnis anklicken, Suchession beenden. Die Zufriedenheitsrate lag in der Studie hier bei 81%. In den Ergebnissen zeigte sich auch, dass bei mehrfacher Veränderung der Anfragen die höchste Unzufriedenheit aufkam.

3.2 JOACHIMS ET AL. 2005

JOACHIMS ET AL. 2005 führten 2005 Untersuchungen durch, inwieweit impliziertes Feedback, in diesem Fall Klickdaten auf Suchergebnisseiten, für die Relevanzbewertung von Suchtreffern nutzbar ist. Es wurden ausschließlich Suchergebnisseiten von Google getestet.

Insgesamt führten JOACHIMS ET AL. 2005 zwei konsekutive Studien durch, um mehr über das Verhalten und den Interaktionen von Nutzern mit Suchergebnisseiten zu erforschen. Dabei sollten neben einer Aufzeichnung des Nutzerverhaltens auch Erkenntnisse über Entscheidungsprozesse von Benutzern gewonnen werden.

3.2.1 Methodik der Studie

Insgesamt wurden 10 Suchaufgaben (5 navigationsorientierte und 5 informationsoritentierte Aufgaben) bearbeitet. Dabei waren die Testpersonen angehalten Google zu benutzen. Bei der Wahl der Suchanfragen wurden die Probanden nicht eingeschränkt.

Im ersten Teil der Untersuchung wurden 34 Studenten aus verschiedenen Fachrichtungen getestet und das Verhalten sowie alle Interaktionen mit den Ergebnissen, auch mit Eyetracking, aufgezeichnet. Eine zweite Untersuchung wurde mit 22 Studenten durchgeführt. Im Unterschied zu der ersten Gruppe, waren die Suchergebnisse bei einem Teil der Probanden manipuliert. Dabei waren entweder die beiden Topergebnisse vertauscht oder die gesamte Reihenfolge der Ergebnisliste lag in umgekehrter Reihenfolge vor.

Die Daten für die Auswertung wurden über Eyetracking und über einen Proxy gewonnen, der alle Klickinteraktionen aufzeichnete. Für einen Vergleich der Klickdaten mit expliziten Relevanzbewertungen wurden in einem dritten Teil die gesichteten Ergebnisse durch 5 Juroren noch einmal beurteilt.

3.2.2 Ergebnisse der Studie

Die Ergebnisse von JOACHIMS ET AL. 2005 teilen sich in Ergebnisse zum Nutzerverhalten und zu der Frage der Relevanzbewertung über die Klickdaten. Bei dem Nutzerverhalten zeigten sich heute bekannte Daten zum Verhalten von Suchmaschinenbenutzern. Die meiste Aufmerksamkeit richteten die Studenten auf die ersten beiden Ergebnisse, ab dem dritten Ergebnis nahm die Aufmerksamkeit ab und verschlechterte sich weiter zu den Ergebnissen hin, die nur durch Scrollen erreichbar waren. Weiter zeigte sich, dass die Studenten mehr Zeit aufwendeten Textbeschreibungen anzusehen, die sich in der Nähe einen geklickten Treffers befanden als weiter entferntere Beschreibungen.

In Bezug auf die implizierte Relevanzbewertung durch Klickdaten führen JOACHIMS ET AL. 2005 zwei Probleme an. Zum einen gibt es eine Tendenz der Nutzer dem Ranking der Suchmaschine zu trauen und damit eine automatische Bevorzugung der ersten Ergebnisse zu geben und zum anderen besteht eine Beeinflussung des Klickverhaltens durch die Qualität des Rankings. Sie meinen dadurch lassen sich Klickdaten schwer für eine Relevanzmessung verwenden. Durch die angesprochenen Effekte ist es schwierig direkte Relevanzurteile aus den Klickdaten zu generieren, da Markeneffekte und Qualität der Retrievalfunktionen des Systems bekannt sein müssten. Sie schlagen daher vor, Klicks nicht als implizites Feedback sondern als Präferenzen zu sehen und neben geklickten Treffern, auch nicht geklickte Treffer einzubeziehen.

Insgesamt kommt die Studie zu dem Ergebnis, das ein nachvollziehbarer Zusammenhang zwischen implizierten Feedback durch Klicks und expliziten Beurteilungen durch Juroren besteht, auch wenn dieses insgesamt etwas niedriger als bei einer expliziten Relevanzbewertung eines Dokuments durch zwei Juroren ausfällt.

3.3 HUFFMAN / HOCHSTER 2007

Huffman und Hochster sind in ihrer Studie der Frage nachgegangen, inwieweit die Relevanz von Suchergebnissen Einfluss auf die Nutzerzufriedenheit hat. Beide waren Mitarbeiter von Google als sie die Studie durchführten.

3.3.1 Methodik der Studie

Für die Studie wurden 200 zufällige Suchanfragen von Google.com aus der Mitte des Jahres 2006 verwendet und an Google geschickt. Die Top 3 Ergebnisse zu jeder Suchanfrage wurden dann insgesamt von 7 Juroren bewertet. Dafür wurde eine abgestufte Skala verwendet. Ferner wurden die Anfragen auch mit Hilfe von 9 Bewertern kategorisiert. Dabei ging es darum, ob die Anfrage falsch geschrieben war, die Anfrage entweder navigationsorientiert, transaktionsorientiert, informationsorientiert sind, die Themen speziell oder breit sind und weiteres. Die Kategorisierung wurde nach einem Mehrheitsvotum durchgeführt. Eine weitere Erweiterung der Suchanfragen fand durch eine Einschätzung der Informationsbedürfnisse hinter den Anfragen statt. Da die bloße Suchanfrage keine Informationen zu den Absichten und Zielen liefert. Damit auch die Mehrdeutigkeit von Anfragen berücksichtigt wird, sollten jeweils ein primäres und ein sekundären Informationsbedürfnis angegeben werden. Anschließend wurden die häufigsten Angaben ausgewertet und sprachlich normalisiert.

In einer ersten Untersuchung wurden nun verschiedenen Gruppen von Personen die Suchanfragen mit den definierten Suchanfragen vorgelegt. Diese sollten sich in die Person mit dem Informationsbedürfnis hineinversetzen und mit Google die Suche in einer Session starten. Die Sessions sollte beendet werden, wenn die Testperson das Gefühl hat, dass das Informationsbedürfnis befriedigt ist oder dass die Person mit dem ursprünglichen Informationsbedürfnis die Suche abbrechen würde. Alle Interaktionen mit dem Webbrowser wurden aufgezeichnet. Am Ende einer Suchsession sollten die Probanden ihre Zufriedenheit einschätzen und Kommentare zu ihren Erfahrungen abgeben. Daraus kalkulierten Huffman und Hochster Werte zwischen 0 und 1. Sie nahmen bei Suchanfragen mit zwei Informationsbedürfnissen Gewichtungen vor, in dem das primäre Informationsbedürfnis mit 0,7 und das sekundäre Informationsbedürfnis mit 0,3 in die Wertung einging.

Anschließend verglichen Huffmann und Hochster die Relevanz der jeweils originären Suchanfrage mit der Zufriedenheit, die sich aus der gesamten Suchsession ergab. Für die Relevanz wurde dafür eine einfache Formel entwickelt:

$$Relevance(q) = \frac{(pos1 + pos2/2 + pos3/3)}{1 + \frac{1}{2} + \frac{1}{3}}$$

Es wurde mit dieser Formel ein gewichteter Durchschnitt einer Anfrage mit den Top-3 Ergebnissen ermittelt. Dabei wurden nur die Ergebnisse der Ausgangsanfrage auf Relevanz beurteilt, nicht die weiteren Anfragen, die während einer Suchsession entstanden.

3.3.2 Ergebnisse der Studie

Insgesamt zeigte sich bei den Messungen eine hoher Zusammenhang zwischen Nutzerzufriedenheit und Relevanz, wobei besonders eine große Zufriedenheit bei hoch relevanten Dokumenten festgestellt wurde. Der Zusammenhang ist auch nicht so stark ausgeprägt bei falsch geschriebenen Suchanfragen, was sich darauf zurückführen lässt, dass Nutzer während einer Sessions auf eine korrigierte Suchanfrage von Google klickten und so relevante Ergebnisse zu der Anfrage finden konnten.

Ferner zeigen Huffman und Hochster auch, dass bei navigationsorientierten Anfragen die Beziehung zwischen Relevanz und Zufriedenheit stark nach der ersten Trefferposition nachlässt, während sie bei nicht navigationsorientierten Anfragen auch bei den Treffern auf dem zweiten und dritten Platz bestehen bleibt.

Insgesamt wurden die Ergebnisse genutzt, um ein Modell zu entwickeln, dass bei der Vorhersage der Zufriedenheit des Nutzers helfen soll. Zum Schluss kommen sie zu einem linearen Modell, dass sich folgendermaßen zusammensetzt:

(Pos1+Pos2+Pos3)*Nav+NumEvents

Dabei werden die Relevanzbewertungen der ersten Positionen, die Art der Anfrage (handelt es sich um eine navigationsorientierte Anfrage oder nicht) und die Nummer der aufgezeichneten Ergebnisse während einer Suchessions berücksichtigt. Insgesamt konnten Huffman und Hochster mit ihrem Ansatz einen starken Zusammenhang zwischen Relevanz der ausgegebenen Suchtreffer zur Anfrage und der Zufriedenheit der ganzen Suchsession aufzeigen.

3.4 Würdigung der Studien

Die oben gezeigten Studien wurden betrachtet, um praktische Untersuchungen von protokollierten Suchessions und damit verbundenen Untersuchungen zum Nutzerverhalten sowie der Verwendung als Relevanzindikatoren zu erhalten. Keine der genannten Studien wurde genutzt, um die Qualität von Suchsystemen vergleichend zu messen. Es ging immer mehr um die Frage, inwieweit sich beispielsweise die protokollierten Klicks zur Bestimmung der Nutzerzufriedenheit eignen bzw. zu untersuchen welcher Zusammenhang zwischen Relevanz und Nutzerzufriedenheit besteht.

In der Studie von Fox et al. 2005 kam dabei sogar ein Softwaresystem zum Einsatz, das in Ansätzen vergleichbar mit dem Prototyp ist, der innerhalb dieser Arbeit entwickelt wird. Auf der einen Seite wurden Nutzerinteraktionen aufgezeichnet und dazu aber explizite Urteile zur Zufriedenheit und Relevanz eingeholt. Der Vorteil dieser Studie ergab sich daraus, dass die Nutzer nicht in einer künstlich geschaffenen Umgebung mit definierten Arbeitsaufgaben getestet wurden. Der Nachteil ergibt sich hier aber aus einer Vergleichbarkeit oder Wiederholung der Studie. Interessanterweise konnte die Berechnung der Zufriedenheit hier aber besonders auf eine bestimmte Anzahl von Indikatoren reduziert werden.

In den Studien von JOACHIMS ET AL. 2005 und HUFFMAN / HOCHSTER 2007 wurden Zusammenhänge zwischen der Relevanz und den implizierten Feedback durch erhobene Klickdaten sichtbar. Beide führen aber verschiedene Bereiche an, die die Aussagekraft der aufgezeichneten Protokolle einschränken. So zeigen Joachims et al. 2005 durch die zusätzliche Aufzeichnung mit Eye Tracking Geräten, dass bei der Sichtung der Suchergebnisliste oft schon das Ranking den Nutzer beeinflusst auch die Markeneffekte sind besonders bei freien Suchsessions ein Faktor, der die Ergebnisse verzerrt. HUFFMAN / HOCHSTER 2007 zeigen weiterhin, dass der Anfragetyp zu berücksichtigen ist.

4. Modell für die Zusammenführung explorativer Suchen und Retrievaltests

Dieser Abschnitt beschäftigt sich mit einigen exemplarischen Modellen, die in der Information Retrieval-Evaluierung eingesetzt werden, um Suchsysteme miteinander zu vergleichen. Dabei wird auch ein Gegensatz aufgezeigt, der sich aus den jeweiligen Evaluierungsschwerpunkten (System oder Nutzer) ergibt.

4.1 Frameworks und Modelle

Bei den folgenden Modellen werden Frameworks zur Evaluierung der Qualität von Suchmaschinen dargestellt und in Bezug auf den Nutzen für die Verbindung von Suchsessions und Relevanztests diskutiert.

4.1.1 Das Cranfield Modell

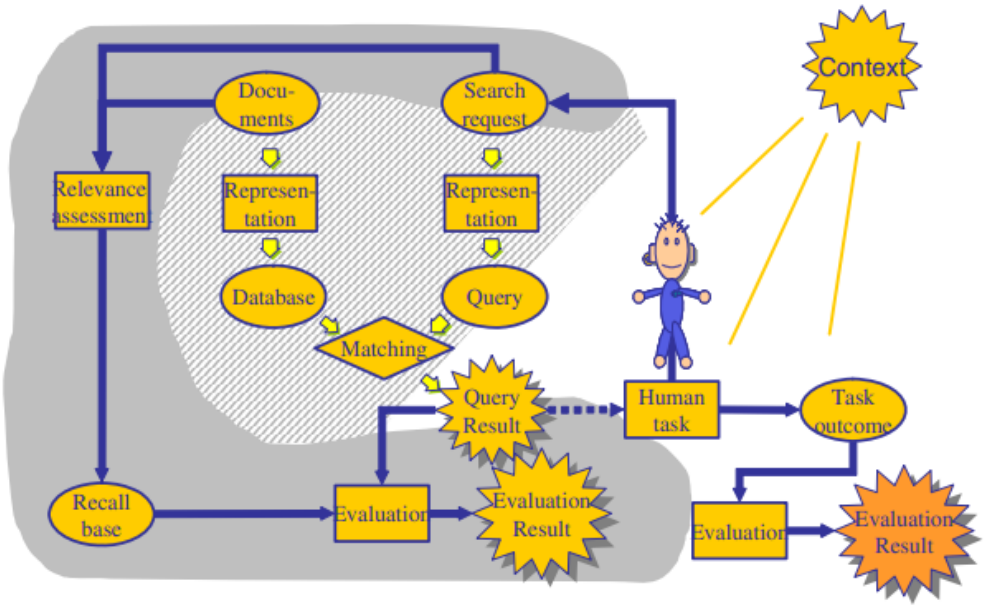
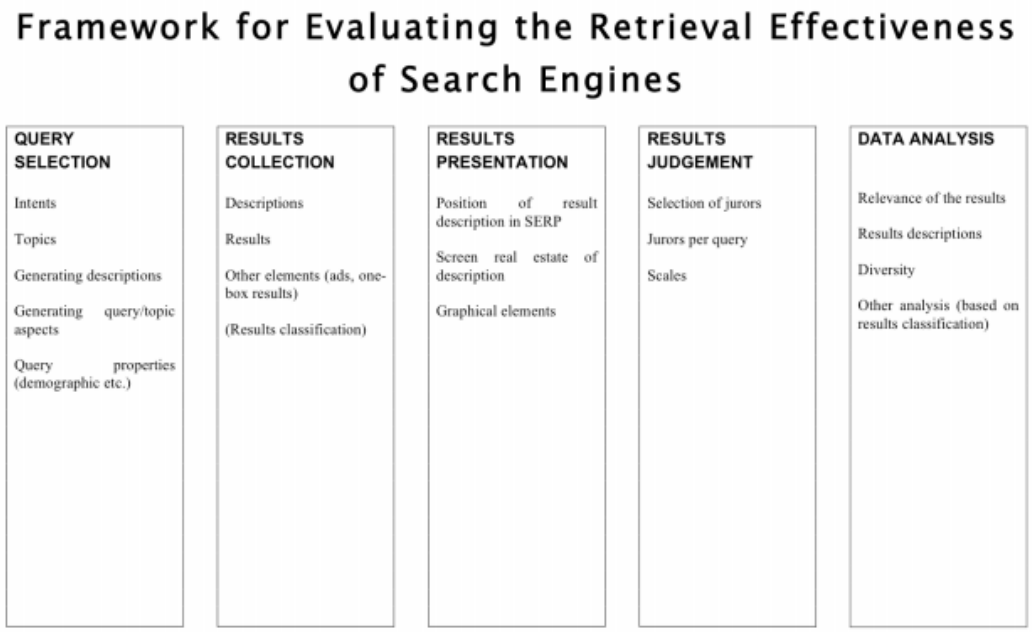


Abbildung 5: Cranfield Model (Quelle: JÄRVELIN 2009, S. 291)

Das Cranfield Modell ist der Klassiker unter den Evaluierungsmodellen von Information-Retrieval-Systemen. Die ersten Cranfield Tests wurden 1953 durchgeführt und bildeten die Grundlage für die heutige Gestaltung von Relevanztests (vgl. GÖDERT ET AL. 2012, S. 333ff.). Der Vorteil dieses Modells liegt darin, dass vergleichbare Ergebnisse produziert werden. Es steht vor allem die Retrievaleffektivität von Systemen im Vordergrund, die mit den Standardmaßen Recall und Precision (oder Abwandlungen dieser Kennzahl) erfasst werden. Die tatsächlichen Interaktionen des

Benutzers mit dem System oder auch die Usability und Funktionalitäten der einzelnen Komponenten der Suchmaschine werden nicht berücksichtigt. Auch TREC basiert weitestgehend auf diesem Modell.

4.1.2 Lewandowski: A Framework for Evaluating the Retrieval Effectiveness of Search Engines



Das Framework von LEWANDOWSKI 2012 dient der Evaluierung der Retrievaleffektivität von Suchmaschinen und ist insgesamt in 5 Teile gegliedert. Im ersten Teil werden die zu evaluierenden Suchanfrage ausgewählt. Dabei sollten die Anfragen nach Borders Anfragetypen gegliedert und breite Themen gewählt werden. Die Anfragetypen sollten dabei nicht im Test vermischt werden.Als Quelle für die Anfragen könnten dabei Logfiles genutzt werden. Im Idealfall sollte auch der Nutzer die Ergebnisse bewerten, die er durch seine Anfrage gefunden hat, da dieser sein Informationsbedürfnis am besten kennt. Da dies in den meisten Testszenarien nicht der Fall ist, müssen die Informationsbedürfnisse entweder durch Suchende direkt abgefragt werden oder bei der Nutzung von Logfiles von anderen Personen eingeschätzt werden (wie beispielsweise bei HUFFMAN / HOCHSTER 2007). Daneben sollten auch die Aspekte der Anfrage berücksichtigt werden. So kann beispielsweise eine Anfrage wie „James Bond“ Aspekte zu den Filmen, Büchern, Schauspielern und so weiter enthalten. Dazu kommen noch weitere Eigenschaften der Suchanfrage, die berücksichtigt werden sollten.

Der zweite Teil des Frameworks bezieht sich auf die Zusammenstellung der zu testenden Ergebnisse und auf weitere Informationen, die Suchmaschinen zu den Ergebnissen angeben. Die Informationen beziehen sich dabei auf die Trefferbeschreibung, Informationen zum Typ des Treffers (Trefferbeschreibung, die zu einem organischen Treffer führt oder direkte Antworten), die Position oder auch den Platz, den der Treffertyp auf dem Bildschirm einnimmt. Trefferbeschreibungen sind dabei ein wichtiges Element, weil diese dem Nutzer dabei helfen zu entscheiden, ob er ein Ergebnis anklickt oder nicht. Trefferbeschreibungen sollten daher bei der Evaluierung einbezogen werden. Bei der Erfassung der Suchergebnisse ist es entscheidend für die spätere Auswertung die originären Trefferpositionen zu erfassen. Wenn die Suchergebnisse nicht immer gleich in der Ergebnisliste dargestellt werden, sowie es beispielsweise im Universal Search der Fall ist, sollten auch Ergebnisse aus den Universal Search Kollektionen (Bilder, Videos usw.) sowie Anzeigen mit in die Trefferzusammenstellung für den Test gehen. Ferner ist auch noch zu berücksichtigen, dass aktuelle Eye-Tracking- und Klickdaten-Studien (siehe u. a. Eye Tracking Studien) zeigen, dass Suchergebnisse von Nutzern nicht immer gleichwertig wahrgenommen werden. So werden die Anzeigen auf der rechten Seite einer Ergebnisseite häufig anders wahrgenommen als die organischen Treffer auf der linken Seite und grafische Elemente in den Trefferlisten haben ebenfalls Einfluss auf die Wahrnehmung.

Der vierte Teil des Frameworks gibt Hinweise für die Rekrutierung der Juroren sowie zur Gestaltung der Skalen für die Relevanzbewertung. Juroren sollten nach Anwendungsfall bzw. Spezialisierung der Suchmaschinen und der Themen in den Suchanfragen ausgewählt werden. Direkte Empfehlungen für die Anzahl der Beurteiler für eine Suchanfrage können nicht gemacht werden. Es hat sich aber gezeigt, dass eine hohe Anzahl Juroren keine stark abweichenden Ergebnisse bei der Bewertung der Relevanz von Dokumenten erzeugt. Bei den Skalen für die Bewertung sollten binäre und abgestufte Skalen verwendet werden.

Im letzten Teil werden die Daten ausgewertet. Dabei werden dann häufig erprobte Retrievalmaße (siehe dazu Kapitel 2.1.4 und Kapitel 2.1.5) verwendet. In diesem Schritt wird auch noch einmal auf die Relevanz von Trefferbeschreibungen eingegangen. Dafür stellt LEWANDOWSKI 2012 vier Mögliche Verbindungen zwischen der Trefferbeschreibung und dem dahinter liegenden Ergebnis dar:

- relevante Trefferbeschreibung → relevanter Treffer: Der Idealfall, der bei jedem Ergebnis zutreffen sollte.
- relevante Trefferbeschreibung → nicht relevanter Treffer: Der Nutzer klickt einen Treffer an, weil die Beschreibung relevant ist und wird zu einem nicht relevanten Treffer geführt. Meistens kehrt der Nutzer auf die Suchergebnisseite mit einem bestimmten Grad an Enttäuschung zurück.

- nicht relevante Trefferbeschreibung – > nicht relevanter Treffer: Der Nutzer klickt aufgrund der nicht relevanten Beschreibung den Treffer nicht an. In diesem Sinne ist die Beschreibung hier hilfreich, wobei diese aber nicht unter den Top Ergebnissen auftauchen sollte.
- nicht relevante Trefferbeschreibung → relevanter Treffer: Die schlechte Trefferbeschreibung sorgt dafür, dass der Nutzer diesen relevanten Treffer verpasst.

Zur Messung der Konsistenz bei den Trefferbeschreibungen und den Suchergebnissen gibt Lewandowski noch vier verschiedene Kennzahlen an:

Maßzahl	**Beschreibung**
Decription-result precision	Misst das Verhältnis aller Ergebnisse bei denen Beschreibung und Treffer relevanz bewertet wurden.
Description-result conformance	Erfasst die Menge der Paare bei denen Treffer und Beschreibung gleich bewertet wurden (relevanz oder nicht-relevant).
Description fallout	Misst das Verhältnis zwischen irrelevanten aber relevanten Treffern. Dadurch wir die Anzahl der Dokumente bestimmt, die durch eine schlechte Trefferbeschreibung verpasst wurden.
Description deception	Über diese Kennzahl wird bestimmt, wie häufig relevante Trefferbeschreibungen zu nicht relevanten Dokumenten führten.

Weiterhin sind auch Analysen denkbar, bei denen Suchergebnisse zunächst nach Typ klassifiziert werden (z. B. Nachrichten, Blogeinträge) um beispielsweise die Nützlichkeit der verschiedenen Dokumenttypen gegenüberzustellen.

Lewandowski gibt mit dem Framework einen ausführliches Modell zur Durchführung von Retrievaltests. Ein Problem dieses Ansatzes liegt darin, dass tatsächliches Nutzerverhalten nicht berücksichtigt wird, dafür gibt dieses Framework aber viele wichtige Hinweise, die bei der Gestaltung von Tests zur Retrievaleffektivität berücksichtigt werden müssen.

4.1.3 Borlund: The IIR Evaluation model: A Framework for Evaluation of Interactive Information Retrieval Systems

BORLUND 2003 entwickelte mit ihrem Modell einen Ansatz zur Evaluierung von interaktive Information-Retrieval-Systemen als Alternative zum systemorientierten Modell von Cranfield. Das Framework soll dabei helfen interaktive Retrieval Daten zu sammeln und zu analysieren.

Insgesamt setzt sich das Modell aus drei Teilen zusammen:

1. Im ersten Teil stellt Borlund Bestandteile vor, die für ein funktionelles, valides und realistisches Umfeld für die Untersuchung des Informationssystems sorgen.
2. Der zweite Teil beschäftigt sich mit empirisch belegten Vorschläge für die Nutzung simulierten Arbeitsaufgaben gemacht.
3. Im letzten Teil werden alternative Kennzahlen dargestellt, die zur Perfomanzmessung der Systeme eingesetzt werden können.

Teil 1: Komponenten der experimentellen Umgebung

Borlund gibt drei Komponenten an, die für den ersten Teil notwendig sind und stark miteinander verknüpft sind:

- Der Test sollte mit potenziellen Nutzern des Systems erfolgen.
- Es sollen Interpretationen von individuellen und dynamischen Informationsbedürfnisse angewendet werden.
- Die Beurteilung der Relevanz soll multidimensional und dynamisch erfolgen.

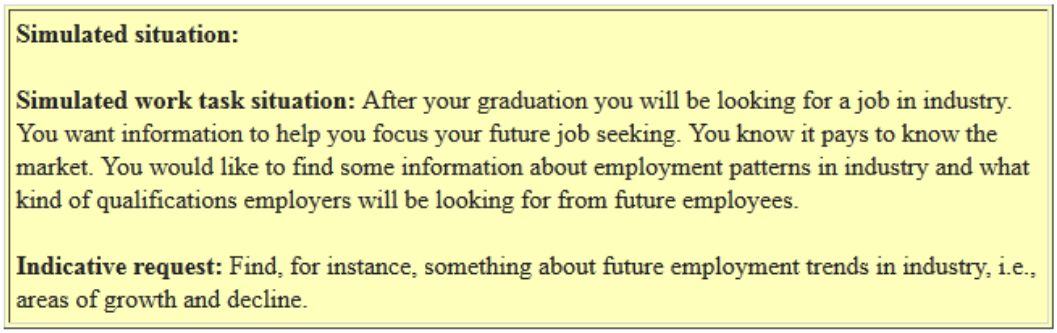
Durch diese Komponenten fallen bei der interaktiven Nutzung mit dem System Daten an, die sowohl für systemorientierte Evaluierungen als auch benutzerorientierten Evaluierung genutzt werden können.

Damit echte Informationsbedürfnisse den Ausgangspunkt der Untersuchung bilden, schlägt BORLUND 2003 vor, die Tests mit simulierten Arbeitsaufgaben durchzuführen. Mit Hilfe von Arbeitsaufgaben werden sowohl realistische als auch kontrollierte Voraussetzungen für die Nutzertests geschaffen.In den Arbeitsaufgaben werden kurze „cover stories“ erzählt, die zur individuellen Nutzung eines Suchdienstes führen. In der „cover story“ wird dabei eine offene Beschreibung des Suchkontextes oder Suchszenarios gegeben.

Mit der Arbeitsaufgabe erhält die Testperson Informationen und Hilfe zu:

- Dem zugrundeliegenden Informationsbedürfnis für das Szenario.
- Dem Umfeld der Situation.
- Dem Problem, das gelöst werden soll.
- Dem Ziel, das mit der Suche erreicht werden soll.

Abb. Zeigt ein Beispiel für eine simulierte Arbeitsaufgabe.



Simulated situation:

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.

Abbildung 6: Beispiel für eine simulierte Arbeitsaufgabe (Quelle: BORLUND 2003)

In der simulierten Situation werden der Zweck und das Ziel der Suche angegeben. Mit dem indicative request werden der Testperson weitere Hinweise gegeben, z. B. wonach gesucht werden könnte. Borlund weißt in diesem Zusammenhang darauf hin, dass die zusätzlichen Hinweise das Suchverhalten des Nutzers und seine formulierten Suchanfragen nicht beeinflussen und das es damit freigestellt ist, diese einzusetzen oder nicht.

Teil 2: Empfehlungen beim Einsatz von simulierten Arbeitsaufgaben

Ergebnisse aus empirischen Untersuchungen zeigen, dass es keine statistischen Unterschiede zwischen realen Informationsbedürfnissen und simulierten Informationsbedürfnisses gibt (vgl. BORLUND 2003). Damit zeigt sich, dass Evaluierungen alleine mit simulierten Arbeitsaufgaben durchgeführt werden können. Es steht dem Testleiter aber auch frei beides innerhalb der Untersuchung zu berücksichtigen. Also eine Mischung aus echten Informationsbedürfnissen und simulierten Informationsbedürfnissen zu testen. Borlund schlägt vor, dass die Arbeitsaufgaben an die Nutzergruppe und an die Informationsumgebung angepasst werden sollten. Wird beispielsweise ein System für medizinische Informationen getestet, sollten sich die Arbeitsaufgaben auch mit medizinischen Themen beschäftigen und nicht allgemeiner Natur sein.

Damit Lerneffekte ausgeschlossen werden können, sollten innerhalb von Laborexperimenten die Arbeitsaufgaben den Testpersonen in verschiedener Reihenfolge vorgelegt werden. Weiterhin sollten alle Probanden die Aufgaben auf verschiedenen Systemen durchführen, wenn ein Vergleich zwischen den Informationssystemen stattfinden soll.

Neben diesen Empfehlungen schlägt BORLUND 2003 noch vor bei der Untersuchung des Nutzerverhaltens nach jeder bearbeiteten Aufgaben kontrollierte Interviews oder Online-Fragebögen zu nutzen, um explizite Informationen zum Suchverhalten zu erheben. Daneben erwähnt sie auch, dass Interviews für die Gestaltung von Arbeitsaufgaben durchgeführt werden könnten. Personen, aus der zu testenden Nutzergruppe könnten nach ihren Informationsbedürfnissen befragt, aber auch in Pilotstudien beobachten werden, damit Hintergründe für Arbeitsaufgaben generiert und echte Informationsbedürfnisse gesammelt werden.

Teil 3: Kennzahlen zur Performanzmessung der Suchsysteme

Borlund stellt heraus, dass sich die Standardmaße Recall und Precision aus der systemorientieren Evaluierung im Kontext mit benutzerzentrierten Untersuchungen nicht nutzen lassen. Recall und Precision beziehen sich auf die binäre Bewertung der Relevanz und berücksichtigen damit nicht verschiedene Arten von Relevanz. Die folgende Tabelle zeigt die Kennzahlen, die Borlund für die Auswertung der Ergebnisse diskutiert:

Maßzahl	Beschreibung
The RR Measure	Bei dem RR Measure werden verschiedene Arten von Relevanz berücksichtigt. Es wird gemessen inwieweit diese Arten (R_1 und R_2) übereinstimmen. Dabei könnte eine Art beispielsweise die algorithmische Relevanz des Systems und die andere Art die subjektive Relevanzbewertung des Nutzers (Pertinenz, Situationsrelevanz) sein. $\textbf{cosine: association } (R_1,R_2) = \left\| \frac{\sum(R_1R_2)}{(\sum R_1^2)^{1/2} * (\sum R_2^2)^{1/2}} \right\|$
The RHL Indicator	Der RHL Indicator bezieht zum einen das algorithmische Ranking der Ergebnisse und zum anderen subjektive Bewertungen von Juroren mit ein. Dabei nimmt die objektive Relevanz eines Dokuments mit der Position stetig ab.

	$$M_g = L_m + \left(\frac{n/2 - \sum f2}{F(med)} \times CI \right)$$ where: L_m = lower real limit of the median class, i.e., the lowest positioned information objects above the median class; n = number of observations, i.e., the total frequency of the assigned relevance values; $f2$ = cumulative frequency (relevance values) up to and including the class preceding the median class; $F(med)$ = the frequency (relevance value) of the median class; and CI = class interval (upper real limit minus lower real limit), commonly in IR = 1. Je niedriger der RHL Wert, desto höher ist die Rankingposition des Dokuments, desto höher ist die Relevanz.
Cumulated Gain (CG)	Mit dem CG wird die kumulierte Zunahme der Relevanz berechnet. Dabei wird auch wie beim RHL Indicator die Position des Treffers berücksichtigt. Diese Kennzahl wurde geschaffen, um nicht binäre Skalen zu berücksichtigen. Darüber hinaus sol damit auch von vornherein ausgedrückt werden, das besser gerankte Treffer wertvoller für den Nutzer als schlechter gerankte Dokumente sind.
Cumulated Gain with discount	Der DCG ist eine Abwandlung des CG, der auch Faktoren wie das Nutzerverhalten einbeziehen kann.

Tabelle 4: Übersicht zu vorgeschlagenen Kennzahlen für die Evaluierung im Interactive Information Retrieval (Quelle: BORLUND 2003)

Zusammenfassend betrachtet zeigt BORLUND 2003 mit ihrem Modell mehr Ansätze den Nutzer stärker in die Evaluierung mit einzubinden und gleichzeitig werden Kennzahlen genannt, die trotzdem die systemorientierte Evaluation berücksichtigen. Besonders sind dabei die Hinweise auf die Verwendung von simulierten Arbeitsaufgaben und Szenarien, die für kontrollierte Evaluationen nutzbar sind. Innerhalb ihres Modells geht Borlund nicht auf die Gestaltung des Experiments im Detail ein. So weißt sie nur daraufhin, dass für die Arbeitsaufgaben Personen aus der Zielgruppe interviewt oder in Pilotstudien getestet werden sollten, um Informationsbedürfnisse und Szenarien für Arbeitsaufgaben zu sammeln auch die Gestaltung der Skalen für die Relevanzbewertung werden nicht näher betrachtet. Sie geht auch auf Methoden zur Datenerhebung ein, nennt aber nur Beispiele wie Videoaufzeichnungen oder Protokollierung in elektronischer Form und gibt keine weiteren Hinweise z. B. zu Vor- oder Nachteilen der Methoden.

4.1.4 BELKIN ET AL. 2009: A model for Evaluation of Interactive Information Retrieval

Einen weiteren Vorschlag für ein Modell zur Evaluierung von Interactive Information Retrieval machen Belkin, Cole und Liu. Ihr vorgeschlagener Ansatz basiert auf dem Konzept des „Information Seeking“ und bezieht sich vor allem auf die „usefulness“

Zunächst erläutern sie das Konzept des „Information Seeking“. Menschen haben Pläne in ihrem Leben und wollen bestimmte Ziele erreichen. Als Grundlage dafür dient ihr persönliches Wissen. Gibt es eine Lücke in diesem Wissen, muss diese geschlossen werden und um Informationen für die Bewältigung der Situation zu erlangen, lassen sich Suchmaschinen einsetzen. Innerhalb der IR-Systeme finden nun Interaktionen zwischen Benutzer und Informationsobjekten statt und diese Interaktionen wiederum können in eine Sequenz von „information seeking strategies“ (ISS) zerlegt werden. Belkin et al. meinen, ein Informationssystem sollte dahin evaluiert werden, inwieweit es den Nutzer bei seiner Zielerreichung unterstützt.

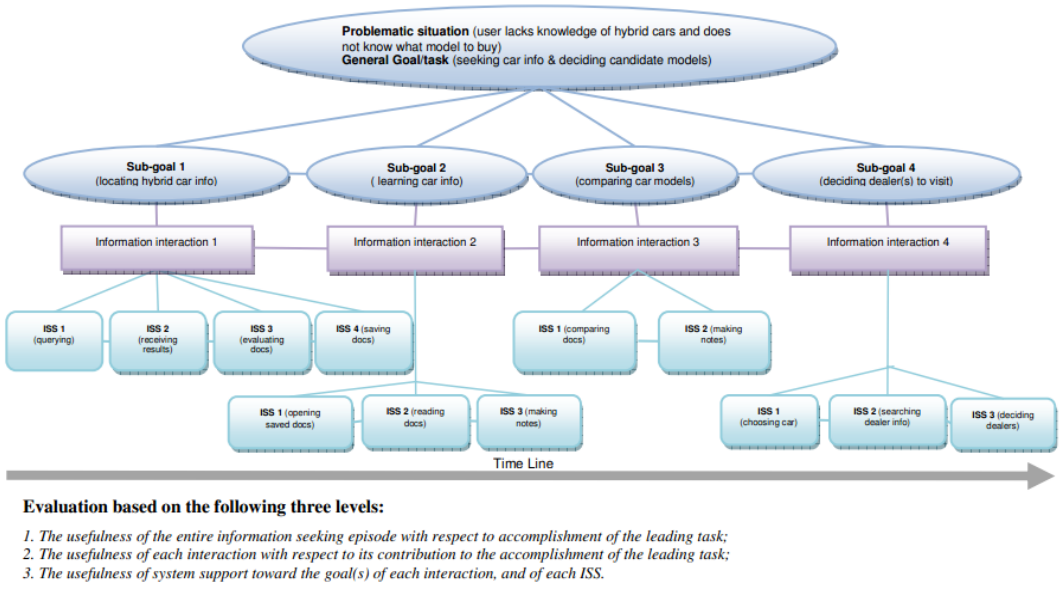


Abbildung 7: Beispielprozesse im Modell zur Evaluierung im Interactive Information Retrieval von Belkin et al. 2009 (Quelle: BELKIN ET AL. 2009)

Abb. 7 zeigt ein Beispiel für ein Informationsproblem mit den Zielen eines Nutzers. Es wird deutlich, dass ein Nutzer mehrere Etappenziele während eines Information Seeking Prozesses verfolgt. BELKIN ET AL. 2009 schlagen eine Evaluierung auf drei Ebenen innerhalb ihres Modells vor:

1. Der ganze Information-Seeking-Prozess sollte bezüglich der Zielerreichung des Nutzers evaluiert werden.
2. Jede Interaktion sollte im Zusammenhang mit der Zielerreichung und Lösung des Problems evaluiert werden.
3. Jede Interaktion und jede ISS sollte in Bezug auf das jeweils einzelne Ziel evaluiert werden.

Als Kriterium für die Evaluierung soll die „usefulness“ dienen, die durch folgende Fragen bestimmt werden soll:

1. Wie nützlich ist der gesamte Information-Seeking-Prozess für die Lösung des Problems?
2. Wie nützlich sind die einzelnen Interaktionen bei der Lösung des Problems?
3. Wie gut konnten die einzelnen Ziele der Interaktionen erreicht werden?

Aus der Sicht des Systems stellen sich folgende Fragen:

1. Wie gut konnte das System den Nutzer bei der Zielerreichung unterstützen?
2. Wie gut ließen sich alle nötigen Interaktionen für die Bearbeitung der Aufgabe innerhalb des Systems durchführen?
3. Wie gut hat das System die Interaktionen unterstützt?

Die Operationalisierung der „usefulness“ soll in diesem Modell auf der Stufe der gesamten Information Retrieval Episode stattfinden; zur Messung des empirischen Zusammenhangs zwischen Interaktion und den Suchergebnissen auf der Stufe der Mitwirkung des Nutzers in Bezug auf die gefundenen Ergebnisse; und auf der Stufe der Ziele zu jeder ISS.

Beispiele auf allen genannten Stufen können gefunden Dokumente auf der Stufe zur Lösung des Problems, Aufgabenbewältigung generell, der Umfang der gesichteten Dokumente auf einer Stufe, die Anzahl nützlicher Dokumente auf Top-Positionen usw. sein. Direkte Kennzahlen werden aber nicht angeben und werden auch als Problem bei der Messung der usefulness genannt.

In diesem Modell werden keine direkten Vorschläge für die Gestaltung einer Evaluation gemacht. Dieses Modell wurde aber an dieser Stelle aufgeführt, um zu zeigen, dass neben der Messung der Relevanz von Dokumenten auch andere Kriterien zur Beurteilung der Qualität eines Suchsystems herangezogen werden können, besonders in Bezug auf den Nutzerkontext.

4.2 Modell für den Prototyp

Der genannte Überblick zu Modellen für die Evaluierung von IR-Systemen machte noch einmal deutlich, dass es viele verschiedene Ansätze gibt, Suchmaschinen zu evaluieren. Das folgende Modell soll nun den Rahmen dafür bilden, wie sich Daten zu explorativen Suchen im Zusammenhang mit klassischen Retrievaltests nutzen lassen könnten. Ziel dieses Modells ist die Verbindung beider Evaluierungsmethoden auch in Bezug auf vergleichende Evaluierungen von Suchdiensten, welche bei vielen genannten Ansätzen besonders im Nutzerkontext schwierig sind. Schwerpunkte des Modells liegen in der Nutzerbeobachtung und der Messung der Qualität der Suchergebnisse, die auf protokollierten Nutzerinteraktionen basieren soll.

Die grundlegenden Methoden innerhalb des Modells sind eine automatische Protokollierung von Nutzerinteraktionen, z. B. Klickdaten, Eingabe von Suchanfragen, Navigation innerhalb des Browsers, manuelle Eingabe von URLs. etc. und klassische Retrievaltests, wie sie im Kapitel 2 beschrieben werden. Ergänzende Methoden wie Eye-Tracking, Videoaufzeichnungen oder andere Verfahren zur Nutzerbeobachtung und Datenerhebung sind zusätzlich zu den genannten grundlegenden Methoden möglich, aber nicht zwingend. Das Prinzip des Modells ist dabei einfach. Benutzer führen simulierte Arbeitsaufgaben in verschiedenen Suchmaschinen durch und werden dabei protokolliert. Anschließend werden diese Daten automatisch analysiert (Klickdaten, Suchanfragen), um Ergebnissets zu generieren, die anschließend von dem Nutzer zusätzlich in Bezug auf die Relevanz oder auch anderen Fragestellungen auf Skalen bewertet werden.

Die folgenden Unterkapitel beschrieben jeweils aufeinanderfolgende Schritte und Komponenten innerhalb des Modells, die notwendig sind, um das oben genannte Prinzip auch in der Praxis umzusetzen. Dabei werden zu jedem Schritt Vor- und Nachteile erläutert.

4.2.1 Auswahl der Suchmaschinen

Am Anfang der Evaluierung kann je nach gesetzten Schwerpunkt und Fragestellung eine Auswahl der Suchdienste und/oder Informationssysteme stehen. Wenn der Schwerpunkt der Untersuchung sich vor allem auf die Nutzerbeobachtung bezieht sollte keine Einschränkung auf bestimmte Systeme erfolgen. Der Nachteil dabei ist aber, dass Vergleiche besonders in Bezug auf die Retrievaleffektivtät damit schwierig durchzuführen sind. Der Testleiter kann vorher nicht wissen, welche Suchdienste von Probanden verwendet werden, wenn diese frei bei der Bearbeitung einer Aufgabe, frei wählbar sind.

Liegt der Untersuchungsschwerpunkt stärker auf einer Evaluierung der Retrievaleffektivität kann die Auswahl der Suchdienste innerhalb des Testszenarios beschränkt werden. Dabei kann auch der Entwickler eines eigenen Systems vergleiche zu anderen Suchdiensten durchführen, wobei hier der Nachteil entsteht, dass der Nutzer durch eine Einschränkung auf bestimmte Dienste an diesem Punkt nicht realistisch abgebildet wird.

Eine Idee diesen Problemen entgegen zu gehen, kann darin bestehen genutzte Suchanfragen in den von Nutzern verwendeten Suchmaschinen für den Retrievaltest automatisch an vorher festgelegt Suchdienste zu senden und diese Ergebnisse explizit durch die Probanden bewerten zu lassen. Der Nachteil dabei ergibt sich aber wiederum dadurch, dass keine Interaktionen mit diesen Suchmaschinen anfallen können. Trotzdem können aufgerufene Dokumente aus den anderen Suchdiensten mit den Ergebnissen der definierten Suchmaschinen abgeglichen werden und dazu auch explizite Relevanzurteile erhoben werden. Dabei ist aber zu beachten, dass der Aufwand einer Testperson immer bei der maximalen Anzahl von zu bewertenden Items zu berücksichtigen ist (siehe dazu Kapitel 2.2.1.2).

4.2.2 Definition der Zielgruppe

Wurde im ersten Schritt entschieden welche Systeme in das Evaluierungssetting eingehen sollen, ist es notwendig die Zielgruppe zu definieren, die innerhalb des Tests untersucht wird. Entscheidungen zur Zielgruppe werden dabei durch folgende Faktoren bestimmt:

1. Die zu untersuchenden Systeme: Welche Systeme sollen in dem Test überprüft werden? Wie bereits erwähnt, ist die Auswahl der Systeme abhängig vom Untersuchungsschwerpunk und von der Fragestellung.
2. Das Nutzerprofil: Welcher Nutzertyp soll innerhalb der Tests abgebildet werden? Werden z. B. allgemeine Websuchmaschinen evaluiert, bezieht sich das Nutzerprofil meistens auf einen Laiennutzer. Das Nuzerprofil verändert sich aber wenn spezielle Suchdienste, Fachdatenbanken usw. getestet werden. In diesem Fall werden in der Regel Experten zu der Domäne benötigt.
3. Den Untersuchungsschwerpunkt
4. und den Untersuchungszeitraum sowie die Umgebung, in der die Daten erhoben werden sollen.

4.2.3 Festlegung des Untersuchungszeitraums und der Untersuchungsumgebung und Auswahl der Testpersonen

Im nächsten Schritt sollten Untersuchungszeitraum und Untersuchungsumgebung festgelegt werden:

- Sollen Langzeitstudien durchgeführt werden?
- Sollen die Probanden und die Informationssysteme in ihrer natürlichen Umgebung oder einer künstlichen Umgebung in Form von Laborexperimenten getestet werden?

Der Vorteil an Erhebungen durch Protokolle liegt darin, dass auch sogenannte Longitudial Studies denkbar sind (vgl. KELLY 2009, S. 25ff.). Der Nutzer könnte sich beispielsweise die Software installieren, die seine Interaktionen aufzeichnet und kann damit über einen längeren Zeitraum beobachtet werden. Ferner findet der Test auch in der natürlichen Umgebung statt, Nachteile, die durch Laborexperimente gegeben sind (vgl. HÖCHSTÖTTER 2009, S. 179), fallen dabei nicht an. Nachteile ergeben sich hier aber bei der Kontrolle der Studie. Arbeitsaufgaben könnten von den Probanden nicht zu Ende geführt werden oder es findet ein kompletter Abbruch der Evaluation statt. Diese Nachteile fallen im Rahmen von Laborexperimenten nicht an. In Laborexperimenten könnten auch neben der Protokollierung der Nutzerdaten und expliziten Relevanzbewertungen weitere Beobachtungen durchgeführt werden.

Bei den genannten Möglichkeiten ist immer der Aufwand zu bedenken, der sich bei der Akquirierung von Probanden und auch bei der Auswertung der Daten einstellt. Daher ist es auch notwendig Überlegungen dazu anzustellen, wie viele Testpersonen an der Evaluierung teilnehmen sollen. Dabei sind die Art des Experiments (natürliche oder kontrollierte Studie im Labor) und Schwerpunkt der Studie zu berücksichtigen. In Anbetracht der Kosten und des Aufwands ist es notwendig Probanden mit großer Sorgfalt auszusuchen. Der Vorteil in diesem Modell besteht aber darin, dass die Bearbeiter der Arbeitsaufgaben auch gleichzeitig als Juroren für die Relevanzbewertungen fungieren. Dadurch fällt eine zusätzliche Gewinnung von Beurteilern weg.

Der vorgestellte Prototyp aus Kapitel 6 soll dabei für beide Anwendungsszenarien einsetzbar sein, wobei auch denkbar ist, den Prototypen im Rahmen von Arbeitsumgebungen, Seminaren im Studium usw. einzusetzen, damit, wie beispielweise bei FOX ET AL. 2005, über längere Zeiträume Daten von tatsächlichen Nutzern mit Informationsbedürfnissen gesammelt werden können. Eine Schwierigkeit würde sich allerdings bei der Vergleichbarkeit und Kontrolle einstellen, da der Einsatz von simulierten Arbeitsaufgaben nur eingeschränkt möglich wäre.

4.2.4 Gestaltung von Pre- und Post-Test Fragebögen

Sind Fragestellung, Zielgruppe und Methodik soweit festgelegt, ist die Ausgestaltung von Fragebögen zur Erhebung von Daten der Testpersonen ein wichtiger Schritt, um diese auf der einen Seite besser charakterisieren zu können und auf der anderen Seite Einschätzungen beispielsweise zur usefulness (vgl. BELKIN ET AL. 2009) einzelner Systeme (bei festgelegten Suchmaschinen für die Tests) zu erhalten. Üblich ist beispielsweise die Erfassung soziodemografischer Daten.

In Bezug auf die Arbeitsaufgaben selber können auch vorher und nachher Einschätzungen erfasst werden. Beispielsweise bei der Einschätzung der Testpersonen zum Schwierigkeitsgrad der Aufgaben. Es lassen sich mit Fragebögen auch Voraussetzungen von Teilnehmern erfassen. Wie in Kapitel 2.3 beschrieben, bringt jede Person eigene Faktoren in die Bearbeitung von Suchaufgaben geprägt durch kognitive Fähigkeiten und das soziale und kulturelle Umfeld mit. Fragebögen oder auch kontrollierte Interviews vor- und nach der Lösung von Suchaufgaben helfen dabei den Nutzer besser einzuschätzen. Mündliche Interviews lassen sich dabei allerdings nur im Rahmen von kontrollierten Laborexperimenten durchführen. Wenn die Studie in einer natürlichen Umgebung, unabhängig von der Anwesenheit, eines Testleiters oder Interviewers, durchgeführt wird, können Online-Fragebögen genutzt werden. Dabei ist im Rahmen von Suchmaschinenuntersuchungen darauf zu achten, dass klare und einfach verständliche Fragen gestellt werden, da ja wie gezeigt, Suchmaschinennutzer eine heterogene Gruppe bilden (vgl. HÖCHSTÖTTER 2009, S. 176 – 177). Fachvokabular sollte daher nur eingesetzt werde, wenn eine dementsprechende Zielgruppe getestet wird.

Neben Pre- und Post-Fragebögen könnten aber auch Zwischenschritte der Benutzer abgefragt werden, wie beispielsweise von FOX ET AL. 2005 realisiert. Dort wurde die Testperson nach jedem geklickten Treffer nach der Relevanz des Treffers befragt. Im Modell von BELKIN ET AL. 2009 sollen Informationen über die Nützlichkeit eines Suchdienstes in verschiedenen Phasen der Suchprozesse ermittelt werden. Ein Problem bei diesen Methoden könnte aber darin bestehen, dass Testpersonen sich überfordert fühlen, wenn diese bei der Durchführung von Recherchen ständig unterbrochen und befragt werden. Eine Idee könnte hier eine Art Online-Notizzettel sein, in dem Probanden ihre Gedanken bei der Bearbeitung von Suchaufgaben eintragen können. Dabei ergibt sich aber wieder die Schwierigkeit einer Vergleichbarkeit der gesammelten qualitativen Daten, da ein solcher Notizzettel keine festen Strukturen aufweisen würde,

4.2.5 Gestaltung von simulierten Arbeitsaufgaben

Wie im Modell von BORLUND 2003 gezeigt wurde, ist die Verwendung von simulierten Arbeitsaufgaben ein probates Mittel, um eine realitätsnahe und kontrollierte Evaluierung von Suchmaschinen durchzuführen. Simulierte Informationsbedürfnisse lassen sich genauso nutzen wie reale Informationsbedürfnisse, da sie statistisch signifikant sehr Nahe bei einander liegen.

Der Vorteil von simulierten vorher definierten Arbeitsaufgaben liegt bei einer Vergleichbarkeit auf der Ebene der Aufgaben, auch führen die festgelegten Aufgaben dazu, Informationsbedürfnisse in Form der Suchanfragen zu sammeln, damit auswertbar zu machen und als eine Grundlage für die Generierung von Ergebnissets zu nutzen, die durch die Testpersonen explizit bewertet werden können. Das Informationsbedürfnis hinter den Anfragen ist durch die Verwendung von Aufgaben festgelegt. Nachteile ergeben sich hier wieder einerseits durch die freie Nutzung von Suchdiensten, wenn diese nicht vorher eingeschränkt wurden, aber auch andererseits durch die Suchanfragen selber. Die zu bewertenden Ergebnisse beziehen sich immer auf die tatsächlich eingegebenen Suchanfragen, die von Testperson zu Testperson variieren können.

Als Kontrolle ist es denkbar, selbst formulierte Suchanfragen zur Generierung von Ergebnissen mit einfließen zu lassen, so dass jede Testperson als Juror auch diese Ergebnisse auf Relevanz bewerten muss. Dabei muss aber wieder der Aufwand bedacht werden, der auf die Testperson zu kommt.

BORLUND 2003 gibt an, dass zu dem eigentlichen Szenario, das hinter der Arbeitsaufgabe steht auch Hinweise angegeben werden können, was zum Informationsbedürfnis an Dokumenten und Informationen gefunden werden sollte. Eine Beeinflussung in Bezug auf das Suchverhalten und der Formulierung der Suchanfragen konnte nicht belegt werden.

Bei der Ausarbeitung der Suchaufgaben sollte weiterhin darauf geachtet werden, welchem Anfragetyp die Aufgabe abdeckt. Handelt es sich um informationsorientierte, navigationsorientierte oder transaktionsorientierte Aufgaben.

Als Quelle für die Inhalte der simulierten Arbeitsaufgaben könnten Pretests mit Personen aus der Nutzergruppe oder Interviews mit Personen aus der Nutzergruppe dienen. So könnten realistische Informationsbedürfnisse ermittelt und als Grundlage genutzt werden.

In Bezug auf Tests mit allgemeinen Suchmaschinen ist es auch denkbar Informationsbedürfnisse aus dem Web TREC zu adaptieren oder wie bei HUFFMAN / HOCHSTER 2007 durch unabhängige Juroren in Bezug auf Suchanfragen entnommen aus Logfiles schätzen zu lassen.

4.2.6 Festlegung von Skalen

Wie bei klassischen Retrievaltests werden auch hier in den automatisch generierten Retrievaltests am Ende einer Suchsession, eine begrenze Anzahl von Suchergebnissen durch die Testperson explizit bewertet.

Wie bei LEWANDOWSKI 2012 gezeigt, sollte die binäre Bewertung der Relevanz zusammen mit einer abgestuften Skala (z. B. 5-Punkte Skala) erfolgen, da eine rein binäre Bewertung nicht immer ausreichend ist. Suchmaschinennutzer tendieren dazu Dokumente mit geringem Informationsgehalt ebenfalls als relevant einzustufen.

Innerhalb des Modells ist eine Beschränkung der Skalen auf die oben genannten aber nicht notwendig. Wenn es gewünscht ist, könnten auch weitere Bewertungskriterien abgefragt werden. Eine nachträgliche Klassifikation der Dokumente durch die Probanden wäre ebenfalls denkbar.

4.2.7 Analyse der explorativen Suche

Die zuvor genannten Schritte bezogen sich auf die Vorbereitung und die Überlegungen zum Testsetting. Sind alle Entscheidungen getroffen, können die Evaluierungen durchgeführt werden.

Die Testperson beginnt den Test damit, demografische Daten einzugeben, die durch ein Formular abgefragt werden. Danach beginnt die Bearbeitung simulierter Arbeitsaufgaben. Während der Bearbeitung einer simulierten Arbeitsaufgabe fallen dann mitprotokollierte Daten an, die in irgendeiner technischen Form (Textdokument, Datenbank etc.) gespeichert werden.

Sind alle Daten zu einer Aufgabe erfasst, ist es notwendig diese im Anschluss zu analysieren, um auf der einen Seite Informationen zum Nutzerverhalten zusammenzustellen und andererseits, um die Grundlage für den anschließenden Retrievaltest zu schaffen. Die Daten zur Generierung der Ergebnissets können dabei folgende sein:

1. Eingegebene Suchanfragen des Benutzers während der Suchsession; mit Hilfe der tatsächlichen Suchanfragen lassen sich Ergebnismengen von zu testenden Suchdiensten erzeugen. Dabei ist zu entscheiden, ob auch Suchanfragen für Retrievaltests genutzt werden sollten, die zu keinem Klick in der Trefferliste durch den Probanden führten. Soll ein bestimmtes System getestet werden, müssen die Suchanfragen in jedem Fall an dieses Suchsystem gestellt werden, wurde die Wahl der Suchdienste freigestellt, ist es trotzdem denkbar die Suchanfragen an bestimmte Suchmaschinen zu schicken und die Ergebnisse dazu bewerten zu lassen. In jedem Fall sollte aber ein Cutoff-Wert für die maximale Anzahl an Treffern pro Suchmaschine festgelegt werden, um die Anzahl der zu bewertenden Items zu begrenzen, da sonst eine zu große Menge durch die Juroren bewertet werden müsste.
2. Geklickte Suchergebnisse durch die Testperson; wird eine explizite Beurteilung von geklickten Treffern gewünscht, müssen diese unbedingt in den Retrievaltest einbezogen werden. Dabei lässt sich dann auch vergleichen ob häufig geklickte Treffer auch eine höhere Relevanzbewertung erhalten.
3. Nicht-geklickte Suchergebnisse; diese Suchergebnisse lassen sich durch die Ergebnislisten zu den gestellten Suchanfragen ermitteln. Es können geklickte Dokumente aus den Trefferlisten ermittelt werden und dadurch kann bis zu einem bestimmten Cutoff-Wert bestimmt werden, welche der Suchergebnisse nicht angeklickt wurden.

Die Durchführung dieser Analyse wird innerhalb des Prototyps (siehe Kapitel 6) automatisch durchgeführt. Es wäre aber auch denkbar eine manuelle Analyse durchzuführen, um die genannten Daten für die Zusammenstellung der zu beurteilenden Dokumente zu verwenden. In jedem Fall gibt es noch einige Entscheidungen, die bei der Datenerhebung für die Retrievaltests getroffen werden müssen. So sollte, wie bereits erwähnt, die Anzahl der zu beurteilenden Items begrenzt werden.

Eine weitere Entscheidung besteht darin, neben den eigentlichen Ergebnissen auch Trefferbeschreibungen bewerten zu lassen. Nach LEWANDOWSKI 2011 sollten diese unbedingt in den Tests berücksichtigt werden, da aufgrund von schlechten Beschreibungen Treffer, relevante Suchergebnisse verloren gehen können. Die Trefferbeschreibungen können das Ergebnisset verdoppeln, fallen aber bei manuell aufgesuchten Webseiten weg.

Weitere Faktoren, die die Anzahl der Suchitems beeinflussen sind Anzahl der Suchmaschinen und Anzahl der Suchanfragen, die durch Nutzer an die Suchdienste geschickt wurden. Wurden bereits Suchanfragen durch Testpersonen erhoben, ist es denkbar, diese als Ausgangspunkt zu nehmen, da dadurch eine stärkere Vergleichbarkeit von Ergebnismengen geschaffen wird. Solche Suchanfragen könnten beispielsweise auch durch Pretests erhoben werden, bei denen Nutzer aus der Nutzergruppe erst mal nur im Rahmen der Nutzerbeobachtung Arbeitsaufgaben bearbeiten, auch Interviews könnten dabei helfen Suchanfragen als Basis zu sammeln.

In jedem Fall besteht bei der Nutzung der protokollierten Daten bzw. der direkten Beurteilung der Items durch die Testperson nach Bearbeitung der Aufgabe der Vorteil darin, dass der Proband gleichzeitig auch Juror der Suchergebnisse ist. Es werden also die vom Nutzer tatsächlich gesuchten und angeklickten Dokumente verwendet. Die Beurteilung erfolgt nicht durch eine andere Person, die sich zunächst in die Aufgabe und das Informationsbedürfnis hineinversetzen muss. Das sogenannte Stellvertreterproblem kann in diesem Zusammenhang nicht auftreten. Bei dem Stellvertreterproblem handelt es sich um die Tatsache, dass ausgesuchte Juroren Ergebnisse bewerten, obwohl diesen das Informationsbedürfnis des eigentlichen Nutzers nicht bekannt ist. Sie müssen sich in die Rolle des ursprünglichen Nutzern hineinversetzen. (vgl. MÖHR 1980, S. 135)

Als Nachteil ergibt sich hier aber je nach festgelegten Testbedingungen ein Problem bei der Vergleichbarkeit der Daten. Arbeiten die Nutzer alle mit verschiedenen Suchdiensten und nutzen sehr unterschiedliche Informationsstrategien, um die Aufgaben zu bearbeiten, fallen keine vergleichbaren Daten an.

4.2.8 Zusammenführung und Analyse der Daten aus den explorativen Suchen und den Retrievaltests

Im nächsten Schritt werden die Daten aus den explorativen Suchen und aus den Retrievaltests zusammengeführt. Dabei sollten für die Übersicht alle Daten, die sich nur auf die explorative Suche und den Nutzer beziehen (demografische Daten, Anzahl der einzelnen Events, Bearbeitungszeit einer Aufgabe, genutzte Suchdienste, Anzahl der Suchanfragen für die Bearbeitung einer Suchaufgabe, manuell aufgerufene Webseiten, geklickte Webseiten, gebookmarkte Webseiten, Einschätzungen über Pre- und Post-Fragebögen) pro Testperson zunächst einzeln aufgeführt werden. Anschließend sollte dann die Aufstellung der Werte aus dem Retrievaltest im Verbindung mit Daten aus den Nutzerinteraktionen (URL des Dokuments, Trefferbeschreibung, Suchmaschinen, Position im Ranking, Skalenwerte, geklickter Treffer, gebookmarkter Treffer, manuell aufgerufener Treffer) erfolgen.

Dabei bietet sich auch eine Zusammenfassung der Werte aus dem Retrievaltest bezogen auf alle Nutzer an, um diese als Datengrundlage für die Berechnung verschiedener Retrievalmaße zu nutzen. Die Auswahl der Kennzahlen für die Messung und Auswertung hängt dabei stark von der Ausgangsfrage der Untersuchung ab, z. B. könnten die vorgeschlagenen Maße von BORLUND 2003 verwendet werden, wenn durch eine Auswertung der subjektiven Relevanz oder Zufriedenheit der Benutzer durch Fragebögen und den Klickdaten berechnet wird und mit der algorithmischen Relevanz, die durch klassische Retrievalmaße erfasst werden kann, verglichen wird. Die erfassten Daten könnten auch mit Kennzahlen ausgewertet werden, die speziell im Kontext von Websuchmaschinen nutzbar sind (siehe Kap 2.1.5).

Je nach verwendeten Datenspeicher (Textdateien, Datenbanken etc.) und technischen Werkzeugen für die Studie, können die Daten manuell oder automatisch zusammengeführt werden.

4.2.9 Zusammenfassung des Modells

Das beschriebene Modell ist insgesamt ein Vorschlag für die Untersuchung der Retrievaleffektivität von Suchdiensten in Verbindung mit dem tatsächlichen, realen Benutzer. Als große Vorteile ergeben sich eine Generierung von Ergebnissets auf Grundlage der Nutzeraktionen in seiner Suche, die anschließend auch von diesem Nutzer bewertet werden. Dabei ist aber immer noch zu bedenken, dass Intentionen des Nutzers nicht direkt erfasst werden können, wenn protokollierte Logdateien als Grundlage dienen. Daher ist eine Befragung der Testpersonen ein sehr wichtiger Faktor bei der Durchführung der Tests. Trotzdem bietet das Modell die Möglichkeit, dass viel kritisierte Cranfield-Paradigma (vgl. JÄRVELIN 2009 S. 291ff.), um die wichtige Komponente des Nutzer zu erweitern.

5. Tools in der Information-Retrieval Evaluierung

In diesem Abschnitt werden vor allem die Softwareanwendungen vorgestellt, die für die Entwicklung des Prototyps wichtig sind. Zum einen wird das Relevance Assessment Tool (RAT) dargestellt, das für die Gestaltung und Durchführung von Relevanztests an der HAW Hamburg entwickelt wurde und zum anderen wird der Search Logger präsentiert, der für die Protokollierung von Nutzerinteraktionen innerhalb explorativer Suchen am Institut für Computerwissenschaften an Universität Tartu programmiert wurde. Der Prototyp soll beide Technologien verbinden, damit das oben angeführte Modell in der Praxis eingesetzt werden kann.

5.1 Relevance Assessment Tool (RAT)

Wie in Kapitel 2.2.1 gezeigt, sind Relevanztests das gängige Verfahren bei der Messung der Retrievaleffektivität von Suchmaschinen und Informationssystemen. Die Gestaltung und Durchführung dieser Tests ist mit einem hohen Aufwand z. B. bei der Beschaffung der zu testenden Ergebnisse oder der Erhebung der Relevanzurteile, verbunden. Damit dieser Aufwand verringert werden kann, wurde an der Hochschule für Angewandte Wissenschaften Hamburg am Department Information ein Online-Tool entworfen und programmiert, das diese Aufgaben erleichtert (vgl. LEWANDOWSKI / SÜNKLER 2011).

Das Relevance Assessment Tool bietet Komponenten und Eingabemasken, die das Studiendesign und die Bewertung von Dokumenten erleichtern und standardisieren. Dadurch können Tests auch systematisch wiederholt und verglichen werden.

5.1.1 Technische Voraussetzungen

Das Relevance Assessment Tool ist eine Webanwendung, die sich nach der Installation in einer Serverumgebung, über jeden herkömmlichen Webbrowser aufrufen lässt. Für die Serverumgebung werden dabei die Skriptsprache PHP in der Version 5.3 und das Datenbankmanagementsystem MySQL in der Version 5.1 benötigt. PHP wurde als Sprache gewählt, da diese zum einen kostenlos verfügbar ist, alle Programmbibliotheken mitbringt, die die gewünschten Funktionen des Tools unterstützen und weil die meisten herkömmlichen Webspaceanbieter eine PHP Umgebung mit Anbindung zu einer MySQL Datenbank bereitstellen. Daneben ist noch das Vorhandensein der Programmbibliothek cURL notwendig, welche weiter unten näher beschrieben wird. Ein weiterer Vorteil ergibt sich noch bei der Guten Dokumentation der Programmiersprache und der sehr starken Verbreitung. (vgl. KOFFLER / ÖGGL 2010, S. 16 ff.)

Die Installation der Anwendung wird dabei durch das Kopieren der Programmdateien auf den Webspace und die Installation der MySQL-Datenbank realisiert.

5.1.2 Komponenten des Relevance Assessment Tools

Im aktuellen Entwicklungsstand stellt das Relevanztool drei Komponenten zur Verfügung, die für die Gestaltung und Durchführung der Tests notwendig sind. Für die automatische Erfassung von Suchergebnissen zu den Testanfragen wurde ein Suchmaschinenscraper entwickelt, der stetig angepasst und erweitert werden muss. Die Administration der Projekte wird mit Hilfe eines Backends durchgeführt und zuletzt bewerten die Testpersonen alle Suchitems mit Hilfe eines speziell programmierten Userinterface. Eine Komponente zur automatischen Auswertung der Testergebnisse befindet sich zur Zeit noch in der Entwicklung (vgl. Lewandowski / Sünkler 2011).

5.1.2.1 Der Suchmaschinenscraper

Ein Scraper (auch: Wrapper) ist allgemein eine Anwendung, die Informationen aus Textdateien extrahiert und in definierten Formaten speichert. Bei Webseiten kann ein Scraping folgendermaßen aussehen. Zunächst wird die Webseite automatisch aufgerufen und nach festgelegten Kriterien analysiert, bevor eine Extrahierung der gewünschten Daten (z. B. die Links auf der Seite) erfolgt. Beispiele für Algorithmen zur Gestaltung von Scrapern stellt Kuschmerick 2000 (aus Lewandowski / Sünkler) vor.

Für das Relevanztool wurde ein spezieller Suchmaschinenscraper entwickelt, der dazu in der Lage ist, zum einen automatisch Ergebnissets zu definierten Suchanfragen von marktführenden Websuchmaschinen zu generieren und zum anderen auch ein Import von Suchergebnissen in Form von Excel-Dateien unterstützt. Dadurch lassen sich dann auch Ergebnisse bewerten, die von spezielleren Suchdiensten bereitgestellt werden.

Der Suchmaschinenscraper basiert im wesentlichen auf der PHP Programmbibliothek cURL (Client for URLs) [5], mit dem es möglich ist Web-Clients, um so automatisiert Datenabfragen ans Internet zu senden. Dabei werden alle gängigen Protokolle wie HTTP, HTTPS, FTP und FTPs unterstützt. Für den Suchmaschinenscraper ist dabei vor allem die Funktion relevant Quelltext von Webseiten mit definierten Anfragen zurückzugeben, um diesen anschließend auf relevante Bestandteile auszuwerten. Neben den cURL-Funktionen im Scraper werden aber auch Filter benötigt, die ein automatisches Auslesen der Suchaergebnisse von Webseiten ermöglichen.

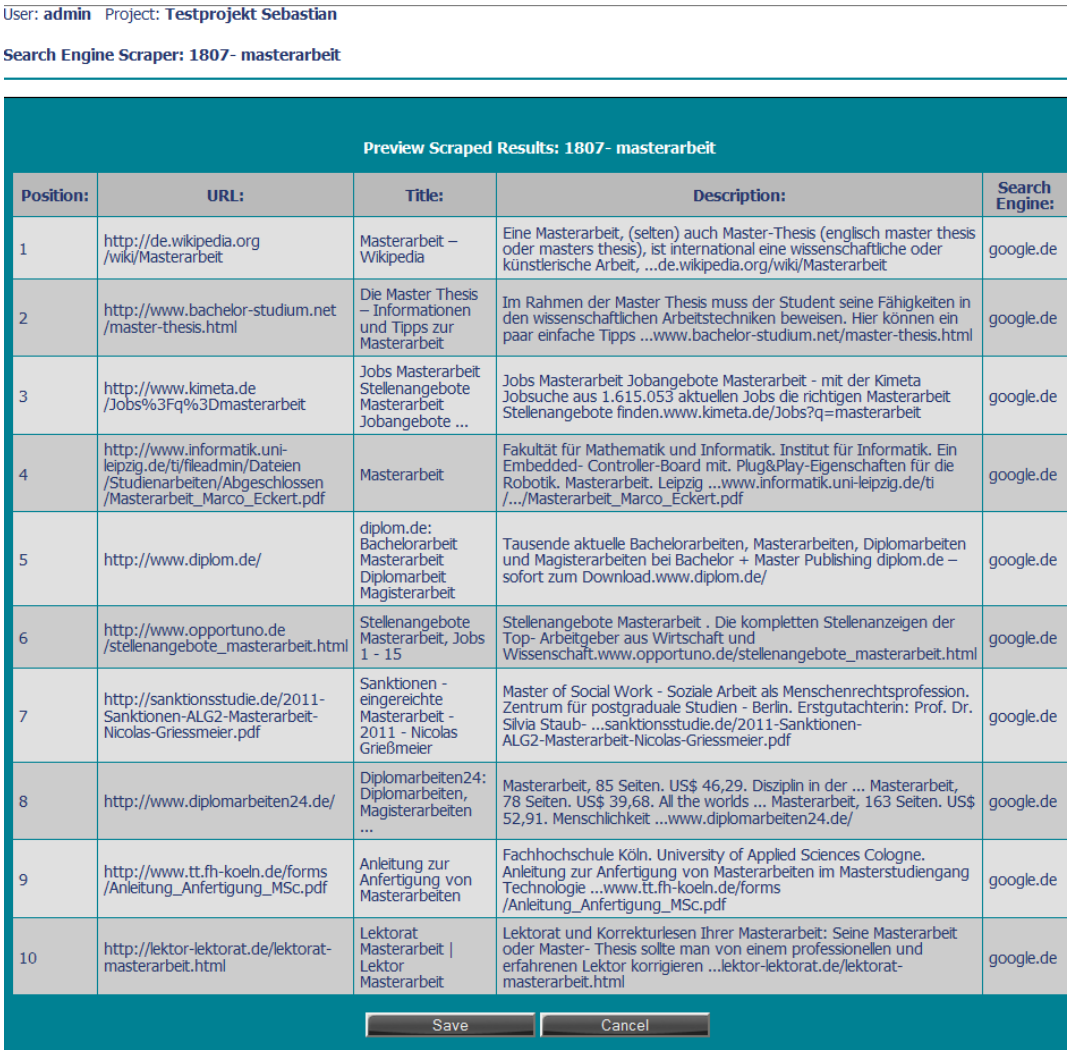
5 Eine Übersicht zu allen Funktionen von cURL findet sich unter http://php.net/manual/de/book.curl.php

Innerhalb des Relevance Assessment Tools arbeitet der Scraper in einer vorgegebenen Reihenfolge. Zunächst legt der Projektadministrator Suchanfragen fest, für diese er eine bestimmte Anzahl an Suchergebnissen wünscht. Diese werden anschließend über den Scraper an ausgewählte Suchmaschinen gesendet und dann durch definierte Filter ausgewertet (Anhang A 1 zeigt am Beispiel von Microsoft Bing, wie so ein Filter im Tool aussieht). Die Suchmaschinen geben ihre Trefferliste als HTML-Datei zurück, die geparst und durch XML in einer Baumstruktur gespeichert wird. In der Baumstruktur sind alle Elemente der geparsten Trefferseite als hierarchische Knoten zugänglich und auslesbar. Die relevanten Elemente der Suchergebnisliste (URL, Trefferüberschrift und Trefferbeschreibung) werden über die XML Abfragesprache XPath mit Hilfe der HTML-Gestaltungstags erfasst. Jede Suchmaschine nutzt dabei andere Strukturen und Gestaltungselemente für die Suchergebnisseite, auch wenn sich einige Standards in der Ergebnispräsentation entwickelt haben (vgl. LEWANDOWSKI / HÖCHSTÖTTER 2009). Beispielsweise nutzt Google andere HTML-Tags zur Formatierung der Trefferbeschreibungen als Yahoo oder Microsoft Bing, auch wenn alle auf dieses Gestaltungselement zurückgreifen. Die Filter sorgen im Moment auch dafür, dass zunächst nur organische Treffer, im Gegensatz zu Universal Search oder Sponsored Links ausgelesen werden. In Zukunft ist aber eine Anpassung der Filter für die Auswahl bestimmter Trefferelemente geplant.

Wurden alle wichtigen Trefferelemente ausgelesen und gefiltert, werden in einem nächsten Schritt lokale Kopien der Dokumente, wieder mit Hilfe von cURL, angelegt. Dieser Verfahren wird genutzt, um auf der einen Seite zu garantieren, dass auch tatsächlich die Ergebnisse bewertet werden, die zum Zeitpunkt der gestellten Suchanfragen durch die Suchmaschinen gefunden wurden und zum anderen lassen sich Elemente wie manipulierendes Javascript wieder über Filter entfernen (z. B. nutzen einige Anbieter sogenannte Framebreaker, die eine Bewertung des Dokuments innerhalb des Nutzerinterfaces nicht möglich machen, weil die Framebreaker, dass Interface überdecken). Bei der Speicherung des Dokuments erkennt der Scraper automatisch das Dateiformat, so kann beispielsweise zwischen HTML-Dateien und PDF Dokumenten differenziert werden.

Der Weg über die direkte Analyse der Suchergebnisseiten aus den Suchmaschinen wurde für das Tool gewählt, um die tatsächlichen Suchergebnisse zu Anfragen zu erhalten. Denkbar wäre auch der Einsatz sogenannter APIs von den Suchmaschinenanbietern, um Zugriff auf die Suchfunktionen und Ergebnisse zu erhalten (vgl. Tosques Mayr). Ein großer Nachteil ergibt sich hier aber daraus, dass bei den APIs häufig ein anderer Index von den Suchmaschinen angeboten wird, was die Suchergebnisse und das Ranking beeinflussen würde. Daneben bietet auch nicht jede Suchmaschine APIs an (vgl. LEWANDOWSKI / SÜNKLER 2012).

Momentan ist es möglich den Scraper einzeln für jede Suchanfrage und Suchmaschine zu nutzen, mehrere Suchanfragen und Suchmaschinen gleichzeitig zu scrapen oder auch Ergebnislisten für andere Zwecke zusammenzustellen. Die Suchergebnisse müssen nicht zwingend innerhalb von Relevanztests genutzt werden.



User: **admin** Project: **Testprojekt Sebastian**

Search Engine Scraper: 1807- masterarbeit

Preview Scraped Results: 1807- masterarbeit

Position:	URL:	Title:	Description:	Search Engine:
1	http://de.wikipedia.org/wiki/Masterarbeit	Masterarbeit – Wikipedia	Eine Masterarbeit, (selten) auch Master-Thesis (englisch master thesis oder masters thesis), ist international eine wissenschaftliche oder künstlerische Arbeit, ...de.wikipedia.org/wiki/Masterarbeit	google.de
2	http://www.bachelor-studium.net/master-thesis.html	Die Master Thesis – Informationen und Tipps zur Masterarbeit	Im Rahmen der Master Thesis muss der Student seine Fähigkeiten in den wissenschaftlichen Arbeitstechniken beweisen. Hier können ein paar einfache Tipps ...www.bachelor-studium.net/master-thesis.html	google.de
3	http://www.kimeta.de/Jobs%3Fq%3Dmasterarbeit	Jobs Masterarbeit Stellenangebote Masterarbeit Jobangebote ...	Jobs Masterarbeit Jobangebote Masterarbeit - mit der Kimeta Jobsuche aus 1.615.053 aktuellen Jobs die richtigen Masterarbeit Stellenangebote finden.www.kimeta.de/Jobs?q=masterarbeit	google.de
4	http://www.informatik.uni-leipzig.de/ti/fileadmin/Dateien/Studienarbeiten/Abgeschlossen/Masterarbeit_Marco_Eckert.pdf	Masterarbeit	Fakultät für Mathematik und Informatik. Institut für Informatik. Ein Embedded- Controller-Board mit. Plug&Play-Eigenschaften für die Robotik. Masterarbeit. Leipzig ...www.informatik.uni-leipzig.de/ti/.../Masterarbeit_Marco_Eckert.pdf	google.de
5	http://www.diplom.de/	diplom.de: Bachelorarbeit Masterarbeit Diplomarbeit Magisterarbeit	Tausende aktuelle Bachelorarbeiten, Masterarbeiten, Diplomarbeiten und Magisterarbeiten bei Bachelor + Master Publishing diplom.de – sofort zum Download.www.diplom.de/	google.de
6	http://www.opportuno.de/stellenangebote_masterarbeit.html	Stellenangebote Masterarbeit, Jobs 1 - 15	Stellenangebote Masterarbeit . Die kompletten Stellenanzeigen der Top- Arbeitgeber aus Wirtschaft und Wissenschaft.www.opportuno.de/stellenangebote_masterarbeit.html	google.de
7	http://sanktionsstudie.de/2011-Sanktionen-ALG2-Masterarbeit-Nicolas-Griessmeier.pdf	Sanktionen - eingereichte Masterarbeit - 2011 - Nicolas Grießmeier	Master of Social Work - Soziale Arbeit als Menschenrechtsprofession. Zentrum für postgraduale Studien - Berlin. Erstgutachterin: Prof. Dr. Silvia Staub- ...sanktionsstudie.de/2011-Sanktionen-ALG2-Masterarbeit-Nicolas-Griessmeier.pdf	google.de
8	http://www.diplomarbeiten24.de/	Diplomarbeiten24: Diplomarbeiten, Magisterarbeiten ...	Masterarbeit, 85 Seiten. US$ 46,29. Disziplin in der ... Masterarbeit, 78 Seiten. US$ 39,68. All the worlds ... Masterarbeit, 163 Seiten. US$ 52,91. Menschlichkeit ...www.diplomarbeiten24.de/	google.de
9	http://www.tt.fh-koeln.de/forms/Anleitung_Anfertigung_MSc.pdf	Anleitung zur Anfertigung von Masterarbeiten	Fachhochschule Köln. University of Applied Sciences Cologne. Anleitung zur Anfertigung von Masterarbeiten im Masterstudiengang Technologie ...www.tt.fh-koeln.de/forms/Anleitung_Anfertigung_MSc.pdf	google.de
10	http://lektor-lektorat.de/lektorat-masterarbeit.html	Lektorat Masterarbeit | Lektor Masterarbeit	Lektorat und Korrekturlesen Ihrer Masterarbeit: Seine Masterarbeit oder Master- Thesis sollte man von einem professionellen und erfahrenen Lektor korrigieren ...lektor-lektorat.de/lektorat-masterarbeit.html	google.de

Save Cancel

Abbildung 8: Vorschauseite des Suchmaschinenscraper

Abb. 8 zeigt ein Beispiel für die Vorschauseite zum Scrapen der Suchergebnisse innerhalb der Anwendung mit allen zu speichernden Elementen. Der praktische Nutzen des Scrapers auch in spezielleren Projekten sowie Hinweise zur Weiterentwicklung durch andere Suchdienste wird weiter unten im Text erläutert.

5.1.2.2 Das Backend zur Projektverwaltung

Das Backend zur Gestaltugn von Projekten / Relevanztests setzt sich aus einigen Optionen zusammen, die für das Testdesign notwendig sind. Der Aufruf erfolgt dabei über die Menüleiste über den Navigationslink „Projects“.

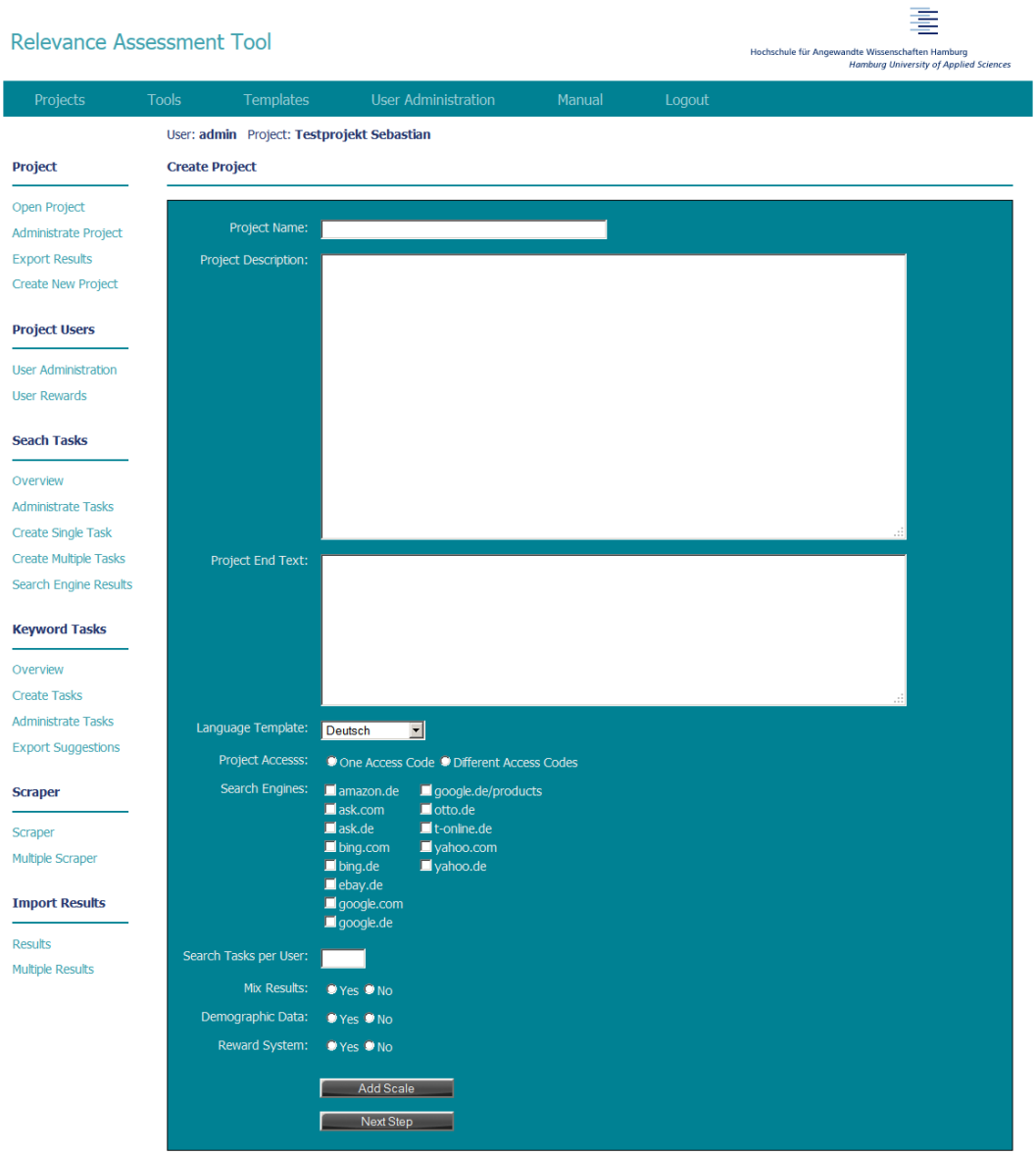


Abbildung 9: Ausschnitt des Backends mit Formular zur Gestaltung von Projektinhalten

Die Abbildung zeigt das Formular zur Gestaltung neuer Projekte sowie auf der linken Seite alle Optionen, die zum Beispiel für die Verwaltung der Testpersonen, der Suchanfragen, dem Importieren und Scrapen von Suchergebnissen dienen.

Nach dem Login in das Backend kann ein neues Projekt angelegt werden. Dabei werden insgesamt 11 Angaben gemacht, die das Design der Studie bestimmen:

1. Eingabe des Projektnamens
2. Beschreibung des Projekts / Begrüßungstext für die Juroren (wird im Nutzerinterface angezeigt)
3. Abschlusstext des Projekts (wird im Nutzerinterface angezeigt)

4. Auswahl eines Sprachtemplates für das Projekt (Templates für Sprachen und Bewertungsskalen werden separat angelegt. Dadurch ist eine Wiederverwendbarkeit gegeben.
5. Auswahl über den Zugang zu den Suchaufgaben. Dem Projektleiter stehen hier zwei Möglichkeiten zur Auswahl. Entweder wird ein Access-Code generiert, bei dem beliebig viele Testpersonen mit Eingabe im Nutzerinterface Zugriff auf die Suchaufgaben erhalten oder es lassen sich „Different Access-Codes“ festlegen. Diese können dann auf verschiedene Nutzergruppen verteilt werden und sind limitiert. Also z. B. könnten 15 Zugangscodes für Experten und 30 Codes für Laien bei Suchmaschinentests generiert werden. Die Zugangscodes werden in einem Schritt bei der Erstellung des Projekts ausgegeben und können als HTML, Text- oder CSV-Datei exportiert werden, auch ein direkter Ausdruck ist möglich.
6. Im nächsten Schritt wird über die Suchmaschinen im Test entschieden. Sollen bei Suchanfragen, aber nur bestimmte Suchmaschinen getestet werden, lässt sich dieses noch explizit dort festlegen.
7. Über „Search Task per User“ wird definiert, wie viele Suchaufgaben ein Proband maximal bearbeiten soll, wobei Suchaufgabe hier meint, wie viele Suchergebnisse zu einer Anfrage von einem Juror bewertet werden sollen.
8. Ferner wird noch über die Darstellung der Treffer im Interface entschieden. So können bei Einbeziehung der Trefferbeschreibungen diese entweder immer zunächst angezeigt werden, bevor das dahinter liegende Dokument bewertet wird oder es lässt sich auch einstellen, ob die Trefferbeschreibungen und Suchergebnisse separat bewertet werden (also zuerst alle Trefferbeschreibungen und dann alle Ergebnisse).
9. Daneben kann auch bestimmt werden, ob demografische Daten im Test erhoben werden sollen.
10. Das Tool bietet daneben noch die Möglichkeit ein Belohnungssystem zu nutzen, dass Testteilnehmern bei Angabe einer E-Mail Adresse automatisch einen Gutscheincode zusendet, welcher in einer Liste innerhalb der Datenbank abgelegt ist.
11. Am Ende müssen noch Skalen ausgewählt werden, auf denen die Beurteiler die Bewertung der Suchergebnisse vornehmen. Die Skalen werden, wie auch die Sprache global definiert, damit diese auch in anderen Projekten wiederverwendet werden können.

Nachdem alle Entscheidungen getroffen wurden, ist es notwendig Suchaufgaben zu definieren. Dabei legt der Administrator des Projekts fest, welche Beschreibung zu der Aufgabe im Nutzerinterface sichtbar sein soll, ob Trefferbeschreibung und Suchergebnis bewertet werden sollen und auch ob die Suchanfrage sichtbar sein soll oder nicht. Nachdem alle Anfragen definiert wurden, müssen diesen mit Hilfe des

Scrapers Suchergebnisse zugewiesen werden. Alle Eingaben werden innerhalb der MySQL-Datenbank für die Projekte gespeichert.

5.1.2.2 Das Nutzerinterface

Im Nutzerinterface bzw. Frontend des Tools, werden die Bewertungen der Suchergebnisse durch die Testpersonen durchgeführt. Der Nutzer sieht dabei den Fortschritt seiner Bearbeitung, die Suchaufgabe, auf der linken Seite die Skalen zur Bewertung und rechts das zu bewertende Ergebnis. Nachdem alle Skalen ausgefüllt sind, kann auf den Button für das nächste Ergebnis geklickt werden. Fällt es der Testperson schwer eine Bewertung des Ergebnisses vorzunehmen, lässt sich dieses auch Überspringen.

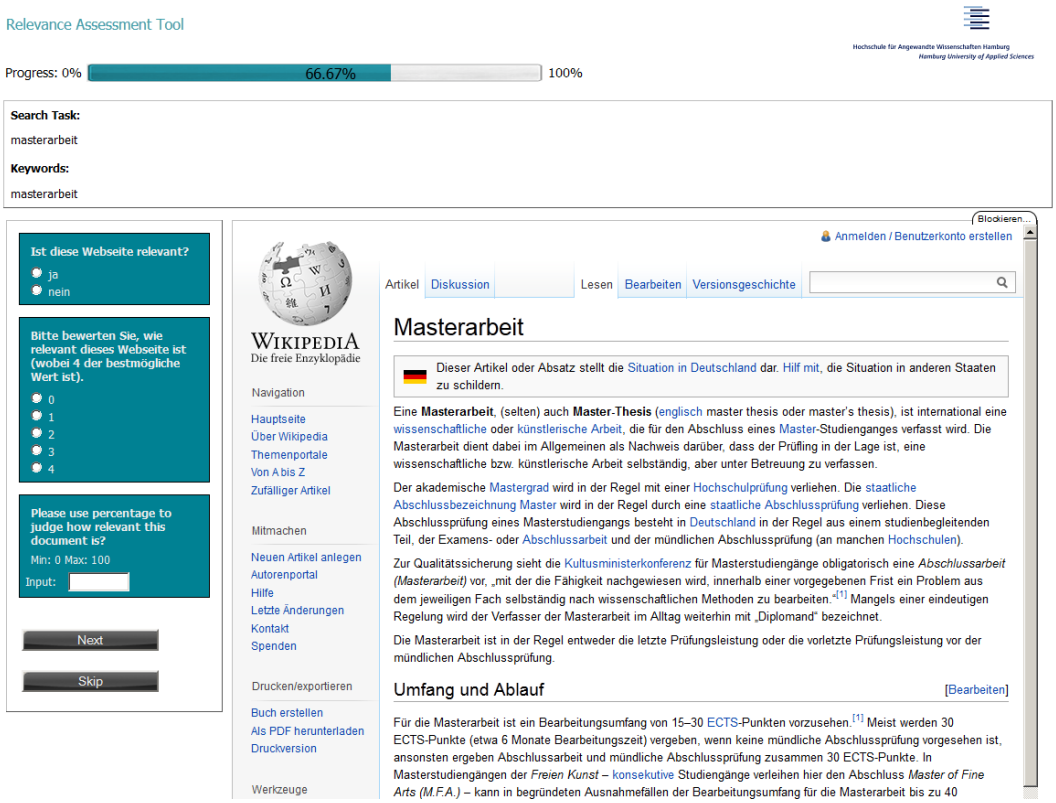


Abbildung 10: Nutzerinterface des Relevance Assessment Tools

5.1.3 Kurzer Überblick zum praktischen Einsatz des Relevance Assessment Tools

Das Relevance Assessment Tool wurde bisher überwiegend von Studenten im Rahmen von Abschlussarbeiten oder Projekten an der Hochschule für Angewandte Wissenschaften eingesetzt. Im Folgendem werden kurz die durchgeführten Untersuchungen und die Rolle des Tools dabei allgemein dargestellt.
Zweck dieser Darstellung ist eine Übersicht zu den vielseitigen Einsatzmöglichkeiten des Tools zu geben.

November 2010 (Masterarbeit über die Analyse von Suchanfragen und Suchergebnisseiten, die keine Klicks auslösten):
Innerhalb dieser Arbeit wurden Suchanfragen erforscht, die keine Klicks auf Suchergebnisseiten auslösten. Die Studentin nutze dabei eine Reihe von Methoden, um ihrer Forschungsfrage nachzugehen. Neben eine deskriptiven Analyse und einer Online-Befragung zur Erfassung der Gründe für die Unterlassung von Klicks wurde das Tool zur Relevanzmessung von Suchergebnissen und Trefferbeschreibungen nicht angeklickter Dokumente eingesetzt.

November 2010 (Einsatz in einem Seminar zum Thema Semantic Search im Masterstudiengang Informationswissenschaft und -management):
Der Einsatz des Tools beschränkt sich nicht auf Abschlussarbeiten und Projekten, auch innerhalb von Seminaren zu Fragestellungen der Websuche oder spezieller Suchdienste ist eine Nutzung denkbar und möglich. In dem Seminar „Semantic Search“, dass im Wintersemester 2010/2011 als Wahlpflichtmodul im Masterstudiengang Informationswissenschaft und -management angeboten wurde, haben zwei kleinere Gruppen aus Studenten das Tool als Basis für die Erarbeitung verschiedener Fragestellungen genutzt. Bei den Miniprojekten ging es immer um Vergleiche von semantischen Suchdiensten mit konventionellen, auf statisch-linguistisch basierenden Suchmaschinen. Eine Gruppe setzte sich dabei mit allgemeinen semantischen Suchmaschinen auseinander, während die andere Gruppe einen Vergleich medizinischer Suchdienste durchführte.

Januar 2011 (Studentisches Projekt für einen Auftraggeber für den Vergleich von Download Musikdiensten im Internet):
Im Wintersemester 2010/2011 führten Studenten im Studiengang Medien und Information ein Projekt durch, bei dem ein Anbieter Vergleichsdaten zur Relevanz verschiedener Musik-Download-Dienste im Internet benötigte.

Das Relevance Assessment Tool wurde zu diesem Zweck angepasst und durch Scraper der getesteten Dienste erweitert. Insgesamt wurden innerhalb des Tools 50 verschiedene Anfragen zu Künstlern und den jeweils produzierten Ergebnissen erfasst und durch verschiedene Testpersonen mit Hilfe entsprechender Skalen bewertet.

August 2011 (Masterarbeit am Department für Health Sciences and der HAW Hamburg):

Im Rahmen dieser Masterarbeit wurde das Tool das erste Mal außerhalb des Departments Information durch eine Studentin für ihre Abschlussarbeit eingesetzt. In der Arbeit ging es um Lebensmittelkontrolle im Internet und mit Hilfe des Tools und vor allem des Scrapers sollten Ergebnisse kategorisiert werden. Die Studentin hat mit Hilfe des Nutzerinterfaces die gefunden Ergebnisse, besonders zu Suchanfragen, die sich auf Inhaltsstoffe in Lebensmittel, bezogen, durch eigens definierte Skalen eingeordnet.

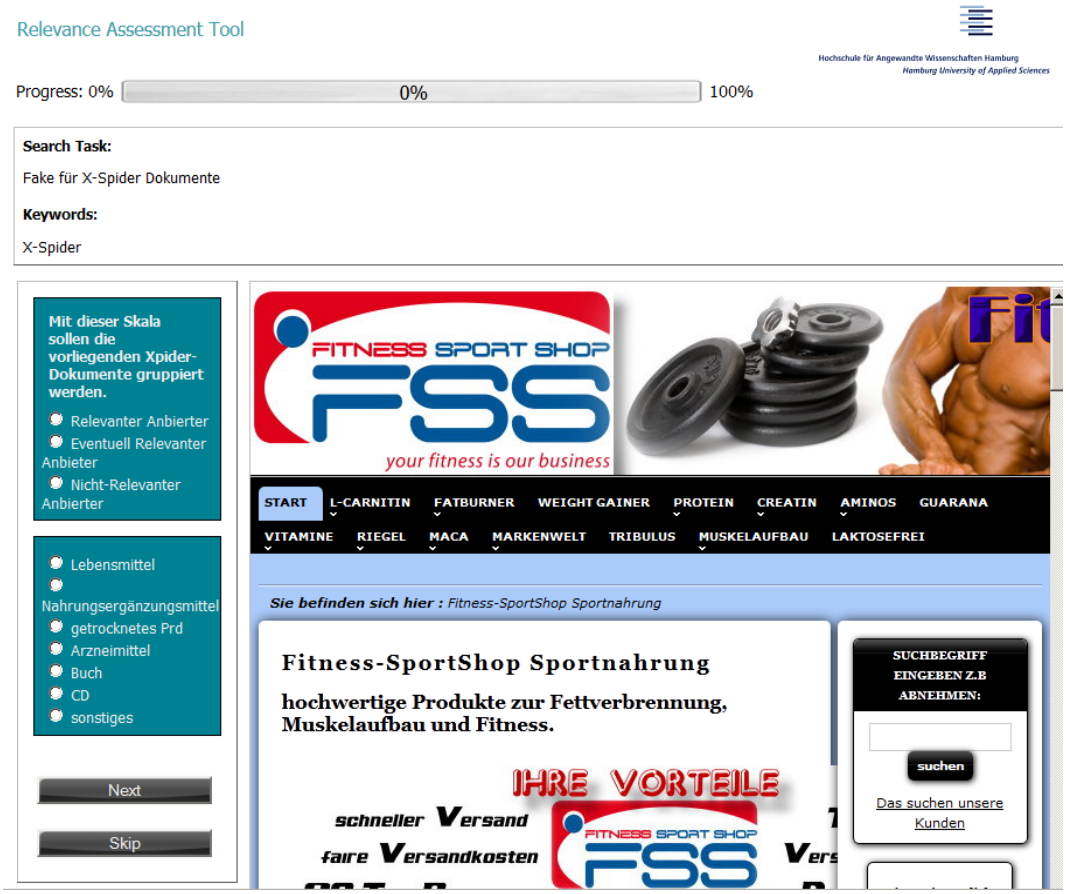


Abbildung 11: Einsatz des Relevance Assessment Tools zur Kategorisierung von Webseiten

Eine Relevanzbeurteilung der Dokumente wurde nicht angestrebt. Dieser Einsatz zeigt die Flexibilität des Tools in Bezug auf Fragestellungen abseits der Retrievaleffektivität von Informationssystemen.

Dezember 2011 (Studentisches Projekt zur Untersuchung von automatisch generierten Suchvorschlägen verschiedener Suchmaschinen):
Durch dieses studentische Projekt, das im Wintersemester 2011/2012 durchgeführt wurde, ergaben sich eine Reihe von Erweiterungen, die das Relevance Tool im Funktionsumfang stark erweiterten. Das Toll sollte Studenten dabei helfen Relevanzbewertungen von Suchvorschlägen auf durch die studentische Gruppe definierten Skalen durchzuführen. Daneben wurde auch eine Anwendung im Tool gewünscht, die eine Excel-Liste mit Suchvorschlägen zu vorher importierten Suchanfragen generiert. Das folgende Beispiel veranschaulicht den Zweck der Anwendung:

Suchanfrage: auto
autoscout
autokraft
auto
autokennzeichen
autohus
auto motor sport
auto wichert

Tabelle 5: Ausschnitt von Google Suchvorschlägen zur Suchanfrage „auto“

Das jeweils generierte Keywordset wurde den Testpersonen innerhalb des Nutzerinterfaces von RAT zur Bewertung vorgelegt. Damit sollten Daten zu der Effektivität der durch die Suchdienstanbieter verwendeten Methoden zur Suchvorschlagsgenerierung gesammelt und analysiert werden.

Januar 2012 (Bachelorarbeit zu der Frage inwieweit die Bildschirmauflösung von Suchergebnisseiten Einfluss auf die Suchmaschinennutzer hat):
Der bisher letzte Anwendungsfall von RAT fand innerhalb der Bearbeitung einer Bachelorarbeit statt. Für die Vergleichbarkeit der Relevanz von Suchergebnisseiten von Google bei verschiedenen Bildschirmauflösungen wurden dementsprechende Projekte angelegt und Screenshots von den Suchergebnissseiten als Ergebnisse importiert.

5.1.4 Ähnliche Tools

Eine Marktsichtung zu ähnlichen Anwendungen zeigt, dass es zwar Tools gibt, die zum Teil Funktionen anbieten, die durch RAT abgedeckt werden, aber das es zumindest keine frei auffindbaren Werkzeuge für die systematische Durchführung von Relevanztests gibt.

Tools, die für das Scraping von Suchergebnissen entwickelt wurden, verfolgen häufig andere Ziele als der vorgestellte Suchmaschinenscraper. Beispielsweise werden diese für die Suchmaschinenoptimierung eingesetzt (vgl. Stokdyk 2009) oder für die Erkennung potenzieller Sicherheitslücken in Webseiten (vlg. Peteron 2006). Daneben kommen diese auch zum Einsatz bei Studien, die ein Harvesting von Webseiten benötigen. LEWANDOWSKI / HÖCHSTÖTTER 2009 nutzen einen Scraper, der zur Analyse der Quellen und Trefferdarstellungen bei US-Versionen von Suchmaschinen eingesetzt wurde.

In der Studie von FOX ET AL. 2005 kam ein Werkzeug zum Einsatz, das sowohl implizite und explizite Urteile von Testpersonen zu Suchanfragen und Ergebnissen sammelte. Dieses war wie der folgende Search Logger ebenfalls ein Browser-Plugin, allerdings für den Internet Explorer. Auch andere Tools z. B. zum Loggen von Benutzerinteraktionen werden in diversen Studien eingesetzt. Ein Problem ist dabei häufig, dass die Anwendungen nicht verfügbar sind oder auch nur wage Angaben zu diesen gemacht werden.

5.2 Search Logger

Die zweite Anwendung, die bei der Umsetzung des Modells eingesetzt wird, ist der sogenannte Search Logger, der zur Untersuchung und Datenerhebung von Nutzerverhalten innerhalb explorativer Suchen eingesetzt wird. Für die Durchführung von Search Logger Studien lassen sich dabei Suchaufgaben definieren.

Ein großer Vorteil beim Search Logger ist, dass dieser auch bei Langzeitstudien eingesetzt werden kann. Eine Bearbeitungszeit von Suchaufgaben ist nicht vorgegeben. So können auch Experimente mit Nutzern über einen längeren Zeitraum beispielsweise mit sehr umfangreichen Arbeitsaufgaben durchgeführt werden.

5.2.1 Technische Voraussetzungen

Die Nutzung vom Search Logger benötigt in der aktuellen Version sowohl eine lokale Installation in den Webbrowser Firefox, da es sich bei der Anwendung um ein Browserplugin handelt. Daneben wird wie beim Relevance Assessment Tool eine Webumgebung mit PHP benötigt, auf der sich die MySQL-Datenbank installieren lässt.

Eine Einschränkung der aktuellen Version bei der lokalen Installation liegt darin, dass diese nur in Verbindung mit einer älteren Version des Firefox ist (Version 3.6). An einer aktualisierten Fassung, sowie an einer Variante mit vorgeschaltetem Proxy, welcher eine zentrale Erfassung der Benutzerdaten ermöglicht und eine lokale Installation auf allen Testsystemen überflüssig macht.

Bei der Entwicklung des Prototyps lag noch keine aktuelle Version vor, daher wurde übergangsweise Downloadmöglichkeiten der portable Versionen (portable Anwendungen können ohne Installation genutzt werden) des Firefox 3.6 im Prototyp umgesetzt, mit dem alle Grundfunktionalitäten zugänglich sind.

5.2.2 Komponenten des Seach Loggers

Der Search Logger ist wie bereits erwähnt ein Browserplugin für Mozilla Firefox. Ein sogenanntes Addon, dass die Funktionalität erweitert. Technisch betrachtet besteht dieses Addon aus HTML, CSS, XML und Javascript Dateien und lässt sich durch einfaches Drag und Drop installieren.

Der Search Logger setzt sich im wesentlichen aus drei Komponenten zusammen. Zum einen gibt es den Search Logger selbst, der als Javascript Datei vorliegt und Browseraktivitäten (eine Übersicht zu allen logbaren Aktionen zeigt Tabelle) erfassen kann. Einstellungen und Optionen für das Addon werden über XML Dateien gesteuert, die sämtliche Menüpunkte enthalten.

Die zweite Komponente ist eine Datenbank, die durch PHP gesteuert wird und in MySQL vorliegt. Das Browserplugin sendet im Hintergrund Befehle in PHP an die Datenbank (Abb. Zeigt die Architektur hinter der Anwendung).

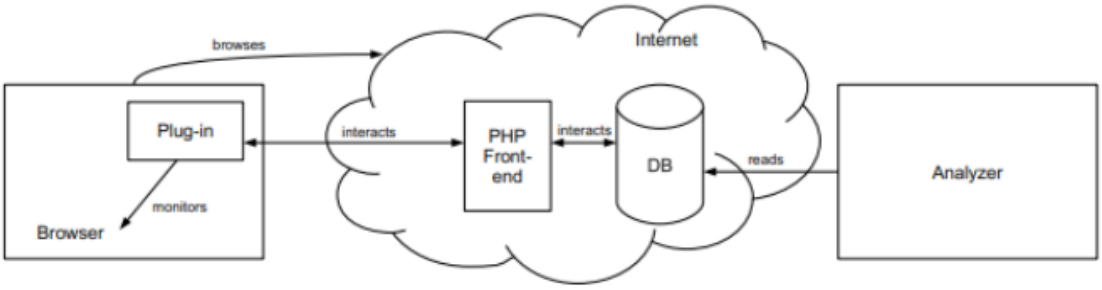


Abbildung 12: Systemarchitektur des Search Loggers (Quelle: Singer et al. 2011, S. 753)

Die dritte Komponente sind die eigentlichen Suchaufgaben, die als HTML-Dateien definiert werden. Insgesamt benötigt jede Aufgabe drei Dateien:

1. Die Einleitung zur Aufgabe mit Beschreibung und frei definierbaren Formular zur Erfassung von Nutzerangaben und Einschätzungen.
2. Die HTML Seite, die angezeigt wird, wenn die Bearbeitung einer Aufgabe fortgesetzt wird (z. B. nachdem der Webbrowser einmal geschlossen wurde).
3. Der Abschlusstext mit einem weiteren Formular zur Befragung des Benutzers.

Vor jedem Test werden demografische Daten über ein einfaches HTML-Formular erfasst. Alle genannten Formulare lassen sich mit einem beliebigen HTML-Editor anpassen. Auch die Suchaufgaben werden so erstellt. Dabei ist zu beachten, diese anschließend innerhalb der Konfigurationsdatei für das Addon bekannt zu machen.

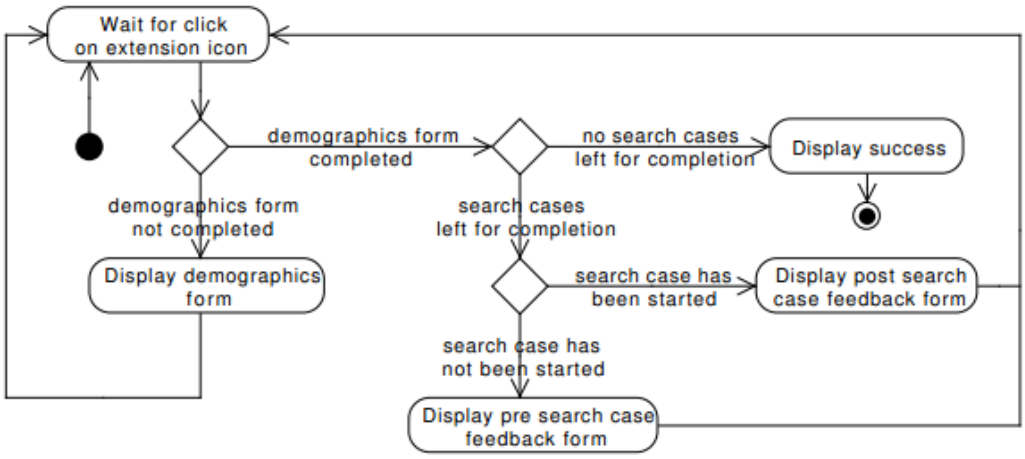


Abbildung 13: Benutzerprozesse im Search Logger (Quelle: Singer et al. 2011, S. 753)

Abbildung 13 stellt die Prozesse des Nutzers dar. Dabei zeigt sich, dass die Bearbeitung von Suchaufgaben erst beginnen kann, wenn das Formular für die demografischen Daten ausgefüllt wurde. Beim Aufruf des Firefox erfolgt automatisch eine Prüfung, ob noch weitere Aufgaben durch die Testperson zu bearbeiten sind. Erst wenn alle Suchaufgaben abgeschlossen sind, erfolgt eine Meldung über den Erfolg.

Die folgende Tabelle zeigt alle Aktionen, die durch den Search Logger erfasst werden können.

Aktivität	**Erklärung**
Suchaufgabe gestartet	Eintrag in die Log-Datenbank, dass eine neue Aufgabe gestartet wurde.
Suchaufgabe beendet	Eintrag in die Log-Datenbank, dass eine Suchaufgabe beendet wurde.
Search Logger gestartet	Beginn mit der Bearbeitung der Suchaufgaben.
Search Logger beendet	Alle Suchaufgaben wurden abgeschlossen.
Webseite besucht	Es wird erfasst, dass die Testperson eine Webseite aufgerufen hat.
Suchanfrage abgesendet	In den Logfiles wird gespeichert, dass der Nuter eine Suche abgeschickt hat (dabei können die Suchmaschinen von denen Anfragen gespeichert werden sollen, festgelegt werden.

Link angeklickt	Der Nutzer hat einen Link angeklickt, z. B. auf der Suchergebnisseite einer Suchmaschine.
Neuer Tab geöffnet	Es wurde im Firefox ein neuer Tab geöffnet.
Tab geschlossen	Es wurde im Firefox ein Tab geschlossen.
Bookmark gesetzt	Die Webseite wurde von der Testperson als Lesezeichen gespeichert.
Bookmark gelöscht	Ein Lesezeichen wurde wieder entfernt.
Text kopiert	Die Zwischenablage wurde verändert.
Search Logger pausiert	Der Nutzer hat den Search Logger deaktiviert.
Einführung eines neuen Benutzers	Es wird festgehalten, dass ein neuer Nutzer mit den Suchaufgaben beginnt.
Demografieformular wird angezeigt	Der Nutzer sieht sich das Formular für die soziodemografischen Daten an.
Demografieformular abgesendet	Der Nutzer hat das Formular für die soziodemografischen Daten abgesendet.
Pre Search Task Formular ausgefüllt	Die Testperson hat das Formular zur Erfassung von Einschätzungen zur Suchaufgabe ausgefüllt.
Post Search Task Formular ausgefüllt	Die Testperson hat das Formular nach Abschluss einer Aufgabe ausgefüllt.

Tabelle 6: Nutzeraktionen, die durch den Search Logger erfasst werden

Die genannten Aktivitäten sind eine wichtige Grundlage für den Prototypen, der im Kapitel 6 dargestellt wird. Besonders relevant für die Untersuchung explorativer Daten ist die Erfassung expliziter Daten durch die Bereitstellung von Formularen. Dadurch können neben der automatischen Generieung von Klickdaten, auch direkte Meinungen und Einschätzungen von den Testpersonen eingeholt und beispielsweise als Grundlage für die Einschätzung und Messung der Zufriedenheit genutzt werden.

5.2.3 Weitere Tools

Neben der vorgestellten Addons gibt es noch weitere Tools, die hauptsächlich für die Erfassung von Nutzerinteraktionen mit Anwendungen entwickelt wurden. Dabei steht meistens keine bestimmte Software im Vordergrund. So wurde der Wrapper von JANSEN 2006 als Desktop-Anwendung konzipiert, um alle Interaktionen mit Anwendungen zu loggen, die eine Relevanz für Informationsspezialisten haben (z. B. auch Microsoft Office oder E-Mail Clients). Das Ziel ist dabei die Untersuchung explorativer Suchsysteme. Ein Nachteil bei diesem System ist aber, dass keine speziellen Suchaufgaben definiert werden können. Ein anderes Browser Plugin ist der Lemur's Query Log Toolbar[6], der für Firefox und für den Internet Explorer angeboten wird. Mit diesem Plugin lassen sich alle Interaktionen im Browser loggen. Eine Nutzung von Suchaufgaben ist aber nicht möglich.

6 Die Toolbar kann unter http://www.lemurproject.org/querylogtoolbar/ heruntergeladen werden.

Ein Tool, dass ähnliche Funktionen, wie der Search Logger anbietet, ist der HCI Browser von CAPRA 2010. Mit diesem ist auch die Gestaltung von Formularen und Suchaufgaben explizit möglich. Der HCI Browser wurde ebenfalls als Firefox-Plugin entwickelt. Im Gegensatz zum Search Logger lassen sich aber hier keine Suchaufgaben abbilden, die nicht in der Bearbeitungszeit beschränkt sind. Langzeitstudien innerhalb von natürlichen Umgebungen können damit nicht durchgeführt werden.

6. Entwicklung des Prototyps

Nachdem in den vorangegangen Kapiteln die wissenschaftlichen Grundlagen und die technischen Tools, die als Basis für die Entwicklung einer neuen Anwendung dienen sollen, vorgestellt und diskutiert wurden, wird in diesem Kapitel die Entwicklung des Prototyps mit den gestellten technischen und funktionalen Anforderungen und allen entwickelten Komponenten erläutert.

Software-Prototypen können verschieden klassifiziert (vgl. POMBERGER / WEINREICH 1994), wobei jede Art andere Ziele verfolgt. Im Rahmen der Masterarbeit wurde die Entwicklung eines funktionalen Prototyps angestrebt, in dem bereits alle wichtigen Komponenten implementiert sind, die eine tatsächliche Anwendung in der Praxis ermöglichen. Funktionale Prototypen enthalten bereits alle grundlegenden Routinen und Komponenten, die das fertige Produkt enthalten soll. Fehlende Bestandteile, die aus dem Prototypen eine vollwertige Software machen würden, werden ebenfalls in diesem Kapitel erläutert.

Der folgende Abschnitt beschreibt die Anforderungen sowohl technisch als auch funktional, die für die Gestaltung des Prototyps wichtig waren. Der fertige Prototyp lässt sich unter

http://electronicpassion.de/Search_Logger/

Mit den Zugangsdaten:

User: search logger
Pass: search logger

nutzen. Die Nutzung wird dabei innerhalb der Diskussion zur Gestaltung des Prototyps gezeigt.

6.1 Technische Anforderungen an den Prototyp

Der neue Prototyp basiert im wesentlichen auf den Technologien, die innerhalb vom Relevance Assessment Tool und beim Search Logger angewendet werden. Die wichtigste Schnittstelle stellen dabei die Datenbanken dar, die beide in MySQL vorliegen und mit Hilfe von SQL-Statements, die in PHP verarbeitet werden, operationalisiert werden.

Für den Prototypen wurden die Tabellen vom Search Logger in die vorhandene Datenbank von RAT integriert, damit eine schneller und einfacher Zugriff auf die anfallenden Daten möglich ist. Technisch gesehen, agieren zwei verschiedene Datenbanken in einem System, da diese praktisch gleiche Anforderungen erfüllen müssen. Für eine Differenzierung der Projektarten (Klassische Retrievaltests und Search Logger Projekte) wurden eigene Datentabellen definiert.

Die Programmierung des Relevanztools und des Search Logger PHP-Frontends zum Senden der Nutzeraktivitäten an die Datenbank sind eine weitere wichtige Schnittstelle. Dadurch sind werden technische Anforderungen in Bezug auf Serveraktivitäten gegeben. So kann direkt abgefragt werden, in welchem Status sich die Bearbeitung einer Aufgabe befindet (z. B. ob diese gestartet wurde oder beendet wurde). Dieser Umstand ist relevant, damit der Prototyp den Befehl erhalten kann, Suchanfragen und Ergebnisse in Bezug auf das Nutzerverhalten zu analysieren und Ergebnisse für den Retrievaltest zusammenzustellen.

Für die tatsächliche Nutzung des Prototyps und der Verbindung von RAT und dem Search Logger ist weiterhin die lokale Installation des Plugins notwendig, da ja wie bereits in Kapitel 5.2 erwähnt, der Search Logger in der aktuellen Version nicht über das Web gesteuert werden kann.

Für die Anwendungsentwicklung des Prototyps mussten im wesentlichen neue PHP Dateien und Webformulare entwickelt werden, die eine Administration von Search Logger Projekten ermöglichen und im Anschluss an bearbeitete Suchaufgaben auch die Bewertung von Suchergebnissen, basierend auf den angefallenen Daten ermöglichen.

Ein Ziel war dabei eine direkte Integration in das Relevance Assessment Tool, damit beide Anwendungen mit Hilfe einer Oberfläche administrierbar sind und auch bei zukünftigen Studien darüber entschieden werden kann, welche Art von Studie durchgeführt werden soll

.

6.2 Anforderungsanalyse des Funktionsumfangs

Bevor eine Anwendung in der Praxis entwickelt wird, werden erstmal die Anforderungen identifiziert und analysiert, die an eine Software gestellt werden. Die Anforderungsanalyse wird häufig in Form einen Lastenhefts durchgeführt, das in der DIN 69901-1 als „vom Auftraggeber festgelegte Gesamtheit der Forderungen an die Lieferungen und Leistungen eines Auftragnehmers innerhalb eines Auftrages“ definiert wird . Das Lastenheft kann in der Praxis unterschiedlich gegliedert sein.. Es enthält aber in jedem Fall eine Beschreibung des Ist-Zustandes, ein Soll-Konzept , Beschreibung von Schnittstellen, Funktionale- und Nichtfunktionale-Anforderungen (z. B. Benutzbarkeit, Zuverlässigkeit, Effizienz, Wartbarkeit) sowie Angaben zur Systemarchitektur und Abnahmekriterien des Auftraggebers.

6.2.1 Der Ist-Zustand

Der Ist-Zustand wurde bereits durch die Beschreibung der Ausgangssituation angezeigt. Im Moment existieren Anwendungen, die innerhalb der Information-Retrieval- und Suchmaschinenevaluierung eingesetzt werden können, aber dabei verschiedene Schwerpunkte setzen. Während auf der einen Seite systemorientierte Evaluierungen im Vordergrund stehen, werden andere Tools eingesetzt, um das Suchverhalten zu evaluieren. Eine Anwendung, die beide Testarten zusammenführt existiert entweder nicht oder ist nicht frei verfügbar.

6.2.2 Das Soll-Konzept

Das Ziel bei der Entwicklung des Prototyps ist die automatische Verknüpfung benutzerorientierter Studien und Tests zur Messung der Retrievaleffektivität. Testdesignern und Administratoren soll es möglich sein innerhalb einer Benutzeroberfläche Projekte anzulegen und zu administrieren als auch Arbeitsaufgaben für die Bearbeitung festzulegen.

Ferner soll eine automatische Analyse der angefallenen, geloggten Benutzerdaten als eine Grundlage für die Generierung des Ergebnissets im Relevanztest nutzbar sein. Die Logfile-Analyse ist das Mittel der Wahl benutzerorientierter Untersuchungsverfahren, da mit der angestrebten Webanwendung Studien auch innerhalb natürlicher Umgebungen (z. B. bei der Testperson zu Hause) durchführbar sein sollen. Weiterhin werden Aufwand und Kosten durch die Logfile-Analyse begrenzt, da weder Videoaufzeichnungen und/oder Eye-Tracking genutzt wird. Der Idealzustand wäre eine Anwendung, die unabhängig von einer lokalen Installation funktioniert und in jeder Umgebung nutzbar wäre.

6.2.3 Funktionale Anforderungen

Die Administration von Projekten soll über jeden herkömmlichen Webbrowser verfügbar sein, damit jeder Zeit Veränderungen durchführbar sind. Für die Speicherung der Benutzerdaten wird ein Browser-Plugin genutzt, dass eine Verwaltung von Suchaufgaben beinhaltet sowie Abfragen zu Nutzereinschätzungen und Meinungen speichern kann.

Daneben soll auf der Nutzerebene ein Interface bereitgestellt werden, dass die Bewertung von Suchergebnissen mit Hilfe vordefinierter Skalen realisiert. Diese Einheitlichkeit ist notwendig für die Datenerfassung.

Die Anwendung muss alle technischen Bedingungen erfüllen, die notwendig sind, um Projekte und Suchaufgaben anzulegen. Es soll eine Auswahl von Suchmaschinen möglich sein sowie ein Suchmaschinenscraper, der die Ergebnisse von den Suchmaschinen aufgrund fest definierter oder von nutzer generiertet Suchanfragen durch Benutzer speichern und für die Relevanztests aufbereiten kann. Weiterhin müssen Funktionen bereitstehen, die auch Einschränkungen für die Tests ermöglichen, z. B. eine Begrenzung von Suchmaschinen, von definierten Suchanfragen und Cutoff-Werten für die Ergebnismenge, aber auch Entscheidungsmöglichkeiten über die Inhalte der Ergebnisse (sollen Trefferbeschreibungen ebenfalls bewertet werden?).

Die funktionalen Anforderungen sehen zusammenfassend folgendermaßen aus:

Anforderung	**Beschreibung**
Bereitstellung eines Administrationsinterfaces zur Gestaltung von Projekten	Die Anwendung benötigt einen standardisierten Zugang für die Entwicklung von Projekten. Dazu werden Eingabemasken für die Datenbankinhalte benötigt, die alle Daten zu den Projekten erfassen und innerhalb der Datenbank speichern.
Optionen zur Entwicklung von Arbeitsaufgaben.	Neben der elementaren Funktion zur Gestaltung von Projekten, muss die Anwendung Mittel zur Verfügung stellen Arbeitsaufgaben zu definieren. Dazu zählt auch die Gestaltung der Fragebögen, die im Test verwendet werden sollen. So wie Angaben über die zu bewertenden Items und eine Begrenzung der Anzahl der Items.
Export-Funktion für die automatische Erstellung der Browser-Extension	Projekte und Suchaufgaben sollten als Gesamtpaket exportierbar sein, damit diese anschließend direkt im Webbrowser installiert werden können, der als Testumgebung dient.

Automatische Zusammenstellung der Ergebnissets für die Retrievaltests	Der Prototyp benötigt einen Suchmaschinenscraper, der Ergebnisse automatisch speichern kann.
Funktionen für eine automatische Analyse der Logdaten	Die Funktionen sorgen dafür, dass die generierten Benutzerdaten ausgewertet werde, um aus den eingegebenen Suchanfragen und geklickten Suchergebnissen Ergebnissets zu generieren.
Verbindung der Suchaufgaben aus der explorativen Suche und dem Retrievaltest	Es muss technisch umgesetzt werden, dass Probanden nach Abschluss einer Suchaufgabe im Browser einen Retrievaltest definierter Ergebnissets durchführen können.
Anpassung der Sprache für die Test	Die Sprache des Tests soll veränderbar sein. Dafür sollten konfigurierbare Vorlagen angeboten werden.
Definition von Skalen für die Bewertung der Suchergebnisse	Für die Retrievaltests muss die Möglichkeit bestehen Skalen festzulegen, auf denen die Ergebnisse bewertet werden können.
Export-Funktion der Ergebnisse aus den Projekten	Mit Hilfe des Exports der Daten, lassen sich diese für Messungen und andere Anwendungszwecke nutzen. Angestrebt ist ein Export der Ergebnisse als Excel-Tabelle.
Systemunabhängigkeit	Bei der Programmierung der Anwendung soll eine weites gehende Unabhängigkeit des Betriebssystems gewährleistet sein.

Tabelle 7: Notwendige Funktionen für den Prototypen

6.2.4 Nichtfunktionale Anforderungen

Der Prototyp sollte in seiner Anwendung, besonders in Bezug auf die Testperson nachvollziehbar und einfach bedienbar sein. Es sollen keine zu hohen kognitiven Voraussetzungen gestellt werden, da für die Tests verschiedenartige Gruppen gewünscht sein könnten. Dieser Umstand muss auch bei der Durchführung des Retrievaltests berücksichtigt werden. Das Nutzerinterface dafür soll selbsterklärend und die zu bewertende Ergebnismenge steuerbar sein.

Bei der Effizienz ist darauf zu achten, dass bestimmte Prozesse (aufwendige Datenbankabfragen, lokale Speicherung von Webdokumenten) nicht zu viel Zeit in Anspruch nehmen.

In Bezug auf die Administrationsebene ist ebenfalls auf eine hohe Benutzbarkeit zu achten. Menüs und Formulare sollten möglichst selbsterklärend sein sowie eine hohe Effizienz gewährleistet sein, damit es nicht zu ständigen Verzögerungen bei der Benutzung kommt.

6.2.5 Systemarchitektur des Prototyps

IDie Systemarchitektur des Prototyps soll eine Erweiterung der Search-Logger Architektur (vgl. SINGER ET AL. 2011, S. 753) durch eine Komponente für Retrievaltests sein. In Abbildung 14 wird die Kombination aus beiden Anwendungen mit allen grundlegenden Bestandteilen dargestellt.

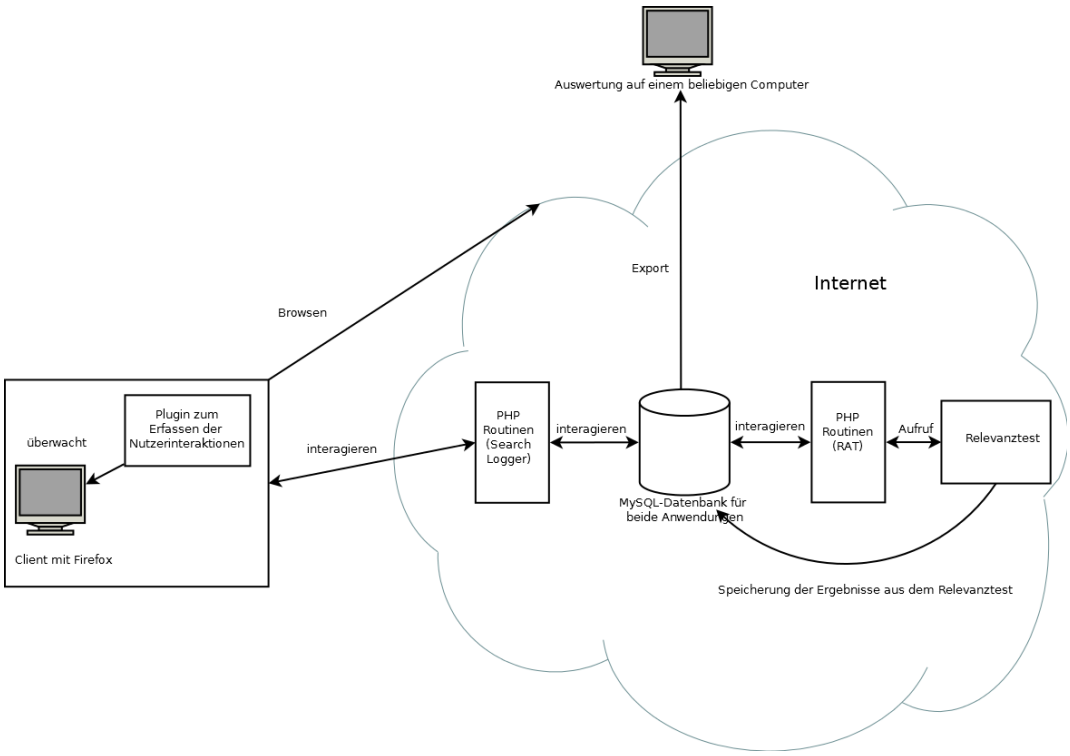


Abbildung 14: Systemarchitektur des Prototyps

Das System basiert, wie bereits oben beschrieben, auf einem Client, auf dem lokal Firefox mit dem Browser-Addon zugegriffen wird. Dieses Plugin überwacht sämtliche Eingaben und Interaktionen des Benutzers und sendet diese an das dafür angepasst PHP-Frontend, das die gesammelten Daten in Echtzeit in der gemeinsamen Datenbank beider Anwendungen speichert. Dabei greifen PHP Routinen vom Search Logger und vom Relevance Assesment Tool parallel auf die Daten zu. Die PHP Routinen vom Search Logger beschränken sich dabei vor allem auf die Verarbeitung und Speicherung der Nutzeraktivitäten.

Wird eine aktuelle Suchaufgabe durch einen Probanden soweit beendet, beginnen die Routinen des Relevanztools auf die Datenbank zuzugreifen und diese zu analysieren. Es werden dabei die geklickten, die manuell aufgerufenen URLs, die Suchanfragen sowie besuchte Suchmaschinen ausgewertet, um daraus elementare Daten für die Zusammenstellung von Suchergebnisses zu generieren. Sind diese Daten erhoben beginnt ein Suchmaschinenscraper im Hintergrund mit dem automatischen Speichern von Suchergebnisses und leitet anschließend zum Nutzerinterface für den Retrievaltest. Wird dieser Test abgeschlossen, kann der Proband, für den Fall, das weitere Arbeitsaufgaben vorliegen, diese bearbeiten so lange bis alle Aufgaben und die dazugehörigen Relevanztests bearbeitet wurden.

Wie die Abbildung zeigt ist eine ständige Verbindung zum Internet und zur Datenbank elementar für die Durchführung der Studien. Da eine ständige Kommunikation zwischen den einzelnen Komponenten stattfinden muss.

Weitere Anforderungen fallen zunächst nicht an, da keine direkten Abnahmekriterien festgelegt werden. Der Prototyp wird im allgemeinen entwickelt, um eine Anwendung anzubieten, die verschiedene Evaluierungsmethoden im Information-Retrieval berücksichtigt.

6.3 Technische Umsetzung

Die technische Realisierung des Prototyps basierte auf den oben genannten Anforderungen an die Funktionalität und auf dem theoretischen Modell, das im Rahmen der Zusammenführung explorativer Suchen und Retrievaltests entwickelt wurde. Als Basis-Design wurden für die Verwaltung der Projekte und Tests die Strukturen und die farbliche Gestaltung von RAT übernommen, damit sich der Prototyp direkt integrieren lässt. Auch die Gestaltung des Search Loggers als Browser-Extension wurde übernommen, da es vor allem zunächst darum ging einen funktionalen Prototypen zu entwickeln.

Am Anfang der technischen Umsetzung und tatsächlichen Programmierung mussten zunächst der Quelltext, die Struktur und Datenbanken beider Anwendungen, aus denen der Prototyp entstehen sollte, analysiert werden, um Schnittstellen zwischen den Anwendungen zu erkennen oder zu schaffen. Basierend auf der Datenbank und den möglichen anfallenden Daten, die für beide Anwendungen benötigt werden, wurde eine Datenbank zusammengestellt, die alle Anforderungen berücksichtigt. Für das Relevance Assessment Tool bestand dabei bereits eine Datenbank zur Speicherung von

Projekten, Suchaufgaben, Benutzern, Skalen, Sprachtemplates und Suchaufgaben. Diese Datenbank wurde als Grundlage verwendet, um eine auf Search Logger Projekte zugeschnittene Datenbank zu designen. Für einen schnellen Zugriff und einer schnellen Verknüpfung, wurde die Entscheidung getroffen alle Tabellen in einer Datenbank zu speichern, auch wenn dies in Bezug auf die Übersichtlichkeit und Wartbarkeit Probleme bereiten könnte. Für den Anwendungszweck des Prototyps ist diese Vorgehensweise aber soweit annehmbar, da zunächst nur Testdaten erfasst werden.

Wie bereits erwähnt basierten beide Anwendungen auf PHP und MySQL-Datenbanken und nutzen daneben auch Javascript und HTML für erweiterte Funktionalitäten.

Javascript und HTML sind für den Search Logger sogar elementar, da Firefox Plugins darauf basieren. Mit HTML werden Suchaufgaben und alle Formulare für Studien im Search Logger gestaltet. Zusammenfassend zeigte sich, dass vor allem folgende Strukturen und Daten innerhalb des Browser-Plugins wichtig sind und für eine gemeinsame Nutzung mit dem Relevance Assessment Tool, angepasst werden mussten:

1. Die Datei mit dem Javascript-Quelltext für den Search Logger: In dieser Datei sind alle Funktionen vorhanden, die für die Erfassung von Nutzeraktionen notwendig sind sowie die Beschriftungen sämtlicher Buttons in der Anwendung.
2. Die Konfigurationsdatei: Innerhalb dieser Datei werden die Suchaufgaben festgelegt, welche Aufgaben bereits bearbeitet wurden, ob die demografischen Daten schon erfasst wurden sowie der Status der Anwendung (pausiert oder nicht) und die Webadresse für die Datenbank zum Speichern der Logdaten.
3. Die HTML-Dateien für die Suchaufgaben: In diesen Dateien sind die Suchaufgaben abgelegt (insgesamt sind es drei Dateien pro Aufgabe).
4. Die PHP Datei zum Senden der Events an die MySQL-Datenbank.

Nach Identifizierung alle notwendigen Dateien für die Verbindung mit dem Relevanztool erfolgte eine Planung über alle notwendigen Komponenten mit entsprechenden Funktionen und Formularen innerhalb der Administrationsoberfläche. Dabei war vor allem zu berücksichtigen, dass die oben genannten Dateien automatisch nach Anforderungen an die Projekte angepasst werden müssen. Es werden im Rahmen der Arbeit nicht alle Eingabemasken ausführlich dargestellt, da diese nach einem wiederkehrenden Prinzip funktionieren.

Der Nutzer gibt etwas ein, die Eingaben werden überprüft und bei Fehlerfreiheit in die Datenbank übertragen. Abbildung 17 auf Seite 96 zeigt ein Beispiel für einen solchen Prozess.

6.3.1 Die Datenbank des Prototyps

Damit eine reibungslose Funktionsweise gegeben sein kann, benötigt der Prototyp eine Datenbank, die sowohl alle Daten für die Projekte (Name, Beschreibung, Suchaufgaben usw.) enthält als auch Tabellen zur Erfassung der anfallenden Protokolldaten und der Relevanzurteile durch die Testpersonen. Abb. 15 zeigt das Datenbankmodell, dass für den Prototypen entwickelt wurde. Die Tabellen sind dabei normalisiert und optimiert (vgl. KOFFLER / ÖGGL 2010, S. 307) damit schnelle und sinnvolle Zugriffe über SQL-Statements möglich sind. Alle in der Datenbank erfassten Informationen, bilden am Ende auch die Grundlage für die Erstellung der Tabellendokumente über die Funktionen der Klasse PHPExcel[7].

7 Dokumentation und Download unter http://phpexcel.codeplex.com/

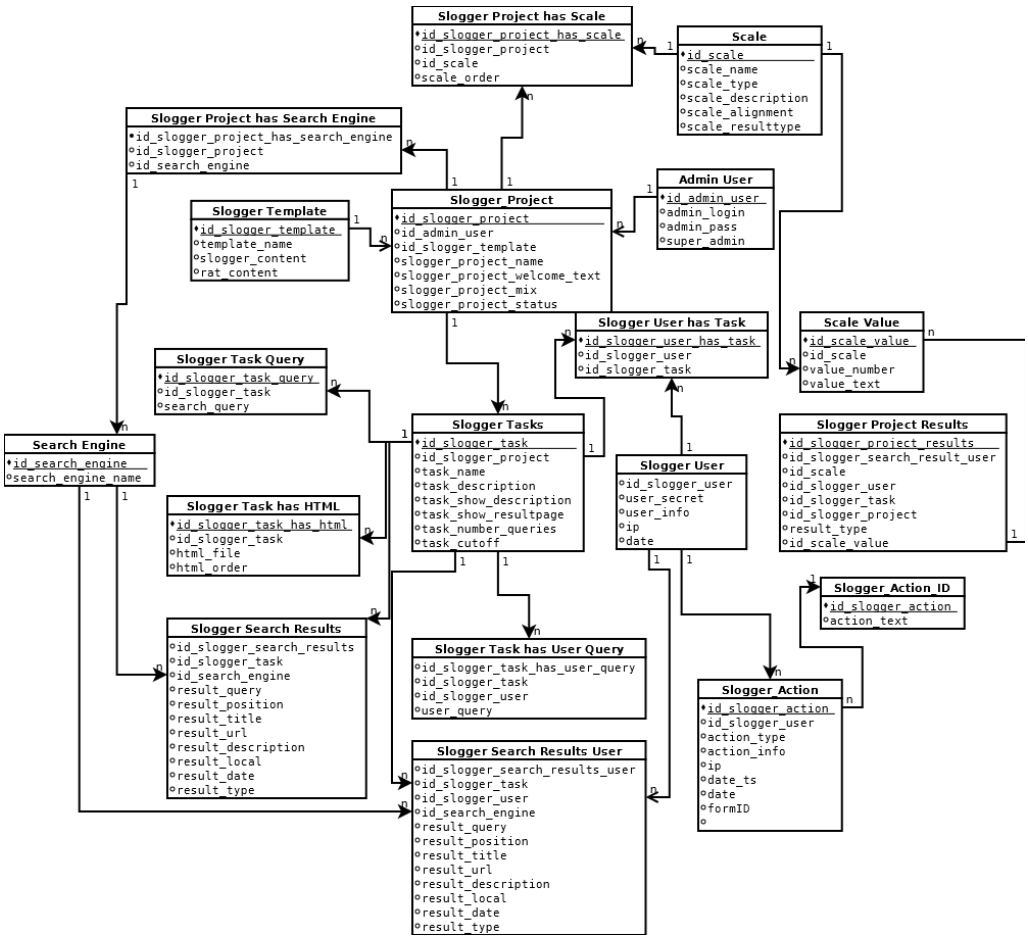


Abbildung 15: Komplette Datenbank der Software

Die Darstellung der Datenbank zeigt die Komplexität der Anwendung. Es bestehen eine Menge Verbindungen zwischen den Tabellen.Im Zentrum der Datenbank steht dabei zunächst das Projekt selber mit den Daten zum Projekttitel, dem Sprachtemplate und den dazugehören Arbeitsaufgaben, die über eine Zwischentabelle zugewiesen werden müssen. Dabei können dem Projekt Daten aus Vorlagentabellen für die Skalen zur Bewertung, der Projektsprache und Suchmaschinen zugewiesen werden, die untersucht werden sollen. Ein Administrator hat dabei die Option mehrere Projekte anzulegen. Eine weitere sehr elementare Tabelle wird für die Suchaufgaben verwendet. Den Aufgaben werden durch zwei Tabellen Suchergebnisse zugewiesen, die sich dadurch unterscheiden, dass sie einerseits über Suchanfragen erstellt werden, die einer Arbeitsaufgabe zugewiesen werden können oder über die Suchanfragen gescraped werden, die ein Nutzer während einer Suchsession eingibt.

Die Suchergebnisse, die durch die Handlungen der Nutzer zur Stande kommen werden mit entsprechenden Scripten generiert.

Die genannten Tabelle sind hauptsächlich für die Verwaltung und Erstellung von Projekten zuständig. Die Tabellen „Slogger Actions“ und „Slogger Project Results“ bilden den Speicher für die Ergebnisdaten, die erhoben und analysiert werden. Die Tabelle „Slogger Project Results“ enthält dabei fast ausschließlich nur Fremdschlüssel aus anderen Tabellen, da die Ergebnisse dynamisch generiert werden.

6.3.2 Der Suchmaschinenscraper

Wie auch bei der Standardnutzung des Relevanztools, ist der Suchmaschinenscraper auch ein sehr wichtiger Bestandteil innerhalb des Prototyps. Im Gegensatz zu dem Anwendungszweck im Relevance Assessment Tool wird dieser hier aber teilweise dynamisch auf Basis von Nutzeraktionen verwendet. Während bei den klassichen Retrievaltests für direkt zu vorher eingegebenen Suchanfragen Ergebnisse gescraped und für Tests technisch aufbereitet werden, müssen im Prototypen in den meisten Fällen direkt nach der Bearbeitung von Suchaufgaben Ergebnissets zu den Suchanfragen und gesichteten Treffern der Testperson zusammengestellt werden. Daneben ist aber auch die Definition von Suchanfragen für Arbeitsaufgaben möglich, die ebenfalls einen direkten Scraping Prozess auslösen. Dadurch zeigt sich, dass der Scraper, wie auch beim Relevanztool eine der wichtigsten Komponenten der Anwendung ist.

Für die automatische Generierung der Dokumentensets für die Probanden im Test war es notwendig PHP Funktionen zu schreiben, die eine Analyse der generierten Interaktionen durchführen und die gesammelten Daten anschließend an den Scraper senden. Dabei werden dann auch verschiedene Einstellungen zu der Aufgabe berücksichtigt, die das Ergebnisset begrenzen. Dazu zählen eine Beschränkung der maximalen Suchanfragen der Testperson und die Reduzierung der Treffer auf einen bestimmten Cutoff-Wert pro Suchmaschine.

Innerhalb einer PHP-Datei wurden die Funktionen integriert, die das automatische Scrapen ermöglichen. Die Funktionen werden in einer fest definierten Reihenfolge folgendermaßen abgearbeitet:

1. Der Benutzer beendet eine Suchaufgabe. Wurde das Post-Formular für das Feedback zur Suchaufgabe abgeschickt, wird die Datei „seach_logger_end.php“ aufgerufen.
2. In der search_logger_end wird die User ID, des Benutzers ermittelt, der die letzte Aufgabe bearbeitet hat. Über eine SQL Abfrage wird die ID des Nutzers ermittelt, der zuletzt einen Event verursacht hat:
   "SELECT max(userID) FROM slogger_actions“
3. Mit Hilfe dieser Information werden anschließen die Projektnummer und die ID der Suchaufgabe ermittelt. Die Projektnummer wird benötigt, um die Suchmaschinen, die für das Projekt zugewiesen wurden, zu identifizieren.
4. Es findet ein Abgleich mit den geloggten Aktionen des Nutzers und den Suchmaschinen statt, um die vom Nutzer eingegebenen Suchanfragen zu finden. Für diesen Vorgang wurden im Prototyp Teile der Suchmaschinen URLs hinterlegt, z. B. sieht der Filter für Google.de so aus:
   http://www.google.de/search?hl=de&q=
   Die Variable *hl* zeigt dabei die Sprache (hier de für deutsch) am und hinter *q*= sind die Suchterme angegeben.
5. Nachdem die Suchanfragen gefiltert und in die Datenbank eingetragen wurden. Werden die Suchergebnisse, die vom Testleiter einer Suchaufgabe zugewiesen wurden, in das Ergebnisset des Nutzers kopiert. Ein Scraping findet nicht statt, da bereits bei der Definition der Suchaufgaben, Ergebnisse lokal gespeichert wurden. Würden die Ergebnisse immer neu gescraped werden, würde dies eine Vergleichbarkeit beeinflussen.
6. Als nächstes werden die URL in das Testset aufgenommen und gescraped, die manuell aufgerufen und angeklickt wurden.
7. Im letzten Schritt werden anschließend die Suchergebnisse geholt, die der Benutzer über die Suchanfragen gefunden hat. Dabei werden die Anfragen an die festgelegten Suchmaschinen geschickt und die festgelegte maximale Anzahl sowie der Cutoff Wert beachtet, damit die Testpersonen nicht zu viele Items bewerten muss.
8. Es erfolgt nun der Aufruf des Nutzerinterfaces „slogger_user_interface.php“ dort werden je nach Präferenzen, dem Benutzer Skalen, Ergebnisdokumente und / oder Trefferbeschreibungen zur Bewertung vorgelegt.

Der Scraper prüft bei allen Scraping Aktionen, ob das aufzurufende Dokument im Netz erreichbar ist. Dabei wird der HTTP-Status Code verwendet, der von Webservern zurückgegeben wird.[8]

Alle Interaktionen mit Webseiten erfolgen dabei auch über die cURL-Library, die im Kapitel 5.1 vorgestellt wurde.

6.3.3 Das Administrationsinterface

Das Administrationsinterface kann, wie bereits bei der Originalversion des Relevance Assessment Tools über jeden herkömmlichen Webbrowser aufgerufen und genutzt werden.Da PHP eine serverbasierte Skriptsprache ist, werden alle Prozesse, die zum Programm gehören in der Serverumgebung ausgeführt, während der Nutzer nur HTML Seiten als Ausgabe erhält. Prüfungen von Eingaben oder auch die Interaktionen mit der MySQL-Datenbank erfolgen nur auf dem Server. Im Gegensatz zu anderen Varianten des Prototyping, wurde in diesem Fall auf ein eigenes Design verzichtet, da für eine zukünftige Nutzung die Integration in das Relevance Assessment Tool vorgesehen ist. Es wurden also die Gestaltung der Formulare sowie alle grafischen Komponenten für den Prototyp übernommen, dies trifft auch auf das Nutzerinterface zu. Daraus ergibt sich auch der Vorteil, dass bei einer Veränderung von RAT, die neue Anwendung stufenlos angepasst wird.

Für die Gestaltung von Search Logger Projekten wurden insgesamt vier wichtige Kategorien identifiziert, die alle bestimmte Funktionen enthalten müssen:

1. Projektverwaltung: Durch die Projektverwaltung soll es dem Projektadministrator möglich sein neue Projekte anzulegen, bestehende Projekte zu verändern, Ergebnisse aus den Projekten zu exportieren, aber auch alle Inhalte eines Projekts direkt als Firefox-Erweiterung herunterzuladen, um diese anschließend zu installieren. Daneben wird hier auch eine Funktion angeboten, um das Formular für die Erfassung demografischer Daten anzupassen. Eine Anleitung dazu befindet sich direkt in der HTML-Datei mit dem Formular.
2. Projektnutzer: Dieses Menü soll sich besonders auf die Interaktionsdaten beziehen. Dem Testleiter soll es hier möglich sein, unabhängig vom Relevanztest Daten zum Nutzerverhalten zu exportieren.

8 Das W3C hat unter http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html eine Übersicht zu möglichen HTTP-Statuscodes.

Innerhalb des Prototyps wurde diese Auswertung zunächst nur auf geklickte und aufgerufene Webseiten und Suchergebnisse sowie Suchanfragen von Benutzern beschränkt.

3. Arbeitsaufgaben: In dem Menü für die Arbeitsaufgaben sollen alle Optionen bereitstehen, die benötigt werden, um Aufgaben für die Probanden zu definieren. Dabei wird vor allem der Suchmaschinenscraper aus dem Relevance Assessment Tool als eine Basistechnologie verwendet. Das Anlegen und Bearbeiten von Aufgaben hat eine direkte Anpassung der HTML-Dateien für die Suchaufgaben zur Folge. Ferner werden hier auch Funktionen angeboten, um Fragebögen zu gestalten, die vor und nach der Bearbeitung einer Suchaufgabe aufgerufen werden.
4. Vorlagen für die Projektsprache: Buttons und Hinweise im Nutzerinterface und in der Browser-Extension sollten anpassbar sein. Daher wurde im Tool im Menü „Templates“ eine neue Option zur Gestaltung von eines Search Logger Templates entwickelt. In der Eingabemaske kann dabei festgelegt werden, welche Beschriftungen der Nutzer bei der Bearbeitung der Suchaufgaben sieht.

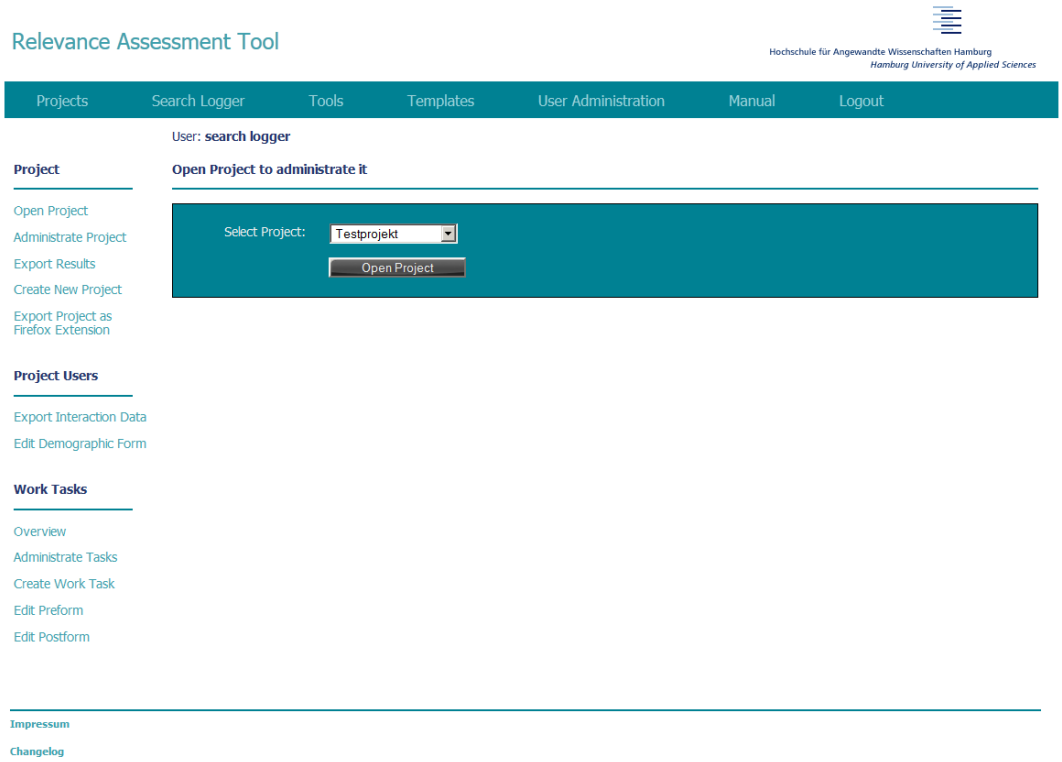


Abbildung 16: Administrationsoberfläche für den Prototypen

Für eine direkte Integration in das Relevance Assesment Tool, wurde der Menüpunkt „Search Logger“ hinzugefügt. Es ist mit dieser Anpassung möglich parallel zu klassischen Retrievaltests, Tests für explorative Suchen und Relevanztests anzulegen.

Technisch umgesetzt wurde das Menü zunächst mit einer PHP Datei mit der Bezeichnung „search_logger.php“ .In dieser Datei befinden sich alle Funktionen zur Überprüfung der Daten aus Formularen sowie der Zugriff auf den Suchmaschinenscraper. Die Eingabemasken der Daten wurden innerhalb einer Ordnerstruktur abgelegt und sind der Übersicht halber getrennt gespeichert. Ruft der Administrator nun beispielsweise das Formular für die Gestaltung eines neues Projekts auf, wird das entsprechende Formular angezeigt. Die Daten- und Eingabeprüfung wird dabei durch die Hauptdatei „search_logger.php“ übernommen. Wurden alle Daten korrekt eingeben, werden diese anschließend in die dahinterliegende Datenbank gespeichert. Routineaufgaben und bestimmte Bildschirmausgaben sind innerhalb von Include Dateien (Klassen und Quelltext Dateien, die wiederkehrende Funktionen enthalten) gespeichert, damit nicht wiederkehrende Operationen ausformuliert werden müssen. Der folgende Ausschnitt zeigt eine Überprüfung über Einträge in einer Datenbank. Als Bedingung wird kontrolliert, ob bereits Suchmaschinen, die für Projekte genutzt werden können, in der Datenbank gespeichert sind. Ist dies nicht der Fall, soll eine Fehlermeldung den Nutzer darüber informieren. Ist die Bedingung erfüllt, können Suchmaschinen für das Projekt ausgewählt werden.

```
// Test if some search engines saved in the database
        if (db_items_exist("search_engine")==false)
        {
           $html_page -> con_header("Create Project");
           echo 'Currently are no Search Engines saved in the database! 
<br/><br/>Please add them to the Database!';
           break;
        }
```

Zu den Besonderheiten bei der Entwicklung der Projektumgebung zählen vor allem die Möglichkeit das Projekt direkt als Erweiterung zu exportieren sowie die Gestaltung von Suchaufgaben und automatische Anpassungen innerhalb des Browser-Plugins.

Firefox-Plugins sind technisch betrachtet einfache Javascript Anwendungen in Verbindung mit HTML und CSS-Dateien, die als Zip-Dateien in die Browserumgebung installiert werden können. Die Aufgabe für den Prototypen bestand nun darin, einen Weg zu finden, angelegte Projekte so zu exportieren, dass diese auch funktionsfähig in einen Firefox Webbrowser installiert werden können. Zu diesem Zweck wurde eine Klasse installiert, die ein Zip Datei aus Daten erzeugen kann, die auf einem Webspace gespeichert sind[9]. Nach einigen Anpassungen konnte diese Klasse genutzt werden, um aus einem Ordner mit den Projektdaten eine funktionsfähige Firefox-Erweiterung zu generieren. Die Projektdaten werden dabei automatisch aus dem Administrationsinterface basierend auf einem Template für die Sprache des Projekts, den Arbeitsaufgaben, definierte Fragebögen, zusammengestellt. Legt ein Projektadministrator ein Projekt im System an, werden diese Daten auf Basis eines leeren Search Logger Projekts erzeugt.

9 Die Klasse kann unter http://www.phpclasses.org/package/2322-PHP-Create-ZIP-file-archives-and-serve-for-download.html heruntergeladen werden.

Der Prozess sieht dabei folgendermaßen aus:

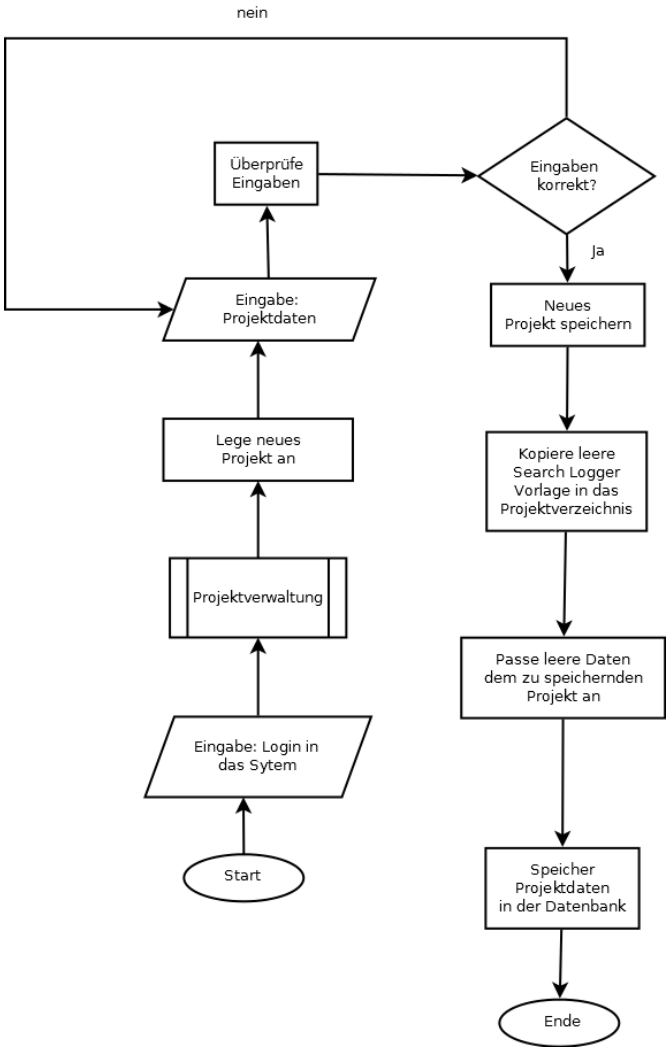


Abbildung 17: Prozess bei der Eingabe der Projektdaten für ein neues Projekt

Erst wenn alle Daten korrekt eingegeben wurden, werden alle Operationen auf Datei und Datenbankebene durchgeführt. Überprüfungen der Daten beziehen sich vor allem auf leere Eingabefelder oder Buchstaben in Zahlenfeldern usw. weiter wird aber auch geprüft, ob Skalen oder Suchmaschinen für das Projekt festgelegt wurden.

Die Dateien für das Projekt in dem jeweiligen Projektverzeichnis werden bei Veränderungen am Projekt oder den Suchaufgaben sowie den Fragebögen verändert. Für diese Veränderung wurde eine eigene Funktion „edit_text_file“ geschrieben, die nach bestimmten Inhalten in einer Text-Datei sucht und diese durch eine festgelegte Zeichenkette ersetzt. Die Funktion enthält dabei Parameter zur Lokalisation der Datei, den gesuchten Text und den neuen Text:

function edit_text_file($target_file, $search, $replace)

Damit gewährleistet ist, dass die neuen Texte an den gewünschten Positionen gespeichert werden, wurden in den Vorlagen für ein Projekt Tags in Form von HTML-Kommentaren gesetzt:

<!-- Edit Task Name here -->

Die HTML-Dateien für die Suchaufgaben werden hingegen dynamisch erzeugt. Dafür werden die Informationen, die innerhalb der Eingabefelder für die Aufgaben definiert wurden als Ausgangspunkt verwendet. Search Logger Search Cases setzen sich aus insgesamt drei Dateien zusammen, die folgendermaßen benannt werden:

- 1_search_case_1.html (Datei für den Beginn einer Suchaufgabe und für den Fragebogen, der vor der Bearbeitung ausgefüllt werden soll.
- 1_search_case_1_2.html (Datei mit Hinweis über die Suchaufgabe, nachdem diese zwischendurch pausiert wurde.)
- 1_search_case_1_3.html (Datei für den Abschluss einer Suchaufgabe mit Fragebogen zur Ermittlung von Nutzereindrücken nach der Bearbeitung einer Aufgabe.)

Für eine bessere Zuordnung wird durch den Prototypen die Benennung dahin geändert, dass „search_case“ durch den Namen der Aufgabe und die „1_“ durch die Datenbank-ID der Suchaufgabe ersetzt wird. Neben den Formularen in den einzelnen Dateien der Arbeitsaufgaben, befindet sich auch in jedem Formular ein Teil in Javascript, der den Zugriff auf den Search Logger sowie den Zugang zum PHP Frontend steuert, damit alle anfallenden Logdaten direkt in der MySQL-Datenbank abgelegt werden können.

Bei den Suchaufgaben selber stehen dem Projektverwalter eine Reihe von Eingabefeldern und Optionen zur Verfügung, die kurz aufgeführt werden:

- Work Taks Name: Festlegung eines Namens für die Suchaufgabe, die auch von der Testperson gesehen wird.
- Work Task Description: Beschreibung der Arbeitsaufgabe. Hier können Informationsbedürfnisse oder auch Hinweise zu relevanten Dokumenten und ähnliches festgelegt werden.
- Max Queries per Search Engine: Durch diese Option wird eine erste Beschränkung für die maximale Anzahl der zu beurteilenden Suchergebnisse eingetragen. Wird beispielsweise eine zwei als Maximum verwendet, werden bevor der Retrievaltest durchgeführt wird nur zwei Anfragen der Testperson automatisch an die Suchmaschinen im Test geschickt und anschließend die Ergebnisse gesichert.
- Cutoff: Hier kann der Administrator festlegen, wie viele pro Suchmaschine bewertet werden soll. Diese Option ist die zweite Möglichkeit die maximale Anzahl an Treffern einzuschränken.
- Number of Queries: Im Gegensatz zu den Max Queries per Search Engine wird hier eingestellt, wie viele selbst definierte Suchanfragen Ausgangspunkt für den Relevanztest sein sollen.
- Description or Resultpage: Hier lest sich festlegen, ob auch die Trefferbeschreibungen oder auch nru die Suchergebnisse selber Gegenstand im Retrievaltest sein sollen.

Neben allen Eingabemöglichkeiten bietet das Administrationsinterface auch Funktionen zur Ausgabe von Projektergebnissen.

4	User performed a search	Location: http://www.google.de/search?hl=de&source=hp&q=angelika&gbv=2&oq=angelika&aq=f&aqi=g4g-s1g5&aql=&gs_s... Title: angelika - Google-Suche	79.246.77.64	2012-02-27 16:45:31	1_merkel_1
4	User performed a search	Location: http://www.google.de/... Title: Google	79.246.77.64	2012-02-27 16:45:29	1_merkel_1
4	User opened a tab	User opened a tab	79.246.77.64	2012-02-27 16:45:27	1_merkel_1
4	User viewed a related link	Location: about:blank... Title:	79.246.77.64	2012-02-27 16:45:27	1_merkel_1
4	User viewed a related link	Location: http://www.spiegel.de/... Title: SPIEGEL ONLINE - Nachrichten	79.246.77.64	2012-02-27 16:45:21	1_merkel_1

Abbildung 18: Auszug aus den Logfiles

Abbildung 18 zeigt die Form, in der das Search Logger-Plugin die Daten erfasst und in der Datenbank speichert. Links in der ersten Spalte ist die User ID, daneben die durchgeführte Aktion. Der Search Logger kann erkennen, wann der Nutzer eine Suchanfrage durchgeführt hat und welche Seite darauffolgend besucht wurde. Neben diesen Daten werden noch die IP-

Adresse, das Datum und die Suchaufgabe geloggt. Zusammen mit den Bewertungen aus dem Retrievaltest werden so die erhobenen Daten generiert und zusammenfasst und lassen sich in Form eines Excel-Dokuments zurückgeben und können für weitere Auswertungen exportiert werden.

User ID:	Demographics:	Pre Questionnaire:	Post Questionnaire:							
4	a=(35))RB(b=(2))RB(RB(d=(1))RB(e=(2))I (k=(3))RB(l=(3))RB(I RB(n=(3))RB(o=(3))I (p=(3))RB(q=(3))RB(	(a=(1))(b=(1))(c=(1)	(a=(1))(b=(1))(c=(1))(d=(1))							
Work Task:merkel										
Started:	Finished:	Search Queries:								
2012-02-27 16:45:06	2012-02-27 16:45:47	angelika								
Scraped Result URLs: Testprojekt										
ID:	User:	Search Task:	Search Query:	Search Engine:	Position:	URL:	Result Date:	Comment:	Viewed by U	Scale Binärskala
1	2	merkel	angela merke	bing.de	1	http://angela-merkel.de/	2012-02-27	definied		ja
2	2	merkel	angela merke	bing.de	2	http://de.wikipedia.org/w	2012-02-27	definied		ja
3	2	merkel	angela merke	bing.de	3	http://www.bundeskanzl	2012-02-27	definied		ja
4	2	merkel	angela merke	google.de	1	http://de.wikipedia.org/w	2012-02-27	definied		ja
5	2	merkel	angela merke	google.de	2	http://www.angela-merk	2012-02-27	definied		ja
6	2	merkel	angela merke	google.de	3	http://www.bundeskanzl /Angela-Merkel/angela-m	2012-02-27	definied		nein
7	2	merkel	schuhe	bing.de	1	http://www.schuhe.de/	2012-02-27	scraped		nein
8	2	merkel	schuhe	bing.de	2	http://de.wikipedia.org/w	2012-02-27	scraped		nein
9	2	merkel	schuhe	bing.de	3	http://www.zalando.de/	2012-02-27	scraped		nein
10	2	merkel	schuhe	google.de	1	http://www.goertz.de/	2012-02-27	scraped		nein
11	2	merkel	schuhe	google.de	2	http://www.imwalking.de	2012-02-27	scraped		nein
12	2	merkel	schuhe	google.de	3	http://www.zalando.de/	2012-02-27	scraped		nein
13	2	wulff	wulff	bing.de	1	http://de.wikipedia.org/w	2012-02-27	definied		ja
14	2	wulff	wulff	bing.de	2	http://www.wulff-gmbh.c	2012-02-27	definied		nein
15	2	wulff	wulff	bing.de	3	http://www.bundespraes	2012-02-27	definied		ja
16	2	wulff	wulff	google.de	1	http://www.spiegel.de/p	2012-02-27	definied		ja
17	2	wulff	wulff	google.de	2	http://de.wikipedia.org/w	2012-02-27	definied		ja
18	2	wulff	wulff	google.de	3	http://www.focus.de/pol Schwiegermutterstuhl_aic	2012-02-27	definied		ja
19	2	wulff	kerzen	bing.de	1	http://de.wikipedia.org/w	2012-02-27	scraped		nein
20	2	wulff	kerzen	bing.de	2	http://www.kerzen-gesch	2012-02-27	scraped		nein
21	2	wulff	kerzen	bing.de	3	http://www.engels-kerze	2012-02-27	scraped	2	nein
22	2	wulff	kerzen	google.de	1	http://www.engels-kerze	2012-02-27	scraped	2	nein
23	2	wulff	kerzen	google.de	2	http://www.amazon.de/K	2012-02-27	scraped		not_reached
24	2	wulff	kerzen	google.de	3	http://www.kuhlmann-m	2012-02-27	scraped		nein
25	4	merkel	angela merke	bing.de	1	http://angela-merkel.de/	2012-02-27	definied		nein
26	4	merkel	angela merke	bing.de	2	http://de.wikipedia.org/w	2012-02-27	definied		ja
27	4	merkel	angela merke	bing.de	3	http://www.bundeskanzl	2012-02-27	definied		ja

Abbildung 19: Auszug aus einer generierten Ergebnisdatei

Der Testleiter sieht innerhalb der Daten, welcher Nutzer, welche Dokumente geklickt, welche Suchanfragen durchgeführt, wie die Fragebögen ausgefüllt wurden, welche Webseiten dieser manuell aufgerufen hat, wie oft etwas in die Zwischenablage kopiert, wie häufig Webseiten als Lesezeichen gespeichert wurden und wie er schließlich im Retrievaltest die Search Items explizit bewertet hat. Es wird insgesamt eine Übersicht zu implizierten Feedback, dem Suchverhalten des Nutzers und expliziten Feedback über den Relevanztest angeboten.

6.3.4 Die Nutzerumgebung

In der folgenden Darstellung wird die Nutzerumgebung anhand eines groben Ablaufs einer Search-Logger-Session mit Erläuterungen zu der technischen Realisierung aufgezeigt. Im Gegensatz zum Administrationsinterface, das über eine Serverumgebung aufgerufen wird, ist wie bereits erwähnt, die Nutzung der Browser-Extension abhängig von der lokalen Umgebung. Das hängt dabei auch von dem Umstand ab, dass sämtliche Daten, die zum Search Logger gehören, lokal vorliegen müssen. Dazu gehören die Arbeitsaufgaben, die in einfachen HTML Dateien abgelegt sind sowie die Konfigurationsdateien, die teilweise in XML vorliegen aber auch die Anwendung selber, die in Javascript geschrieben ist. Die Installation des Search Loggers wird innerhalb des ausgewählten Firefox-Profils durchgeführt. Wie bereits gezeigt, werden diese Dateien durch die Gestaltung der Projekte im Administrationsinterface verwaltet. Sämtliche Buttonbeschriftungen, Hinweise zu den Suchaufgaben sowie weitere Gestaltungselemente im Firefox-Plugin werden mit Hilfe von Daten aus der MySQL-Datenbank und mit Funktionen zur Manipulierung der HTML und XML-Dateien sowie der Javascript-Anwendung verändert. Werden Veränderungen am Projekt oder den Aufgaben vorgenommen ist ein erneuter Export des Projekts elementar sowie eine Neuinstallation des Addons.

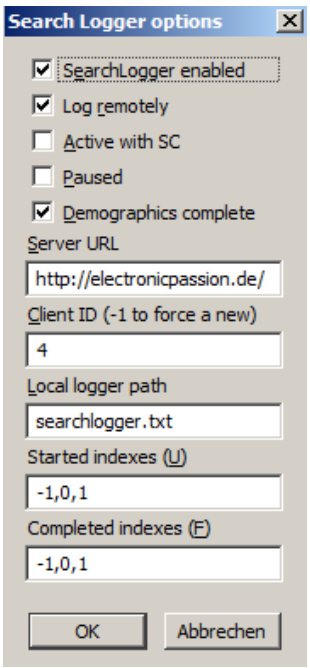


Abbildung 20: Search Logger Optionen innerhalb von Firefox

Ferner ist es auch notwendig bestimmte Einstellungen im Plugin vorzunehmen, wenn beispielsweise Suchaufgaben erneut bearbeitet werden sollen oder auch neue Nutzer den Search-Logger verwenden möchten. Dem Projektleiter stehen dabei folgende Optionen zur Verfügung:

- Search Logger enable: Angabe darüber, ob der Search Logger aktiv sein soll.
- Log remotely: Einstellung, ob neben einer lokalen Speicherung der Interaktionen, auch ein Logging in der Datenbank stattfinden soll. Diese Einstellung ist elementar für den Prototypen.
- Active with SC: Durch diese Einstellung wird dem Search Logger mitgeteilt, dass der Nutzer des Browsers noch aktiv an einer Aufgabe arbeitet. Der Search Logger benötigt diese Angabe, damit mit dem Aufruf keine neue Aufgabe zugewiesen wird.

- Paused: Durch diese Einstellung kann der Search Logger zwischendurch pausiert werden. Alle Eingaben und Browserinteraktionen werden dann nicht geloggt.
- Server-URL: Hier muss die URL eingetragen werden, auf der die Datenbank zum Loggen installiert wurde.
- ClientID: Die ClientID ist in der Datenbank gleichzeitig die NutzerID. Der Vorteil an der lokalen Einstellung im Addon ist, dass so Benutzer des Browser bei festgelegter ID wiedererkannt werden. Diese Einstellung ist nötig für Aufgaben, die einen längere Bearbeitungszeit benötigen.
- Local logger path: Pfad zur lokalen Textdatei, in der die Aktionen geloggt werden.
- Started Indexes (U): Jede Suchaufgabe erhält einen Index in einem Array, über den die Aufgabe angesprochen wird. Damit die Reihenfolge der Bearbeitung der Suchaufgaben zurückgesetzt wird, muss hier als Wert -1 eingetragen werden.
- Completed Indexes (F): Hier ist angegeben, welche Suchaufgaben schon bearbeitet wurden. Auch hier stellt der Wert nur die Position im Array dar. Die IDs der Suchaufgaben in der Datenbank sind unabhängig von diesem Wert.

Wenn ein Projekt mit Suchaufgaben angelegt und für den Einsatz im Firefox exportiert und installiert wurde, kann die Studie beginnen. Dabei ruft der Benutzer über den Browser zunächst den Search Logger auf. Diese Optionen lassen sich in einer XML-Datei anpassen, wobei dann auch entsprechende

Funktionen in der Javascript-Anwendung vom Search Logger für neue Optionen geschrieben werden müssen.

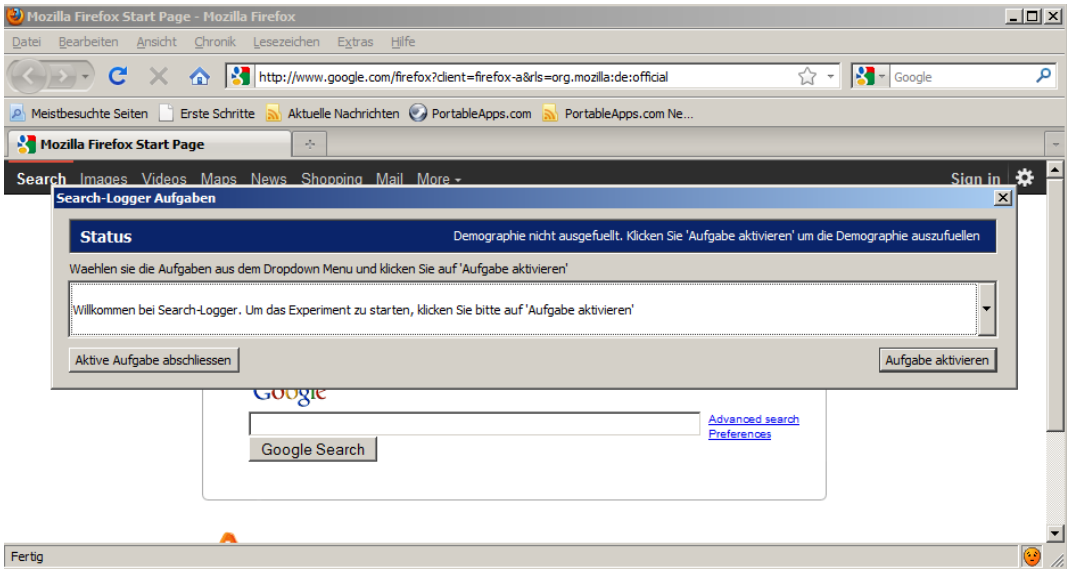


Abbildung 21: Begrüßungstext im Search Logger

Der Search Logger wird durch einen blinkenden Smiley gestartet, der unten rechts im Browserfenster platziert ist. Klickt die Testperson den Smiley an, öffnet sich das Fenster mit einem Begrüßungstext und Buttons, um mit der Aufgabe zu beginnen. Begrüßungstext und Button sind dabei durch das oben genannte Search Logger Template beschriftet. Entscheidet sich der Nutzer dafür, mit dem Test zu beginnen, wird er aufgefordert demografische Daten in einem Formular einzugeben. Alle Browser-Interaktionen werden ab diesem Zeitpunkt aufgezeichnet und in einer Datenbank verarbeitet. Technisch betrachtet wird bei jeder Aktion im Webbrowser immer wieder eine PHP Datei mit Parametern aufgerufen, die eine direkte Speicherung aller Events realisiert. Der Aufruf erfolgt dabei mit folgendem Programmcode:

```
gXMLHttpRequest.open("GET", SERVER_URL + "p.php?" + clientID + "{/}" +
cSecret + "{/}" + type + "{/}" + tabLocation + "{/}" + tabTitle + formID);
```

Mit der Methode XMLHttpRequest.open wird jedesmal (solange der Search Logger aktiviert ist) die aktuelle Webseite erfasst und mit definierten Variablen an die p.php Datei gesendet, in der dann über die GET-Methode (dabei werden Variablen die mit Aufruf einer URL versendet werden) ausgewertet und in einem SQL Statement in die Datenbank zum Loggen der Nutzerhandlungen geschrieben. Folgende Variablen werden dabei übergeben:

clientID = Die aktuell zugewiesene Nutzer-ID der Testperson. Diese bleibt erhalten, wenn die Testperson den Browser schließt und wieder öffnet, um eine Aufgabe fortzusetten.

- cSecret = Ein generierter Wert, der die Eindeutigkeit des Nutzers gewährleistet.
- type = Hier wird übermittelt, welche Aktion durchgeführt wurde.
- tabLocation = Ist die aktuelle URL, die aufgerufen wurde.
- tabTitle = Der Titel des Firefox Tabs.
- formID = Ist der Name der aktiven Suchaufgabe.

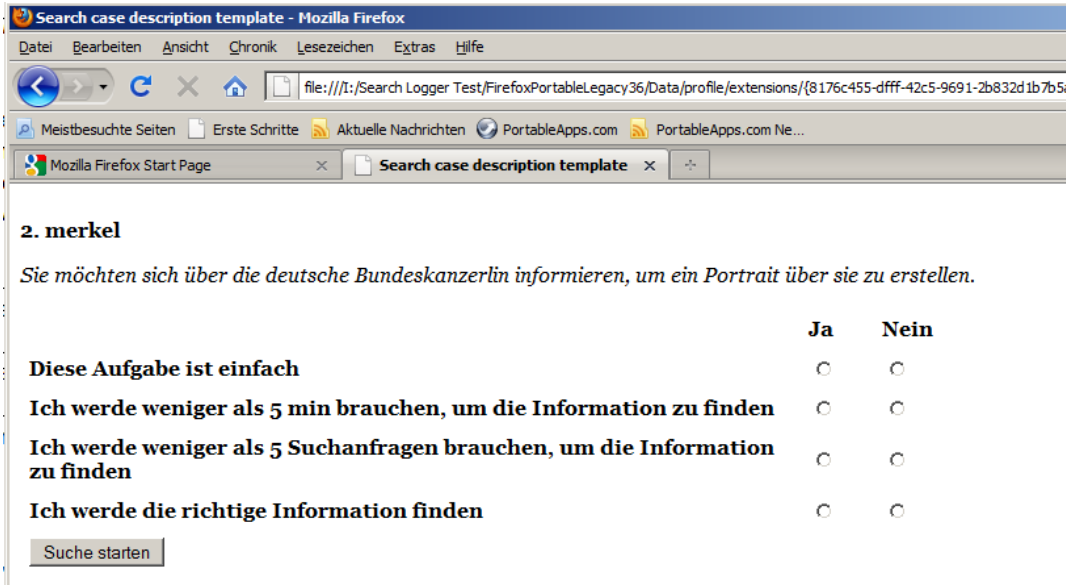


Abbildung 22: Beispiel für einen Fragebogen vor der Bearbeitung einer Suchaufgabe

Wurde das Demografie-Formular ausgefüllt, beginnt die erste Suchaufgabe. Der Proband wird aufgefordert einen Fragebogen zur Aufgabe auszufüllen, bevor dieser durch einen Klick auf den Button eine weitere Webseite aufruft, die mitteilt, dass dieser die Aufgabe begonnen hat und jetzt mit allen Mitteln, die er nutzen möchte, bearbeiten kann.

Ist die Aufgabe soweit bearbeitet, muss diese mit Hilfe des Smiley im Browser vorerst beendet werden. Daraufhin folgt der Fragebogen, der nach Bearbeitung einer Aufgabe ausgefüllt werden soll. Durch den Click auf den Button unterhalb wird der Nutzer wieder weitergeleitet und sieht nun einen Hinweis darüber, dass sich in Kürze eine Seite öffnet, auf der Suchergebnisse, passend zur Aufgabe explizit bewertet werden können.

Abbildung 23: Weiterleitungsseite auf den Retrievaltest

Während die Ladeanimation in diesem Fenster arbeitet, werden im Hintergrund der Anwendung die notwendigen PHP-Routinen durchgeführt, um die Suchergebnisse für den Relevanztest zu generieren (siehe Kapitel 6.3.2 zum Suchmaschinenscraper).

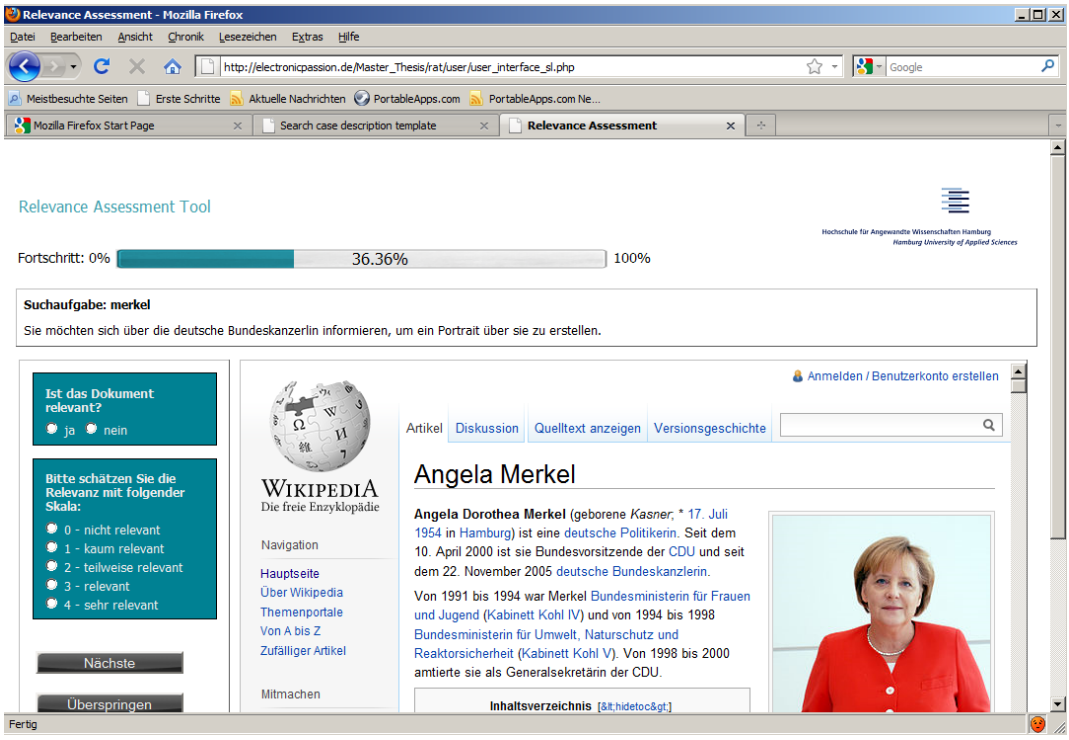


Abbildung 24: Nutzerinterface des Prototyps

Wenn die Suchaufgabe komplett abgeschlossen wurde, erscheint eine Hinweisseite, dass nun eine weitere Aufgabe bearbeitet werden kann. Alle genannten Prozesse wiederholen sich so lange, bis alle Arbeitsaufgaben bearbeitet wurden.

Sind alle Aufgaben bearbeitet, sieht der Nutzer eine Seite dazu, die ihm mitteilt, dass das Experiment abgeschlossen ist. Finden die Tests innerhalb eines Labors statt, müssen nun die lokalen Einstellungen des Search Loggers zurückgesetzt werden (siehe Seite xxx).

6.4 Abgleich mit dem theoretischen Modell aus Kapitel 4.2

Es werden hier noch einmal die Schritte aus dem theoretischen Modell für den funktionalen Prototyp in Bezug auf die technische Umsetzung diskutiert.

6.4.1 Auswahl der Suchmaschinen

Am Anfang einer Studie zur Messung der Qualität der Suchergebnisse steht die Entscheidung über die Auswahl der Suchdienste und Suchmaschinen, die evaluiert werden sollen. Innerhalb des Prototyps lassen sich die Suchmaschinen bei der Gestaltung von Projekten auswählen.

Diese Auswahl bezieht sich dann vor allem auf die Entscheidung von welcher Suchmaschine Suchergebnisse automatisch für den Retrievaltest durch den Suchmaschinenscraper gespeichert werden sollen. Damit direkte Vergleiche zwischen Suchanfragen möglich sind. Kann der Projektleiter den Suchaufgaben verschiedene Anfragen direkt zuweisen.

6.4.2 Definition der Zielgruppe

Die Auswahl der Zielgruppe ist keine Entscheidung, die durch die Technologie des Prototyps abgenommen werden kann. Es stehen keine direkten Funktionen zur Verfügung, die diesen Prozess unterstützen. Allerdings hilft die Abfrage demografischer Daten dabei zumindest Personen bestimmten Zielgruppen zuzuordnen. Der Fragebogen kann durch das Tool für bestimmte Daten angepasst werden.

6.4.3 Festlegung des Untersuchungszeitraums, der Untersuchungsumgebung und die Auswahl der Testpersonen

Der Vorteil des vorgestellten Prototyps ist, dass dieser unabhängig von Untersuchungszeitraum und Untersuchungsumgebung genutzt werden kann. Die Testperson muss das Plugin nur auf einem Computer seiner Wahl installieren und kann anschließend direkt mit der Bearbeitung der Testaufgaben beginnen. Da die Daten zentral in einer Datenbank in einer Serverumgebung gespeichert werden, lassen diese sich jederzeit abrufen.

6.4.4 Gestaltung von Pre- und Post-Fragebögen

Bereits der Search Logger in der Einzelvariante bietet eine Möglichkeit, die bei vielen Anwendungen zur Untersuchung des Nutzerverhaltens bei

Suchmaschinen fehlt. Es können sowohl Fragebögen vor der Bearbeitung von Aufgaben als auch danach über HTML-Seiten zum Ausfüllen angeboten werden. Diese Fragebögen helfen dabei Einschätzungen, Meinungen und Daten zur Zufriedenheit der Testpersonen zu gewinnen. Diese können dabei helfen, z. B. zu untersuchen inwieweit Relevanzbewertungen und Zufriedenheit zusammenhängen (wie beispielsweise bei FOX ET AL. 2005 und HUFFMAN / HOCHSTER 2007 untersucht).

Der Prototyp bietet Funktionen für die Bearbeitung dieser Fragebögen innerhalb von Projekten. Allerdings sind einfache HTML-Kenntnisse notwendig, um diese zu bearbeiten.

6.4.5 Gestaltung von simulierten Arbeitsaufgaben

Im Prototypen ist es möglich und notwendig Arbeitsaufgaben für die Projekte zu erstellen. Dabei können sowohl Beschreibung und Name als auch explizit genannte Suchanfragen festgelegt werden. Die Daten zu den Aufgaben bilden die Grundlage für die Erfassung der Nutzeraktionen im Test als auch für die Schaffung von Ergebnissets zur Bewertung im Relevanztest.

6.4.6 Festlegung von Skalen

Die Definition von Skalen wurde direkt aus dem Relevance Assessment Tool übernommen, da keine besonderen Anpassungen für den Prototyp notwendig waren. Skalen lassen sich innerhalb der Anwendung zum einen durch den Dokumententyp (Trefferbeschreibung, Trefferseite) und zum anderen durch den Fragetyp (offen, geschlossen) festlegen.

Bei geschlossenen Skalen können anklickbare Werte vorgegeben werden, wie es beispielsweise bei 5er-Skalen üblich ist. Eine geschlossene Skala könnte dabei folgende Werte vorgeben, wenn nach dem Grad der Relevanz eines Dokuments gefragt wird:

- 0 = nicht relevant
- 1 = wenig relevant
- 2 = teilweise relevanten
- 3 = relevant
- 4 = sehr relevant

Bei offenen Skalentypen im Prototypen ist es möglich minimal und maximal Werte, z. B. für die Abfrage einer Skale mit Prozentwerten, zu definieren oder einfach ein Textfeld für eine einfache Eingabe anzubieten.

6.4.7 Analyse der explorativen Suche

Auf die Möglichkeiten der automatischen Analyse der Daten aus der explorativen Suche wurde oben, besonders im Bezug auf das Nutzerinterface, eingegangen. Der Prototyp bietet in diesem Fall Mechanismen, zur automatischen Auswertung der Interaktionsdaten. Zum einen für die Schaffung der Ergebnisse für den Retrievaltest und zum anderen für die Zusammenfassung und Aufbereitung der Projektergebnisse.

6.4.8 Zusammenführung und Analyse der Daten aus den explorativen Suchen und den Retrievaltests

In der Anwendung ist es möglich die Ergebnisse aus den explorativen Suchen und den Retrievaltests als Excel-Dokumente auszugeben. Diese Daten lassen sich nun mit Hilfe verschiedener Kennzahlen auswerten. Zu Gründen der Übersicht können auch getrennte Tabellen zu den Nutzeraktionen und zu den Nutzeraktionen in Verbindung mit den Relevanzbewertungen exportiert werden.

6.5 Grenzen des Prototyps

In der jetzigen Entwicklungsstufe gibt es einige technische und funktionale Grenzen, die zusammenfassend aufgezeigt werden in diesem Abschnitt dargestellt werden.

6.5.1 Technische Grenzen

Zu den ersten technischen Einschränkungen des Prototyps zählen in der jetzigen Form vor allem die Beschränkung des Search Loggers auf den Firefox 3.6 als auch Probleme, die bei der Erfassung von Google entstehen, wenn Google-Instant aktiviert ist. Google-Instant ist eine Funktion von Google, die eine Ausgabe von Suchergebnissen bereits während der Eingabe ermöglicht. Dabei werden keine vollständigen URLs gebildet, so dass eine Erfassung der Suchanfrage nicht möglich ist. Der Firefox, der innerhalb der Prototyp-Umgebung heruntergeladen werden kann, wurde daher auf einen User Agent gestellt, der dafür sorgt, dass Google Instant deaktiviert ist. Bewegt man sich in der angebotenen Firefox Version im Netz, erkennen die

Webseiten einem als „Google-Bot“ und für diesen ist Google Instant deaktiviert. Über die manuelle Ausschaltung von Google Instant kann ebenfalls erreicht werden, dass der Search Logger fehlerfrei Google Anfragen erkennt und speichern kann.

Eine weitere Einschränkung durch die lokale Umgebung wurde bereits erwähnt. So ist es nun immer notwendig den Search Logger auf einem Compuer vor Ort zu installieren, damit das Search Logger Plugin und damit auch der Prototyp genutzt werden kann. Für die Zukunft wäre eine Variante erstrebenswert, die auch über das Web oder einem Proxy gesteuert wird. Entwicklungen an einer Proxy-Variante des Search Loggers laufen bereits.

Innerhalb von RAT bestehen daneben technische Einschränkungen vor allem im Bezug auf die Auswertung der angefallenen Daten. So bestehen bisher keine Mechanismen einer automatischen Datenauswertung mit definierten Kennzahlen. Daneben können Relevanztests nun jederzeit abgebrochen werden und dadurch gehen die Daten daraus verloren. Diese Einschränkung kann zu hohen Datenverlusten führen. Der Suchmaschinenscraper im Prototypen ist bisher nur für die Erfassung organischer Suchergebnisse von Suchmaschinen programmiert. Dadurch gehen andere Bestandteile der Trefferliste, wie Universal Search Ergebnisse verloren.Weiter ist der Suchmaschinenscraper auf eine häufige und regelmäßige Wartung angewiesen, da die Anbieter von Suchmaschinen häufig Änderungen an der Gestaltung ihrer Ergebnisseiten vornehmen.

6.5.2 Funktionale Grenzen

Einige genannte technische Grenzen haben einen starken Einfluss auf die funktionalen Grenzen, die bei der Nutzung des Prototyps bestehen. So können in der jetzigen Versionen nur mit viel Aufwand Untersuchungen in natürlichen Umgebungen durchgeführt werden, da auf jedem Computer, der genutzt werden soll, eine lokale Version installiert sein muss. Auch die Erhebung von Einschätzungen durch Testpersonen während der Suchsession ist mit der jetzigen Variante der Prototyps nicht möglich, da nur vor und nach der Bearbeitung von Aufgaben Fragebögen angeboten werden. Diese Vorgehensweise dient trotzdem dazu, ein genaueres Bild vom Benutzer in Bezug auf die Arbeitsaufgaben zu gewinnen. Daneben sind auch noch einige Einschränkungen in Bezug auf die Usability vorhanden.

So muss ein Projektleiter schon Kenntnisse im Umgang mit Formularen für Datenbanken als auch mit HTML-Quelltext haben, um eine Studie zu gestalten. Ferner ist das Fehlen eines Auswertungsprogramm auch bereits im Relevance Assessment Tool ein Schwachpunkt. Durch das Vorhandensein einer solchen Komponente würde der Aufwand von Information-Evaluierungen noch weiter reduziert werden.

7. Fazit und Ausblick

Zusammenfassend betrachtet wurden durch diese Arbeit zwei Schritte unternommen, um den Forschungsbereich der Information-Retrieval-Evaluierung zu erweitern. Auf der einen Seite bietet das Modell, das auf gängigen Verfahren in der Evaluation von Suchmaschinen beruht, eine Anleitung zur Durchführung von Retrievaltests, die auch das tatsächliche Nutzerverhalten einbeziehen. Die Zusammenstellung der Ergebnisse wird bei dieser Art von Evaluierung nur zum Teil oder überhaupt nicht vorgegeben und ermöglicht so eine realitätsnähere Bewertung der Qualität von Suchergebnissen und der Retrievaleffektivität von Suchmaschinen. Allerdings sind, wie auch in dem Modell diskutiert, immer Grenzen zu setzen. So ist eine Vergleichbarkeit der Ergebnisse aus solchen Studien teilweise stark eingeschränkt. Deshalb ist auch bei Studien, die auf dem vorgestellten Modell beruhen sollen, eine genaue Planung aller Bestandteile der Untersuchung wichtig. Neben des theoretischen Wegweisers des Modells, wurde noch ein Prototyp entwickelt, der die Vorteile aus den Evaluierungsmethoden verbinden soll und eine übersichtliche und technisch relativ einfache Methode für die praktische Durchführung von Information-Retrieval-Evaluierungen unterstützen soll. Auch bei dem Prototyp zeigen sich aber einige Grenzen in der Umsetzung. Das Werkzeug ist aber in seiner jetzigen Form bereits als Basis zur Durchführung von Nutzer- und Retrievaleffektivitätsstudien geeignet. Dabei kann das Tool aber bestimmte Entscheidungen für das Testdesign nur unterstützen. Entscheidungen über die Systeme, die getestet werden sollen, maximale Anzahl von Ergebnissen, Art von Ergebnissen und Zielgruppen sind immer noch Teil der Planung durch die Projektleiter.

In der Zukunft wird der Prototyp weiter entwickelt und soll neben einer Komponente zur automatischen Auswertung von Projektergebnissen noch weitere Optionen erhalten, die ein genaueres Bild des Benutzer erstellen können. Daneben soll eine Auswahl der Suchergebnistypen für den Suchmaschinenscraper umgesetzt werden, damit auch andere Ergebnistypen in die Relevanzbewertung mit einfließen.

8. Literaturverzeichnis

AKHIGBE ET AL. 2011a

Akhigbe, Bernard I. ; Afolabi, Babajide S. ; Udo, James I ; Adagunodo, Emmanuel R.: An Evaluative Model for Information Retrieval System Evaluation : A User-centered Approach, In: *Computer Engineering and Intelligent Systems, 4(2),* 2011, S. 34 – 45.

AKHIGBE ET AL. 2011b

Akhigbe, Bernard I. ; Afolabi, Babajide S. ; Udo, Adagunodo, Emmanuel R.: Assesment of Measures for Information Retrieval System Evaluation : A User-centered Approach, In: *Intermational Journal of Computer Applications, 7(25),* 2011.

BAILEY ET AL. 2007

Bailey, Peter ; Thomas, Paul ; Hawking, David: Does brandname influence perceived search result quality? : Yahoo!, Google, and WebKumara. In: *Conference on Human Factors in Computing Systems – Proceedings of the 12th Australasian Document Computing Symposium.* Melbourne, Australia : ACDS, 2007.

BUCKLEY ET AL. 2007

Buckley Chris. ; Dimmick, Darrin ; Soboroff, Ian ; Voorhes, Ellen: Bias and the limits of pooling for large collections. In: *Inf Retrieval 10,* 2007, S. 491 – 508.

BELKIN ET AL. 2009

Belkin, Nicholas J. ; Cole, Michael ; Liu, Jingjing: A Model for Evaluation of Interactive Information Retrieval. In: *SIGIR Workshop on the Future of IR Evaluation.* Boston, USA: SIGIR Workshop on the Future of IR Evaluation , 2009.

BELKIN 2010

Belkin, Nicholas J: On the evaluation of interactive information retrieval s ystems. In *The Janus faced scholar : A festschrift in honor of Peter I ngwersen.* ISSI, S. 13-22.

BAR-ILAN / LEVENE 2011

Bar-Ilan, Judit ; Levene, Mark: A method to assess search engine results. In: *Online Information Review, 35(6),* 2011, S. 854 – 868.

BORLUND 2003

Borlund, Pia.: The IIR evaluation model : a framework for evaluation of interactive information retrieval systems. In: *Information Research, 3(8),* 2003.

BRODER 2002
Broder, Andrei: A taxonomy of web search. In: *SIGIR Forum, Vol. 36*. New York, USA : ACM, S. 3 – 10.

BUSCHER ET AL. 2004
Buscher, Georg ; Dumais, Susan ; Cutrell, Edward: The Good, the Bad and the Random : An Eye-Tracking Study of Ad Quality in Web Search. In: *Proceedings of the 33rd annual international ACM SIGIR conference on Research and development in information retrieval*. Geneva, Switzerland : SIGIR' 10, 2004, S. 42 – 49.

CAPRA 2010
Capra, Robert: HCI Browser : A Tool for Studying Web Search Behavior. In: *Proceedings of the American Society for Information Science and Technology Annual Meeting.* Pittsburgh, USA : ASIS&T, 2010.

DAHM 2006
Dahm, Markus: *Grundlagen der Mensch-Computer-Interaktion*. München : Pearson Studium, 2006.

DE MEY 1977
De Mey, M.: The cognitive viewpoint : Its development and its scope. In: *CC 77: International Workshop on the Cognitive Viewpoint*. Ghent, Belgium : CC 77, 1977.

DIN 9241 1998
Norm DIN EN ISO 9241 T 11: 1998. In: Deutsches Institut für Normung e.V. (Hrsg.): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten Tl. 11: Anforderungen an die Gebrauchstauglichkeit.

FERBER 2003
Ferber, Reginald: *Information Retrieval : Suchmodelle und Data-Mining-Verfahren für Testsammlungen und das Web*. Berlin / Heidelberg : dpunkt-Verlag, 2003.

FOX ET AL. 2005
Fox, Steve ; Karnawat, Kuldeep; Mydland, Mark ; Dumais, Susan ; White, Thomas : Evaluating Implict Measures to improve Web Search. In: *PACM Transactions on Information Systems 2, (23), 2005, S. 147 – 168.*

GRIESBAUM 2000
Griesbaum, Joachim: *Evaluierung hybrider Suchsysteme im WWW*. Konstanz, Universität Konstanz, Informationswissenschaft, Dipl.-Arbeit, 2000

GRIESBAUM 2004
Griesbaum, Joachim: Evaluation of three German search engines : Altavista.de, Google.de and Lycos.de. In: *Information Research 9, (4),* 2004.

GRIESBAUM ET AL.. 2009
Griesbaum, Joachim ; Bekavac, Bernard ; Rittberger, Marc: Typologie der Suchdienste im Internet. In: Lewandowski, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2009, S. 18 – 52.

GÖDERT ET AL. 2012
Gödert, Winfried ; Lepsky, Klaus ; Nagelschmidt, Matthias: *Informationserschließung und Automatisches Indexieren : Ein Lehr- und Arbeitsbuch*. Berlin / Heidelberg : Springer-Verlag, 2012.

GORDON / PATHAK 1999
Gordon, Michael ; Pathak, Praveen: Finding information on the World Wide Web : the retrieval effectiveness of search engines. In: *Information Processing and Management 35, 1999*, S. 141 – 180.

GRANKA ET AL. 2004
Granka, Laura A. ; Joachims, Thorsten ; Gay, Geri: Eye-Tracking Analysis of User Behavior in WWW Search. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. South Yorkshire, UK : SIGIR'04, 2004, S. 25 – 29.

HAWKING ET AL. 2001
Hawking David ; Craswell, Nick ; Bailey, Peter ; Griffiths, Kathy: Measuring Search Engine Quality. In: *Information Retrieval 1(4), 2001*, S. 33 – 59.

HÖCHSTÖTTER 2009
Höchstötter, Nadine: Methoden der Erhebung von Nutzerdaten und ihre Anwendung in der Suchmaschinenforschung. In: Lewandowski, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2009, S. 175 – 203.

HÖCHSTÖTTER / KOCH 2009
Höchstötter, Nadine ; Koch, Martina: Standard parameters for searching behaviour in search engines and their empirical evaluation. In: *Journal of Information Science 35(1), 2009*, S. 45 – 65.

HUFFMAN / HOCHSTER. 2007
Huffman, Scott ; Hochster, Michael: How well does Result Relevance Predict Session Satisfaction?. In: *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM Press : New York, USA, 2007, S. 567 – 574.

INGWERSEN / JÄRVELIN 2005
Ingwersen, Peter; Järvelin, Kalervo: *The Turn : Integration of Information Seeking and Retrieval in Context*. Dordrecht, The Netherlands : Springer-Verlag, 2005.

JÄRVELIN 2009
Järvelin, Kalervo: Explaining User Performance in Information Retrieval : Challenges to IR evaluation. In: L. Azzopardi; Gabriella Kazai; Stephen Robertson; Stefan Rüger; Milad Shokouhi; Dawei Song & Emine Yilma: *Proceedings of the 2nd International Conference on the Theory of Information Retrieval.* Berlin / Heidelberg : Springer-Verlag, 2009, S. 289 - 296.

JANSEN 2006
Jansen, Bernard J.: The Wrapper : An Open Source Application for Logging User–System Interactions during Searching Studies. In: *Logging Traces of Web Activity : The Mechanics of Data Collection : Workshop at WWW 2006.* Edinburgh, Scotland : WWW 2006, 2006.

JANSEN / SPINK 2006
Jansen, Bernard J. ; Spink, Amanda: How are we searching the World Wide Web? : A comparison of nine search engine transaction logs. In: *Information Processing and Management 42, 2006*, S. 248 – 263.

JOACHIMS ET AL. 2005
Joachims, Thorsten ; Granka, Laura ; Pan, Bing ; Hembrooke, Helene ; Gay, Geri: Accurately Interpreting Clickthrough Data as Implicit Feedback. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development ininformation retrieval*. Salvador, Brazil : SIGIR' 05, 2004, S. 154 – 161.

KELLY 2009
Kelly, Diane: Methods for Evaluating Interactive Information Retrieval Systems with Users. In: *Foundations and Trends in Information Retrieval 1 – 2 (3), 2009*, S. 1 – 224.

KOFFLER / ÖGGL 2010
Koffler, Michael ; Öggl, Bernd: *PHP 5.3 MySQL 5.4*. München : Addison-Wesley, 2010.

KUMAR ET AL. 2005
Kumar, Rakesh ; Suri, P. K. ; Chauhan, R. K.: Search Engines Evaluation. In: *DESIDOC Bulletin of Information Technology, 25(2), March* 2005, S. 3 – 10.

LAMM 2008
Lamm, Katrin: Das *Confirmation/Disconfirmation-Paradigma der Kundenzufriedenheit im Kontext des Information Retrieval*. Hildesheim, Universität Hildesheim, Institut für Angewandte Sprachwissenschaft, Magister Arbeit, 2008

LEIGHTON / SRIVASTAVA 1999
Leighton, H. Vernon ; Srivastava, Jaideep: First 20 Precision among World Wide Web Search Services (Search Engines). In: *Journal of the American Society for Information Science 50,* 1999, S. 870-881.

LEWANDOWSKI 2005
Lewandowski, Dirk: Web Information Retrieval : Technologien zur Informationssuche im Internet. In: Ockenfeld, Marlies (Hrsg.): *Reihe Informationswissenschaft der DGI, Band 7*. Frankfurt : DGI, 2005

LEWANDOWSKI 2007
Lewandowski, Dirk: Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen?. In: Machill, Marcel (Hrsg.) ; Beiler, Markus (Hrsg.): *Die Macht der Suchmaschinen : The Power of Search Engines*. Köln : Herbert von Halem Verlag, 2007, S. 243 – 258.

LEWANDOWSKI / HÖCHSTÖTTER 2007
Lewandowski, Dirk ; Höchstötter, Nadine: Qualitätsmessung bei Suchmaschinen: System- und nutzerbezogene Evaluationsmaße. In: *Informatik Spektrum 30(3)*, 2007, S. 159 –169.

LEWANDOWSKI / HÖCHSTÖTTER 2008
Lewandowski, Dirk ; Höchstötter, Nadine: Web Searching : A Quality Measurement Perspective. In: Spink, Amanda (Hrsg.) ; Zimmer, Michael (Hrsg.) ; Beiler, Markus (Hrsg.): *Web Search : Multidisciplinary Perspectives : Springer Series in Information Science and Knowledge Management 14* . Berlin / Heidelberg : Springer-Verlag, 2008, S. 309 – 340.

LEWANDOWSKI 2008A
Lewandowski, Dirk: The Retrieval Effectiveness of Web Search Engines : Considering Results Descriptions. In: *Journal of Documentation, 64(6),* 2008, S. 915 – 937.

LEWANDOWSKI 2009
Lewandowski, Dirk: Standards der Ergebnispräsentation. In: Lewandowski, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2009, S. 204 – 219.

LEWANDOWSKI 2011A
Lewandowski, Dirk: The Retrieval Effectiveness of Search Engines on Navigational Queries. In: *Aslib Proceedings : New Information Perspectives, 63(4),* 2011, S. 354 – 363.

LEWANDOWSKI 2011b
Lewandowski, Dirk: Evaluierung von Suchmaschinen. In: Lewandowski, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen 2 : Neue Entwicklungen in der Web-Suche*. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2011, S. 203 – 228.

LEWANDOWSKI 2012
Lewandowski, Dirk: A Framework for Evaluating the Retrieval Effectiveness of Search Engines. In: Jouis, Christophe (Hrsg.): *Next Generation Search Engines: Advances Models for Information Retrieval*. Hershey, PA: IGI Global, 2012.

LEWANDOWSKI / SÜNKLER 2012
Lewandowski, Dirk ; Sünkler, Sebastian: Relevance Assessment Tool : Ein Werkzeug zum Design von Retrievaltests sowie zur weitgehend automatischen Erfassung, Aufbereitung und Auswertung der Daten [preprint]. In: *Proceedings der 2. DGI Konferenz 2012 : Social Media und Web Science : Das Web als Lebensraum*. Frankfurt/Main : DGI, 2012.

LORIGO ET AL. 2008
Lorigo, Lori. ; Haridasan, Maya ; Brynjarsdóttir, Hrönn ; Xia, Ling ; Joachims, Thorsten ; Gay, Geri ; Granka, Laura ; Pellacini, Fabio ; Pan, Bing : Eye Tracking and Online Srarch : Lessons Learned and Challenged Ahead. In: *Journal of the AmericanSociety for Information Science and Technology 59 59,* 2008, S. 1041 – 1052.

LUNA-PARK 2011A
Ohne Autor: Europa Suchmaschinen Marktanteile [online]. In: http://www.luna-park.de/blog/, 2011. – URL: http://www.luna-park.de/blog/1650-europa-suchmaschinen-marktanteile/ (Abruf: 2012-01-03)

LUNA-PARK 2011B
Ohne Autor: Suchmaschinen Marktanteile weltweit [online]. In: http://www.luna-park.de/blog/, 2011. – URL: http://www.luna-park.de/blog/1175-suchmaschinen-marktanteile/ (Abruf: 2012-01-03)

MACHILL U. A. 2003
Machill, Marcel ; Neuberger, Christoph ; Schweiger, Wolfgang ; Wirth, Werner: Wegweiser im Netz : Qualität und Nutzung von Suchmaschinen. In: Machill, Marcel (Hrsg.): *Wegweiser im Netz : Qualität und Nutzung von Suchmaschinen*. Gütersloh: Verl. Bertelsmann-Stiftung, 2003, S. 13 – 490

MANDL 2006
Mandl, Thomas: *Die automatische Bewertung der Qualität von Internet-Seiten im Information Retrieval*. Hildesheim, Universität Hildesheim, Fachbereich III – Informations- und Kommunikationswissenschaften, Habil-Schr. 2006

MANDL 2008
Mandl, Thomans: Recent Developments in the Evaluation of Information RetrievalSystems : Moving Towards Diversity and Practical Relevance In: *Informatica, 32,* 2008, S. 27 – 38.

MANNING ET AL. 2009
Manning, Christopher D. ; Raghavan, Prabhakar ; Schütze, Hinrich: *An Introduction to Information Retrieval*. Cambridge, Englad: Cambridge University Press, 2009.

MARCHIONINI 2006
Marchionini, Gary: Exploratory Search : From Finding to Understanding. In: C*ommunication of the ACM, 49(4),* 2006, S. 41 – 46.

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

NET APPLICATIONS 2012
Net Apllications: Search Engine Market Share [online]. – URL: http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4 (Abruf: 2012-01-03)

OPPERMANN / REITERER 1994
Oppermann, Reinhard ; Reiterer, Harald: Software-ergonomische Evaluation. In: Eberleh, Edmund (Hrsg.) ; Oberquelle, Horst (Hrsg.) ; Oppermann, Reinhard(Hrsg.): *Einführung in die Software-Ergonomie : Gestaltung graphisch-interaktiver Systeme : Prinzipien, Werkzeuge, Lösungen*. Berlin : Gruyter, 1994, S. 335 – 371.

PETERSON 2006
Peterson, A.: SEAT. In: Midnight Research Labs [online]. - http://www.marketing2oh.com/serps-for-seo-analysis, 2006 (Abruf: 2012-01-03)

POMBERGER / WEINREICH 1994
Pomberger, G. ; Weinreich, R.: The Role of Prototyping in Software Development. Turorial Paper. In: *Conference on the Technology of Object-Oriented Languages and Systems (TOOLS Europe '94). Versailles, France : Tools 94.*

QUIRMBACH 2009
Quirmbach, Sonja: Universal Search : Kontextuelle Einbindung von Ergebnissen unterschiedlicher Quellen und Auswirkungen auf das User Interface. In: Lewandowski, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2009, S. 220 – 248.

ROBERTSON 1981
Robertson, Stephen: Universal Search : The methodology of information retrieval experiment . In: Sparck-Jones, Karen (Hrsg.): *Information retrieval experiment* . Butterworths, 1981, S. 9 – 31.

ROBERTSON 2008
Robertson, Stephen: On the history of evaluation in IR. In: *Journal of Information Science, 4(34), 2008*, S. 439 – 456.

ROBINS 2000
Robins, David: Interactive Information Retrieval: Context and Basic Notions. In: *Informing Science 2(3), 2000*, S. 57 – 61.

SARODNICK / BRAU 2006
Sarodnick, Florian ; Brau, Henning: *Methoden der Usability Evaluation : Wissenschaftliche Grundlagen und praktische Anwendung*. Bern : Verlag Hans Huber, 2006.

SCHMIDT-MÄNZ / BOMHARDT 2005
Schmidt-Mänz, Nadine ; Bomhardt, Christian: Wie suchen Onliner im Internet?. In: S*cience Factory 2/2005*, S. 5 – 8.

SCHMIDT-MÄNZ 2007
Schmidt-Mänz, Nadine: *Untersuchung des Suchverhaltens im Web : Interaktion von Internetnutzern mit Suchmaschinen*. 1. Aufl. Hamburg : Verlag Dr. Kovač, 2007

SINGER ET AL. 2011
Singer, Georg ; Norbisrath, Ulrich ; Vainikko, Eero ; Hannu, Kikkas ; Lewandowski, Dirk: Search-Logger : Analyzing Exploratory Search Tasks. In: *The 26th ACM Symposium on Applied Computing.* TaiChung, Taiwan : SAC2011, 2011.

STOCK 2007
Stock, Wolfgang G.: I*nformation Retrieval: Informationen suchen und finden. Einführung in die Informationionswissenschaft Bd. 1*. München : Oldenbourg, 2007.

STOKDYK 2009
Stokydy, D.: How to Scrape a Search Engine result Page for Your SEO Project. In: Marketing OH: Internet Marketing for Web 2.0, 2006 [online]. - http://www.marketing2oh.com/serps-for-seo-analysis (Abruf: 2012-01-03)

TAGUE-SUTCLIFFE 1992
Tague-Sutcliffe, Jean: The Pragmatics of Information Retrieval Experimentation, revisted. In: *Information Processing & Management, 28(4), 1992*, S. 467 – 490.

TOTH 2006
Tóth, Erzsébet: Exploring the Capabilities of English and Hungarian Search Engines for Various Queries. In: *Libri vol. 56, 2006*, S. 38 – 47.

TREC 2011
Ohne Autor: Web Track Guidelines [online]. In: http://plg.waterloo.ca, 2011. – URL: http://plg.uwaterloo.ca/~trecweb/ (Abruf: 2012-01-03)

TURPIN/SCHOLER 2006
Turpin, Andrew H., Scholer, Falk: User Perfomance versus Precion Measures for Simple Search Tasks In: *Proceeding of the 29th Annual International ACM Sigir Conference on Research and Development.* Seattle, Washington, USA: Sigir '06, S. 11 – 18, 2006.

VAN EIMEREN / FREES 2011
van Eimeren, Birgit ; Frees, Beate: Ergebnisse der ARD/ZDF-Onlinestudie 2011 : Drei von vier Deutschen im Netz : ein Ende des digitalen Grabens in Sicht?. In: *Media Perspektiven* 7-8/2011, S. 334 – 349.

VAN RIJSBERGEN 1979
Van Rijsbergen: *Information Retrieval*. Londong, Englad: Buttersworth, 1979

VAUGHAN 2004
Vaughan, Liwen: New measurements for search engine evaluation proposed and tested. In: *Information Processing and Management 40, 2004*, S. 667 – 691.

VOORHEES / HARMAN 2006
Voorhees, Ellen M. ; Harman, Donna K.: TREC : An Overview. In: *Annual Review of Information Science and Technology 40,* 2006, S. 113 – 155.

WEBBER 2010
Webber, William Edward: *Measurement in Information Retrieval Evaluation*. Melbourne, University of Melbourne, Department of Computer Science and Software Engineering, Diss., 2010

WEBER 2011
Weber, Karsten: Search Engine Bias. In: Lewandowski, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen 2 : Neue Entwicklungen in der Web-Suche*. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2011, S. 265 – 288.

WEBHITS 2011
Web-Barometer [online]. – URL: http://www.webhits.de/deutsch/index.shtml?webstats.html (Abruf: 2012-01-03)

WHITE ET AL. 2006
White, Ryen W. ; Muresan, Gheorghe ; Marchionini, Gary: Report on ACM SIGIR 2006 Workshop on Evaluating Exploratory Search Systems. In: *ACM SIGIR Forum*. Seattle, Washington : USA, SIGIR'06, 2006.

WHITE ET AL. 2008
White, Ryen W. ; Marchionini, Gary ; Muresan, Gheorghe: Evaluating exploratory search systems Introduction to special topic issue of information processing and management. In: *Information Processing and Management 44, 2008*, S. 433 – 436.

WOMSER-HACKER 2004
Womser-Hacker, Christa: Theorie des Information Retrieaval III : Evaluierung. In: Kuhlen, Rainer (Hrsg.), Seeger, Thomas (Hrsg.), Strauch, Dietmar (Hrsg.): *Grundlagen der praktischen Information und Dokumentatiobn. Bd. 1 : Handbuch zur Einführung in die Informationswissenschaft und -praxis*. 5. Aufl, München : Saur Verlag, 2004, S. 227 – 235.

9. Glossar

Browser-Plugin Auch Addon oder Erweiterung. Mit Hilfe solcher Plugins wird die Funktionalität von Webbrowswers erweitert.

Dokumenten kollektion Zusammengestellte Kollektion von Dokumenten, die durch verschiedene Verfahren gewonnen werden. Der indexierte Inhalt des Webs durch eine Suchmaschinen, kann auch als Dokumentenkollektion betrachtet werden.

Ergebnisseite Webseiten oder Webdokumente, auf die die Hyperlinks aus der Trefferliste einer Suchmaschine verweisen.

Indexierung Indexierung ist beim Information Retrieval die Zuordnung von Deskriptoren bzw. Schlagwörtern zu einem Dokument zur Erschließung der darin enthaltenen Sachverhalte.

Indexierungs sprache Methoden zur Erstellung eines Indexes für Dokumente, z. B. statistisch-linguistische Verfahren, die auf Wörter und Grammatik von Sprachen beruhen.

Informations bedarf Der Informationsbedarf ist der Bedarf an handlungsrelevanten Wissen.

Informationsbed ürfnis Ist das subjektive Bedüfrnis eines Nutzers nach Informationen.

Information Retrieval Information Retrieval ist ein Teilgebiet der Informationswissenschaft, das sich mit der computergestützem Suchen nach Inhalten beschäftigt.

HTML Gestaltungs und Strukturierungssprache für Webseiten.

Hyperlink Verweis von einer Webseite auf eine andere Webseite.

Javascript Skriptsprache, um innerhalb von Webseiten Anweisungen ausführen zu können.

MySQL Relationales Datenbankmanagementsystem zur Verwaltung von Daten mit direkter Anbindung an PHP.

PHP Ist eine an die Programmiersprachen C und Perl angelegte Skriptsprache, die zur Programmierung dyanmischer Webseiten und Webanwendungen genutzt wird.

Phrasensuche Bei einer Phrasensuche werden mehrere Suchterme mit Anführungszeichen umschlossen und dadurch als ein Term behandelt (z.B. „usability test“).

Precision und Recall Trefferquote (engl. Recall), Genauigkeit (engl. Precision) sind Maße zur Gütebewertung von Treffermengen einer Recherche beim Information Retrieval.

Ranking Das Suchmaschinenranking bezeichnet die Reihenfolge, in der die bei der Benutzung der Suchmaschine ermittelten Ergebnisse aufgeführt werden. Das Ranking wird durch unterschiedliche Faktoren festgelegt, die das Ziel haben, das möglichst relevante Dokumente ganz oben in der Trefferliste angezeigt werden.

Suchanfrage Eine Suchanfrage besteht aus einem oder mehreren Suchtermen, die in einer Suchmaske eingegeben werden.

Suchsession Beschreibt alle Suchvorgänge (Eingabe von Suchanfragen, Sichtung von Ergebnissen, Reformulierung der Suchanfragne usw.)

Suchterm Ein Suchterm ist eine ununterbrochene Reihe von Zeichen.

Tab Tabs sind eine Sortier- und Navigationshilfe, die der weiteren Unterteilung von Einzelelementen dient.

Trefferliste Geordnete Liste mit Suchergebnissen.

URL Als Uniform Resource Locator werden die Adressen bezeichnet, die auf eine Ressource in einem Netzwerk verweisen.

Webcrawler Webcrawler sind Computerprogramme, die das Internet automatisch nach Webseiten absuchen und analysieren. Sie werden vor allem von Suchmaschinen eingesetzt, um Webdokumente für den Suchmaschinenindex zu sammeln.

Anhang

A 1 Bing Scraper

```
# Scraper Bing
function scrape_bing($search_query, $start, $search_engine)
{

// Analyse der Suchanfrage. Die Suchanfrage aus der Datenbank wird in das Standardformat von Bing konvertiert.
// Zum Beispiel wird bei "term1 term2" das Leerzeichen durch ein + Zeichen ersetzt, da dies der Standard beim Aufruf von Bing mit einer Suchanfrage ist.
// http://bing.com/search?q=term1+term2

$terms = explode(" ", $search_query);
$terms = implode("+", $terms);
$url_ref = 'http://bing.com';

// In diesem Schritt wird die Sprache der Suchmaschine festgelegt, um zwischen dem deutschsprachigen und englischsprachigen Angebot von Bing zu unterscheiden.

if ($search_engine == "bing.de")
{
   $language = "de-De";

}

if ($search_engine == "bing.com")
{
   $language = "en-US";
}
```

```
// Mit der Variable $start wird dem Scraper mitgeteilt, von welcher Trefferseite Ergebnisse gescrapt werden.
// Diese Variable ist notwendig, damit über die ersten 10 Treffer hinaus ein Erfassen von Suchergbenissen möglich ist.

$start = $start + 1;

// Hier erfolgt der Aufruf der Suchmaschine mit Suchanfrage, Sprache und Nummer der Trefferseite

$results = start_curl('http://www.bing.com/search?q='.$terms.'&setmkt='.$language.'&first='.$start);

// Für eine passende Zeichenkonvertierung wird dem Ergebnissen noch der entsprechende Zeichensatz zugeordnet.

$header = '<meta http-equiv="content-type" content="text/html; charset=utf-8">';
$results = $header.$results;

// Über XPath werden die Ergebnisse nun gefiltert. XPath ist eine Programmbibliothek zur Verarbeitung von XML-Dateien.
// Die gescrapten Ergebnisse werden nun als HTML Datei eingelesen und in ein Standard-XML Format gebracht.
// Der Vorteil liegt nun darin, dass die Daten (HTML-Tags mit Inhalten) als Baumstruktur vorliegen, die sich direkt ansprechen lässt.

$doc = new DOMDocument();
@$doc->loadHTML($results);

$doc->saveHTML();

$xpath = new DOMXPath($doc);

// Im ersten Schritt des Filters werden mit Hilfe von Schleifen die einzelnen Treffer innerhalb von einzelen Array-Elementen gespeichert.
// Damit werden Navigation oder andere Komponenten der Trefferliste eliminiert.

$cells = $xpath->query("//li/div[@class='sa_cc']");
```

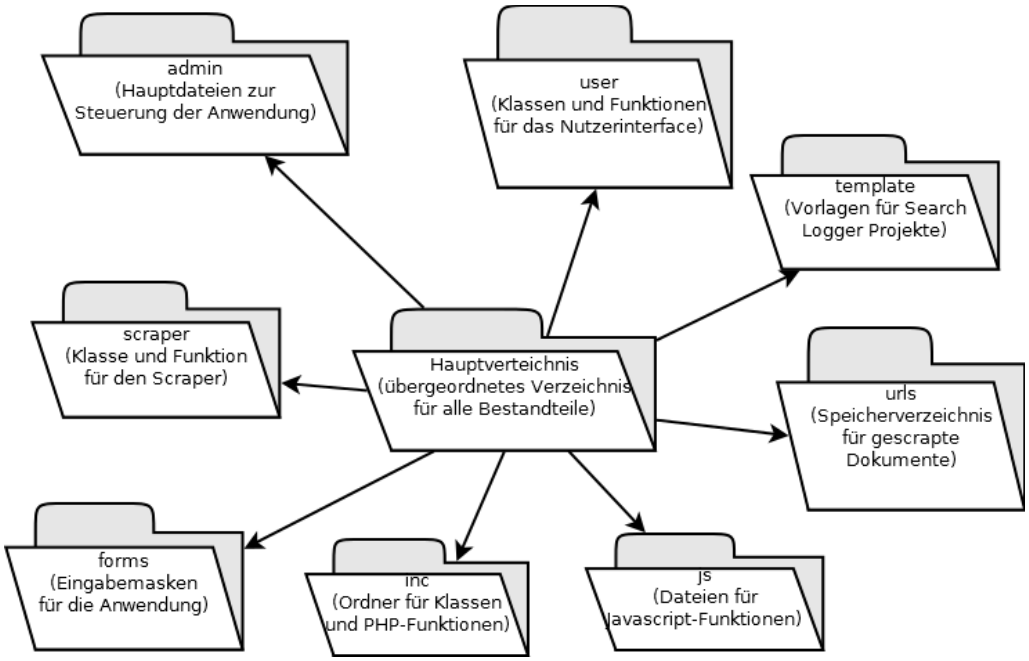
```
foreach ($cells as $cell)
{
   $test[] = $doc->saveXml($cell);
}

foreach($test AS $value)
{
   if(!empty($value))
   {
      $search_results[] = $value;
   }
}


$search_results = implode($search_results);
$results_doc = new DOMDocument();
@$results_doc->loadHTML($search_results);
$capture_xpath = new DOMXPath($results_doc);

// Nachdem die Trefferliste bereinigt wurde, erfolgt das Auslesen der einzelnen Ergebniselemente mit Hilfe der HTML-Tags, die in der Baumstruktur gespeichert worden.
// Jede Suchmaschine benutzt andere HTML-Tags, z. B. zur Formatierung und Gestaltung der Elemente. Daher wird für alle gewünschten Suchmaschinen ein eigener Filter benötigt.

// Speichern der URLs zu den Suchergebnissen.
$hrefs = $capture_xpath->evaluate("//div[@class='sb_tlst']/h3/a");

for ($i = 0; $i < $hrefs->length; $i++)
{
   $href = $hrefs->item($i);
   $result = $href->getAttribute('href');
   $result_link[] = $result;
}


// Speichern der Trefferüberschriten über die Linktexte
```

```
$hrefs = $capture_xpath->evaluate("//div[@class='sb_tlst']/h3/a");

for ($i = 0; $i < $hrefs->length; $i++)
{
   $href_text = $hrefs->item($i);
   $result = $href_text->nodeValue;
   $result_link_text[] = utf8_decode($result);
}

$hrefs = $capture_xpath->evaluate("//div[@class='sa_cc']");

$hrefs_abstr = $capture_xpath->evaluate("//div[@class='sa_cc']/p");

// Speichern der Trefferbeschreibungen mit einer besonderen Schleife zur Überprüfung
leerer Beschreibungen.

$y = 0;

for ($i = 0; $i < $hrefs->length; $i++)
{
   $href_desc = $hrefs->item($i);
   $result = $href_desc->childNodes->item(2)->nodeValue;
   $result_description[$i] = "";

   if(!empty($result))
   {
      $href_desc = $hrefs_abstr->item($y);
      $result = $href_desc->nodeValue;
      $result_description[$i] = utf8_decode($result);
      $y++;
   }
}
```

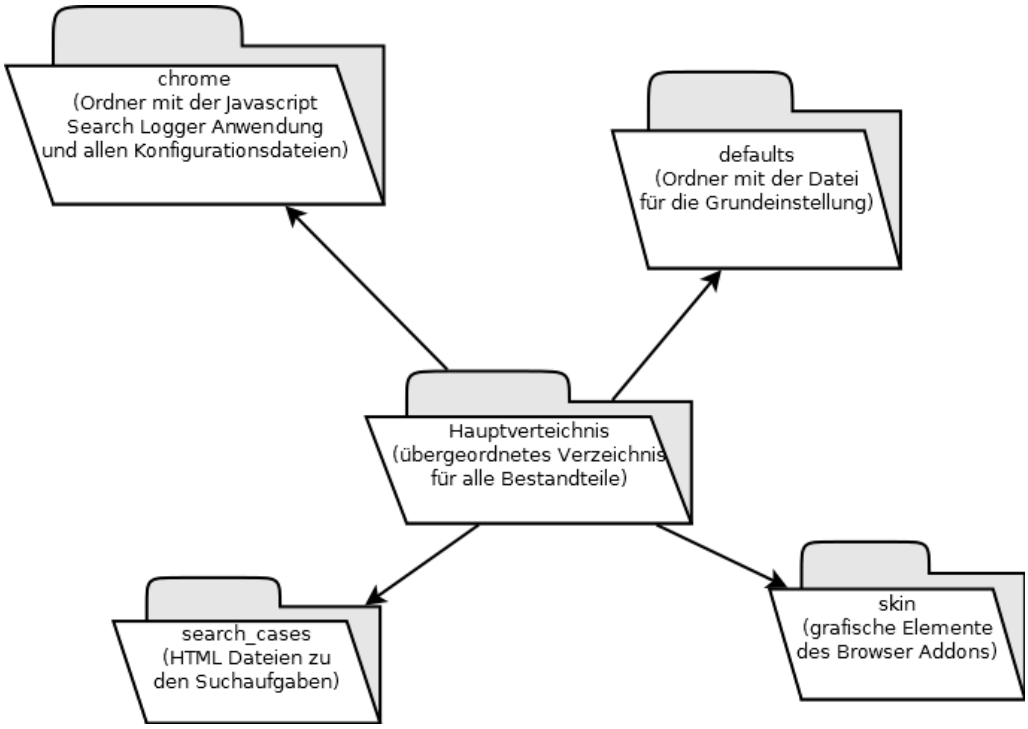
```
// Am Ende erfolgt die Zuweisung der gefundenen Ergebnisse in vorgesehenen Arrays.

$_SESSION['result_link '.$search_engine][] = $result_link;

$_SESSION['result_link_text '.$search_engine][] = $result_link_text;
$_SESSION['result_description '.$search_engine][] = $result_description;

}
```

A 2 Verzeichnisstruktur des Prototyps

A 2.1 Bestandteile des Administrationsinterfaces



A 2.2.2 Bestandteile des Search Loggers



A 3 Kurzanleitung für die Nutzung des Prototyps

A 3.1 Anlegen und Testen von Projekten

1. Öffnen Sie die Webseite http://electronicpassion.de/Search_Logger/rat/admin/

Loggen Sie sich dort mit dem Nutzernamen „search logger“ und dem Passwort „search logger ein“ (beides ohne Anführungszeichen) ein.

2. Clicken Sie auf den Menüpunkt „Templates“ und erstellen sie dort unter „Search Logger“ ein Template, in dem Sie die Formularfelder in der Eingabemaske ausfüllen. Sie finden dort auch bereits einige Standardeingaben, die Sie zum Testen verwenden können.

3. Sie können unter Templates nun auch Skalen definieren, wenn Sie mögen. Es stehen aber auch bereits einige Skalen für eine Verwendung zur Verfügung. Wenn Sie aber selber Skalen eingeben möchten, Klicken sie auf „Create Template“ unter „Templates“. Dort sehen Sie nun eine Eingabemaske. Sie können zwischen geschlossenen (closed), offenen (open) und Kommentar (comment) Skalen wählen. Die Kommentar-Skala erzeugt in der Eingabemaske für den Nutzer ein Textfeld, dass dieser mit beliebigen Inhalt füllen kann. Daneben müssen Sie noch entscheiden, welche Ergebnistypen mit der Skala bewertet werden sollen. Sie können zwischen „Resultpage“, „Description“ und „both“ wählen. Bei „both“ werden beide Ergebnisarten mit der selben Skala bewertet.

4. Klicken sie nun auf den Menüpunkt „Search Logger“ in der Navigationsleiste oben. Wählen Sie dort auf der linken Seite „Create New Project“. Sie können dort nun Angaben zum Projekt machen.

5. Im nächsten Schritt sollten Sie Suchaufgaben anlegen. Klicken Sie dafür auf „Create Work Task“ und füllen Sie das Formular aus. Geben Sie dort auch an, wie viele Suchanfragen insgesamt von einer Testperson bewertet werden sollen, wie viele maximale Suchergebnisse von einer Suchmaschine gescraped werden sollen und ferner ist es möglich auch eigene Suchanfragen zu definieren, die in jedem Fall zu einem Testset führen, das der Proband bewerten muss.

6. Sie haben auch die Möglichkeit Fragebögen für die Aufgaben zu definieren. Klicken Sie dazu für das demografische Formular auf „Edit Demographic Form“, für den Fragebogen vor Bearbeitung der Aufgabe auf „Edit Preform“ und für den Feedback nach Abschluss einer Aufgabe auf „Edit Postform“. Sie können nun eine HTML Datei mit den Formularen herunterladen und bearbeiten. Bitte achten Sie dabei auf die Hinweise in den Dateien. Laden Sie die geänderten Formulare anschließend über das angebotene Upload-Formular hoch.

7. Nachdem Sie ihr Projekt und die Suchaufgaben angelegt haben, müssen Sie dieses als Firefox-Erweiterung herunterladen. Klicken Sie dafür auf „Export Project as Firefox Extension“. Dort können Sie auch eine portable Version 3.6 des Firefox herunterladen, den Sie für die Nutzung der Erweiterung benötigen. Wenn Sie den Firefox herunterladen, erhalten Sie für Ihr System eine Zip-Datei. Die Sie entpacken können, um anschließend den Firefox zu starten. Sie müssen über das angebotene Formular auch die Erweiterung herunterladen. Wenn Sie auf „Export Project“ drücken.
7.

Erhalten Sie die Aufforderung zum Download einer XPI-Datei. Nachdem Sie diese Datei heruntergeladen haben, müssen Sie diese in Ihren Firefox Version 3.6 per Drag and Drop ziehen.

8. Wenn Sie die Datei in das Browserfenster gezogen haben, erscheint eine Aufforderung zur Installation. Bestätigen Sie bitte diese Aufforderung und starten sie anschließend den Firefox erneut.

9. Wenn alles geklappt hat, sehen Sie unten rechts im Browserfenster einen blinkenden Smiley, auf den Sie nun klicken können. Wenn Sie auf den Smiley geklickt haben, erscheint ein Fenster mit einem Willkommensgruß und dem Hinweis, dass sie auf einen Button drücken müssen, um mit dem Search Logger Experiment zu starten.

10. Klicken Sie auf den Button und füllen Sie das Demografie-Formular aus. Danach senden Sie dieses ab und bestätigen die Meldung, dass Firefox Daten auf Ihrem Computer speichern möchte. (Dieser Hinweis bezieht sich nur auf die Textdatei für das Loggen der Interaktionen und stellt keine Bedrohung für Ihr System dar.)

11. Sie werden nun zu der ersten Aufgabe geleitet und sehen das Formular für den Fragebogen, bevor Sie die Aufgabe bearbeiten. Füllen Sie dieses aus und Senden Sie es über den Button.

12. Es erscheint nun eine Hinweisseite, dass Sie eine neue Aufgabe gestartet haben und nun bearbeiten können. Wenn Sie mit der Bearbeitung fertig sind, klicken Sie bitte erneut auf den blinkenden Smiley, um die Bearbeitung abzuschließen.

13. Füllen Sie nun den Fragebogen aus, der nach Abschluss der Aufgabe angezeigt wird und drücken Sie auf den Sendebutton.

14. Sie sehen jetzt eine Seite mit dem Hinweis, dass Sie gleich zum Relevanztest für die Suchaufgabe geleitet werden. Bitte haben Sie etwas Geduld, da im Hintergrund nun die Prozesse zur Generierung des Ergebnissets laufen und unter Umständen etwas Zeit benötigen.

15. Sie sehen jetzt das Nutzerinterface für den Retrievaltest. Klicken Sie auf den Startbutton, um den Test zu beginnen.

16.Sehen Sie sich die Ergebnisse an und füllen Sie die Skalen zur Bewertung aus.

17. Nachdem Sie alle Suchergebnisse bewertet haben, sehen Sie eine Hinweisseite, dass Sie eine weitere Aufgabe bearbeiten können.

18. Wenn Sie alle Suchaufgaben bearbeiten möchten. Führen Sie nun Schritt 9 – 18 durch, bis keine Suchaufgabe mehr übrig ist.

19. Loggen Sie sich erneut mit den Zugangsdaten aus Schritt 1 ein und klicken Sie auf „Search Logger“ → Export Results. Drücken Sie auf den Button und Sie erhalten die Excel-Datei mit den Ergebnissen, die Sie durch die Bearbeitung der Suchaufgaben und des Relevanztests erzeugt habe,

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.

Hamburg, 29. Februar 2012

_____