Größere Zufriedenheit durch bessere Suchmaschinen?

Das Confirmation/Disconfirmation-Paradigma der Kundenzufriedenheit im Kontext des Information Retrieval

Katrin Werner, Hildesheim

In der vorgestellten Studie aus dem Bereich des interaktiven Information Retrieval wurde erstmals die Erwartungshaltung von Suchmaschinennutzern als mögliche Determinante der Benutzerzufriedenheit untersucht. Das experimentelle Untersuchungsdesign basiert auf einem betriebswirtschaftlichen Modell, das die Entstehung von Kundenzufriedenheit durch die Bestätigung bzw. Nicht-Bestätigung von Erwartungen erklärt. Ein zentrales Ergebnis dieser Studie ist, dass bei der Messung von Benutzerzufriedenheit besonders auf den Messzeitpunkt zu achten ist. Des Weiteren konnte ein von der Systemgüte abhängiger Adaptionseffekt hinsichtlich der Relevanzbewertung der Benutzer nachgewiesen werden.

Do better search engines increase user satisfaction? The confirmation/disconfirmation paradigm of customer satisfaction in the context of information retrieval

This article presents a study from the field of interactive information retrieval in which the expectations of search engine users were investigated as possible determinant of user satisfaction. The experimental design is based on a business model that explains the creation of customer satisfaction by the confirmation or disconfirmation of expectations. A central result of this study is that the point in time of the measurement is important with respect to the assessment of user satisfaction. Aside from that user relevance criteria seem to depend on system quality.

1 Einleitung

Die Frage, wann ein Produkt oder eine Dienstleistung einen Kunden zufrieden stellt, fällt in den Kernbereich der traditionellen Marketingforschung. Dort dominiert das anschauliche Confirmation/Disconfirmation(C/D)-Paradigma, das die Entstehung von Kunden(un)zufriedenheit durch die Bestätigung bzw. Nicht-Bestätigung von Erwartungen erklärt (Sauerwein 2000, Scharnbacher & Kiefer 1996). Obwohl diese Überlegung plausibel erscheint, wird die Erwartungshaltung von Benutzern bei der Evaluierung von Suchmaschinen bisher nur wenig beachtet. Im Rahmen der im Folgenden vorgestellten Untersuchung sollte daher geprüft werden, inwieweit sich dieses Konzept auch auf den Information-Retrieval(IR)-Kontext übertragen lässt.

Eine stärkere Einbeziehung von Benutzern in den Evaluierungsprozess von IR-Systemen ist ein Ansatz, der erst allmählich in das Blickfeld wissenschaftlicher Untersuchungen gerät. Die Einführung von Aufgaben zu solch einem interaktiven Retrieval durch die großen Evaluierungsinitiativen wie das Cross-Language Evaluation Forum (CLEF[1]) und die Text REtrieval Conference (TREC[2]) zeigt jedoch, dass in diesem Bereich ein Umdenken stattfindet.

Dass die Berücksichtigung der Benutzer zu neuen Einsichten führen kann, zeigt u.a. eine Untersuchung zu Microsofts Suchmaschine Bing[3], bei der allein durch die Variation der Linkfarbe die Klickrate der Benutzer deutlich erhöht werden konnte (Spiegel Online 2010). Dieses Beispiel macht deutlich, dass Ergebnisse aus systemorientierten im Vergleich zu benutzerorientierten Evaluierungen nur eine beschränkte Aussagekraft haben können, da derartige situative Einflussfaktoren bei der systemorientierten Evaluierung nicht berücksichtigt werden können. Gleichzeitig wird aber auch klar, dass sich die Komplexität des Evaluierungsprozesses durch die Einbeziehung von Benutzern wesentlich erhöht. Weiterhin gibt es im Gegensatz zum systemorientierten Batch-Evaluierungsansatz im Bereich der benutzerorientierten IR-Evaluierung bislang noch keine standardisierten Verfahren, was zur Folge hat, dass Untersuchungsdesigns und Ergebnisse hier gegenwärtig noch sehr heterogen und damit oft schwer vergleichbar sind.

Im Mittelpunkt dieses Beitrags steht die Darstellung einer Untersuchung zum Websuchverhalten, die im Rahmen einer vom Verein zur Förderung der Informationswissenschaft (VFI) mit einem Förderpreis ausgezeichneten Magisterarbeit an der Universität Hildesheim durchgeführt wurde. Eine ausführlichere Darstellung der Thematik findet sich in Lamm (2008).

2 Batch- vs. Benutzerevaluierungen

„Man kann an einem Auto verschiedene Messungen vornehmen und erhält so eine Menge von Messwerten. Ein Messwert ist der Benzinverbrauch. Er beschreibt das Automobil, denn mit ihm kann man abschätzen, wann man wieder tanken muss. Man kann den Benzinverbrauch aber auch zum Vergleichen benutzen, wenn man ihn als Kriterium beim Kauf eines Automobils benutzt.“ (Bollmann & Cherniavsky 1980) Dieses Beispiel veranschaulicht die Grundidee der IR-Evaluierung, verschiedene IR-Systeme anhand abstrahierter Effektivitätsmaße vergleichbar zu machen. Dabei hat es sich bewährt, das System selbst als eine Art *Black Box* zu betrachten, deren innerer Aufbau und Funktionsweise für die Bewertung ausgeblendet werden (Womser-Hacker 2004). So wie in dem einlei-

1 http://www.clef-campaign.org/ (Alle hier angegebenen URLs wurden am 14.07.2010 auf Erreichbarkeit überprüft)
2 http://trec.nist.gov/
3 http://www.bing.com/?cc=de

tenden Beispiel nur der Benzinverbrauch und nicht die Arbeitsweise des Motors berücksichtigt wird, ist bei der Bewertung eines IR-Systems allein sein Input-Output-Verhalten von Interesse.

Hinsichtlich der Evaluierung von Suchergebnissen existieren im Bereich der IR-Evaluierung zwei unterschiedliche Bewertungsansätze: Der system- und der benutzerorientierte Ansatz. Während bei ersterem die Systemperspektive vorherrscht, steht bei letzterem die Perspektive des Benutzers im Mittelpunkt. Dabei entspricht der systemorientierte Ansatz dem einleitenden Beispiel zum Benzinverbrauch, da auch in diesem Fall objektive Messgrößen betrachtet werden. Beim benutzerorientierten Ansatz werden zusätzlich subjektive Messgrößen untersucht. Dies käme einer Unterscheidung von Autos durch das von ihnen vermittelte Fahrvergnügen gleich. Bei diesem Ansatz sind also überdies das Verhalten und Erleben des Benutzers von Interesse.

Die systemorientierte Bewertung hat eine längere Tradition als die benutzerorientierte und stellt den Kern aller wichtigen Evaluierungsinitiativen dar. Bei diesem Ansatz werden automatisiert Testanfragen an ein System gestellt und durch die Auswertung verschiedener Effektivitätsmaße überprüft, wie gut relevante Dokumente gefunden und irrelevante Dokumente zurückgehalten werden. Die Relevanzbewertung der Dokumente erfolgt durch unabhängige Experten mit dem Ziel, eine möglichst einheitliche und objektive Bewertung zu erreichen. Nach den ersten großen Retrievaltests, die mit der sogenannten *Cranfield-Kollektion*[4] durchgeführt wurden, wird dieses Vorgehen in der Literatur als das *Cranfield-Paradigma* der Evaluierung bezeichnet (Buckley & Voorhees 2005, Mandl 2008).

Aus dem Umstand, dass bei dieser Form der Evaluierung die Relevanz der Dokumente bestimmt werden muss, ergibt sich das sogenannte Stellvertreterentscheidungsproblem, da Juroren, welchen das Informationsbedürfnis des Benutzers unbekannt ist, die Qualität des Retrievaloutputs bewerten (Möhr 1980). Tatsächlich wird also gemessen, inwieweit der Retrievaloutput mit den Expertenbewertungen übereinstimmt. Diese Einwände fallen bei der vergleichenden Evaluierung mehrerer Systeme weniger stark ins Gewicht. Deshalb begegnet man „[…] dieser Problematik durch den Einsatz komparativer Evaluierungsverfahren, welche die beteiligten Information-Retrieval-Systeme gleich behandeln, so dass die Ergebnisse im Vergleich ihre Gültigkeit bewahren, jedoch nicht als Einzelbewertung pro System valide sind." (Womser-Hacker 2004)

Aufgrund des Umfangs heutiger Testkollektionen wird bei der vergleichenden Evaluierung durch Evaluierungsinitiativen häufig die sogenannte *Pooling-Methode* angewendet (Harman 1995). Dabei werden nur die von den teilnehmenden Systemen zurückgelieferten Dokumente durch Juroren bewertet. Alle nicht zurückgelieferten Dokumente werden bei diesem Verfahren als irrelevant eingestuft. Wie durch die Pooling-Methode versucht wird, eine möglichst präzise Annäherung an die Gesamtzahl aller in der Kollektion enthaltenen relevanten Dokumente zu einer Suchanfrage zu erreichen, ist in Abbildung 1 graphisch dargestellt. Dabei steht R für die Gesamtzahl der im Dokumentenbestand D zu einer Suchanfrage vorhandenen relevanten Dokumente. Aus der Gesamtzahl der von den einzelnen Systemen A, B und C zurückgelieferten relevanten Dokumente ergibt sich der Schätzwert für R.

Problematisch ist jedoch, dass bei dieser Methode das Risiko besteht, dass die Effektivitätsbewertung der Systeme verfälscht wird, weil die Anzahl der nicht bewerteten Dokumente bei großen Testkollektionen unter Umständen sehr umfangreich sein kann (Turpin & Scholer 2006). Dies birgt in der Folge die Gefahr, zu viele relevante Dokumente als irrelevant einzustufen.

Neben der intendierten höheren Objektivität der Relevanzbewertungen durch unabhängige Experten besteht ein weiterer Vorteil des systemorientierten Bewertungsansatzes darin, dass Retrievaltests ohne die Einbeziehung realer Benutzer mit vergleichsweise geringem Aufwand durchzuführen sind. Allerdings bleibt zu klären, inwieweit sich diese systemorientierten Ergebnisse auf reale Benutzer und deren individuelle Informationsbedürfnisse übertragen lassen. Eine einfache Antwort auf diese Frage wird wohl auch künftig nicht zu finden sein, da sich durch die Einbeziehung der Benutzer, wie bereits angedeutet, völlig neue Evaluierungsmöglichkeiten eröffnen. Al-Maskari und Sanderson (2006) fassen dies prägnant zusammen: „Therefore, the issue in Information Retrieval (IR) shifts from maximizing the retrieval performance by refining IR techniques and methods to maximizing the understanding of users' behaviors and information need representation during retrieval."

Im Unterschied zum systemorientierten verfolgt der benutzerorientierte Bewertungsansatz deshalb das Ziel, die Anwendungssituation von IR-Systemen möglichst realistisch zu simulieren. Dazu werden neben einer Testkollektion reale Benutzer als Versuchspersonen benötigt, die innerhalb eines vorgegebenen Anwendungsszenarios mit dem System interagieren. In der Regel konfrontiert man hierzu die Testbenutzer mit Testaufgaben, die sie mit Hilfe des zu beurteilenden Systems lösen sollen. Im Fokus können je nach zu untersuchender Fragestellung zum Beispiel der Suchprozess, die Qualität der erreichten Lösungen, das persönliche Erleben der Testpersonen oder die Beobachtungen des Versuchsleiters stehen. Daran wird bereits deutlich, dass diese Art der Evaluierung zwar realistischer, aber auch aufwendiger und komplexer ist als im rein systemorientierten Fall.

Ingwersen und Järvelin (2005) nennen in der Einleitung zu ihrem Buch „The turn: integration of information seeking and retrieval in context" zwei Bereiche im Zusammenspiel von IR-Systemen und realen Benutzern, in welchen gegenwärtig ein Umdenken stattfindet:

- Anstatt das Konzept der *Relevanz* als einfaches binäres Konzept zu betrachten, bringen reale Benutzer subjektive und dynamische Relevanzbewertungen in den Evaluierungsprozess mit ein, die zudem an aktuelle Begebenheiten geknüpft sind.
- Im Bereich der *Information-Seeking-Modelle* wird IR als ein Bestandteil im Gesamtprozess der Informationssuche betrachtet, wobei auch hier der Einfluss aktueller Begebenheiten, wie der Suchaufgabe, betont wird.

Wie diese kurze Übersicht zeigen sollte, besteht eine Schwierigkeit der benutzerorientierten Evaluierung darin, dass Benutzertests keine einheitlichen und objektiven, sondern individuelle, durch die Testpersonen subjektiv gefärbte Ergebnisse liefern. In der Regel werden hier der Erfolg und die Zufriedenheit der Benutzer bewertet. Während sich der Erfolg jedoch wie bei der systemorientier-

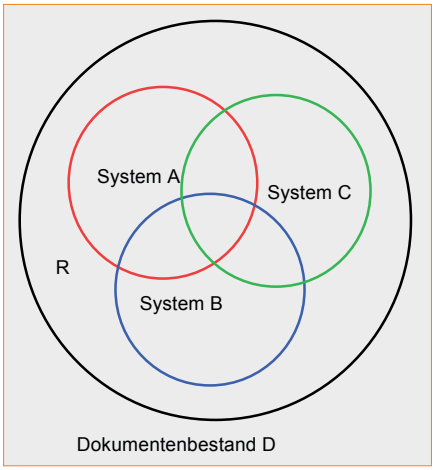

Abbildung 1: Die Pooling-Methode als Verfahren zur Bestimmung der Gesamtzahl aller relevanten Dokumente in einer Kollektion (Quelle: In Anlehnung an Kowalski 1997).

4 Eine solche Testkollektion beinhaltet als wesentliche Elemente eine Sammlung von Dokumenten, eine Zusammenstellung von Testanfragen, sogenannten Topics, sowie die zugehörigen Relevanzurteile, die angeben, welche Dokumente aus der Kollektion für die jeweilige Anfrage relevant sind.

ten Evaluierung anhand der gefundenen Dokumente bestimmen lässt, ist es notwendig, die Zufriedenheit direkt bei den Benutzern zu erfragen. Vor allem in Bezug auf die Erhebung der Benutzerzufriedenheit müssen also geeignete Methoden gefunden und in der IR-Community etabliert werden, denn nur so wird ein Vergleich unterschiedlicher Untersuchungsergebnisse möglich. Diese Schwierigkeit hat in der IR-Forschung dazu beigetragen, dass die systemorientierte Evaluierung immer noch bevorzugt wird. Da jedoch die Suchleistung, die reale Benutzer mit IR-Systemen erreichen, am Ende über deren Anwendbarkeit entscheidet, sollte die benutzerorientierte Evaluierung nicht vernachlässigt werden. Diese Tatsache wird von Järvelin und Ingwersen (2004) wie folgt zusammengefasst: „The real issue in information retrieval systems design is not whether its recall-precision performance goes up by a statistically significant percentage. Rather, it is whether it helps the actor solve the search task more effectively or efficiently."
Zwei wichtige Fragen, die man im Rahmen von Benutzerevaluierungen zu beantworten versucht, sind, ob bessere Suchmaschinen (1) den objektiv messbaren (Benutzerleistung) und (2) den subjektiv wahrgenommenen (Benutzerzufriedenheit) Sucherfolg von Suchmaschinennutzern erhöhen. Zur Übersicht werden im Folgenden vier Studien, die diese Fragestellungen untersuchen, skizziert.

2.1 Benutzerleistung im IR

Für die Bewertung der Benutzerleistung gelten ähnliche Maßstäbe wie in der systemorientierten Evaluierung. Die Grundfrage ist jedoch in diesem Zusammenhang, wie gut verschiedene Benutzer in der Lage sind, mit einem zu evaluierenden System relevante Dokumente zu finden. Dazu wird auch in diesem Fall eine Testkollektion, bestehend aus Aufgaben, Dokumenten und Relevanzurteilen, benötigt. Zwei wesentliche Erkenntnisse aus bisherigen Studien sind, dass Benutzer teilweise in der Lage sind, systemseitige Leistungsunterschiede zu kompensieren und dass die Relevanzkriterien von Benutzern stark kontextabhängig sind.
In einem 3 x 3 faktoriellen Design mit zwei Experimentalgruppen und einer Kontrollgruppe haben Smith und Kantor (2008) das Suchverhalten von 36 Testpersonen untersucht. Ihre Ausgangshypothese war, dass Benutzer Systemunterschiede so gut kompensieren können, weil sie ihr Verhalten anpassen. Google-Ergebnisse dienten als Dokumentenbasis. Dazu sendete das von Smith und Kantor verwendete Testsystem die Suchanfragen der Teilnehmer an Google und lieferte je nach Testbedingung Ergebnislisten, beginnend an unterschiedlichen Rankingpositionen, zurück. Teilnehmer der Kontrollgruppe bekamen die Ergebnisse so dargeboten, wie sie von Google zurückgeliefert wurden. Bei den Experimentalgruppen gab es eine Gruppe mit konsistent hohen Rankingpositionen (beginnend ab Rang 300) und eine mit inkonsistenten Rankings. Die Ergebnisse von Smith und Kantor (2008), dass Nutzer des schlechteren Systems ihre Relevanzkriterien zu lockern und somit im weiteren Verlauf auch weniger relevante Dokumente zu akzeptieren scheinen, stimmen mit der hier vorgestellten Studie hinsichtlich der Kontextabhängigkeit von Relevanzurteilen in Benutzerevaluierungen überein.
Zu einem ähnlichen Ergebnis kommen auch Smucker und Jethani (2010) in ihrer erst kürzlich veröffentlichten Studie zum Vergleich von Retrieval Precision und Benutzerleistung. In einem Within-Subjects-Design untersuchten sie das Suchverhalten von 48 Testpersonen. Insgesamt wurden von den Testpersonen acht Aufgaben aus dem TREC 2005 Robust Track bearbeitet, je vier pro Testphase. Während die Teilnehmer in der ersten Phase gebeten wurden, jedes Dokument in der Ergebnisliste zu bewerten, konnten die Teilnehmer in der zweiten Phase frei entscheiden, welche Dokumente sie ansehen und bewerten wollten. Die Ergebnislisten wurden so manipuliert, dass

die Precision für das gute System 0,6 betrug, für das schlechte System hingegen 0,3. Auch hier zeigte sich ein adaptives Verhalten der Benutzer. Nutzer des besseren Systems neigten in beiden Testphasen dazu, strengere Relevanzkriterien auf bessere Ergebnislisten anzuwenden. Im Gegensatz zu anderen Studien konnten die Testpersonen in dieser Studie den Systemunterschied nicht kompensieren, was sich an den signifikant besseren Leistungen der Testpersonen mit dem besseren System zeigt (Smucker & Jethani 2010).

2.2 Benutzerzufriedenheit im IR

In Bezug auf die Evaluierung der Benutzerzufriedenheit haben sich bisher noch keine einheitlichen Methoden etabliert. Häufig wird die Zufriedenheit über zusätzliche Fragebögen im Anschluss an die Bearbeitung der Testaufgaben ermittelt. Dabei variieren Art und Form der verwendeten Einzelfragen und Skalen noch stark zwischen den verschiedenen Studien. Auf die Frage, wie Zufriedenheit entsteht und welche Determinanten vorrangig die Benutzerzufriedenheit und den Sucherfolg erklären, wird meist nicht explizit eingegangen.
Eine Studie, die diese Fragestellung aufgreift, wurde von Szajna und Scamell (1993) durchgeführt. Diese Studie weist sowohl hinsichtlich der überprüften Fragestellung als auch methodisch viele Parallelen zu der in diesem Artikel beschriebenen Untersuchung auf. Szajna und Scamell untersuchten den Effekt von Benutzererwartungen im Kontext eines Informationssystems. Im Gegensatz zu der hier vorgestellten Studie wurde die Theorie der kognitiven Dissonanz (siehe auch Abschnitt 3) verwendet, um die Reaktionen der Teilnehmer vorherzusagen. Diese von Festinger (1978[5]) entwickelte Theorie geht davon aus, dass Menschen nach kognitiver Harmonie suchen. Um kognitive Dissonanzen bzw. Ungleichgewichte zwischen Erwartungen und Wahrnehmungen zu reduzieren, tendieren Menschen dazu, entweder die eigenen Erwartungen zu relativieren oder aber die Wahrnehmung zu korrigieren. In einer Längsschnittstudie kontrollierten Szajna und Scamell die Erwartungen von 159 Testpersonen als hoch, moderat oder niedrig. Zur Untersuchung des Zusammenhangs zwischen Erwartungen, Zufriedenheit und Benutzerleistung, wurde die Zufriedenheit der Teilnehmer an drei Messzeitpunkten mit Intervallen von einer und drei Wochen erfasst. Die Ergebnisse zeigen einen Zusammenhang zwischen dem Realismus von Benutzererwartungen und der wahrgenommenen Qualität des Informationssystems. Kein Zusammenhang zeigte sich hingegen hinsichtlich der Benutzerleistung. Besonders interessant in Bezug auf die hier vorgestellte Untersuchung zum Informationssuchverhalten ist die Beobachtung, dass Benutzererwartungen mit der Zeit verblassen (Szajna & Scamell 1993).
Eine weitere interessante Studie, die sich mit dem Einfluss von Benutzererwartungen auf die Systemevaluierung befasst, wurde von Jansen et al. (2007) realisiert. In einem Laborexperiment mit 32 Testpersonen untersuchten sie den Effekt von Marken bei Suchmaschinen. In einem Within-Subjects-Design wurde die Markenwahrnehmung der Teilnehmer in Bezug auf drei bekannte Suchmaschinen (Google, MSN[6], Yahoo[7]) und eine den Testpersonen unbekannte Suchmaschine verglichen. Jede Testperson bearbeitete mit jeder der vier Suchmaschinen eine Suchaufgabe. Die angezeigten Ergebnisdokumente waren für jede Suchmaschine pro Suchaufgabe identisch, die Markenelemente auf der Suchergebnisseite waren für jede Suchmaschine unterschiedlich. Die Wirkung der Markenerwartung wurde über die Relevanzbewertung der Suchergebnisse erfasst. Auch hier zeigen die Ergebnisse, dass Erwartungen einen deutlichen Einfluss auf die Qualitätsbewertung von Suchergebnissen haben. Im Vergleich schneidet die den Teilnehmern unbekannte Suchmaschine am Schlechtesten ab (Jansen et al. 2007).

3 Theoretische Überlegungen zur Zufriedenheit

Zu Beginn dieses Abschnitts soll das Auto-Beispiel aus Abschnitt 2 noch einmal aufgegriffen werden. Man stelle sich vor, ein Sport- und ein Kleinwagenfahrer machen beide eine Probefahrt mit einem Mittelklassewagen. Anschließend werden beide nach ihrer Zufriedenheit mit dem Beschleunigungsverhalten des Wagens gefragt. Vermutlich wird der Sportwagenbesitzer weniger zufrieden gestellt sein als der Kleinwagenfahrer. Diese Abhängigkeit des Zufriedenheitsurteils von Vorerfahrungen und Erwartungen an ein Produkt wird auch in der Kundenzufriedenheitsforschung untersucht.
Dabei hat sich das sogenannte *Confirmation/Disconfirmation (C/D) - Paradigma* etabliert (Sauerwein 2000, Scharnbacher & Kiefer 1996). Dieses Modell erklärt den Entstehungsprozess von Zu- oder Unzufriedenheit gerade als einen solchen individuellen Vergleichsprozess zwischen den Erwartungen an ein Produkt einerseits (Soll-Komponente) und der wahrgenommenen Produktqualität andererseits (Ist-Komponente). Die Grundlage für die Entstehung von Zu- oder Unzufriedenheit folgt dann tatsächlich aus der Bestätigung (confirmation) oder eben Nicht-Bestätigung (disconfirmation) der Kundenerwartungen.

Das Prinzip des C/D-Paradigmas wird in Abbildung 2 veranschaulicht. Werden im Rahmen eines Soll-Ist-Vergleichs die Erwartungen des Kunden erfüllt, entspricht also die Ist- der Soll-Leistung, ist der Kunde zufrieden. Man bezeichnet dies als *Konfirmation*. Im Fall einer Nicht-Bestätigung der Kundenerwartungen unterscheidet man zwei Situationen: Werden die Erwartungen des Konsumenten übertroffen, übersteigt also die Ist- die Soll-Leistung, wird in der Literatur von *positiver Diskonfirmation* gesprochen. Werden die Erwartungen hingegen enttäuscht, liegt also die Ist- unter der Soll-Leistung, ist der Kunde unzufrieden. Man bezeichnet dies als *negative Diskonfirmation*. Dieses Modell diente als theoretischer Bezugsrahmen für die im Folgenden beschriebene Studie.
Ebenso wie die Erwartungen der Kunden mitunter subjektiv gefärbt sind, können auch bezüglich der Wahrnehmung der Produktleistung Verzerrungen entstehen. Sauerwein (2000) zufolge ist die Theorie der kognitiven Dissonanz eine der wichtigsten Theorien, die zur Interpretation solcher Wahrnehmungsverzerrungen entwickelt wurde. Wie schon im vorherigen Abschnitt angedeutet geht das von dem Sozialpsychologen Festinger eingeführte Erklärungsmodell davon aus, dass Wahrnehmungsverzerrungen eine Folge nichtbestätigter Erwartungen sind. Gemäß dieser Theorie ist der Mensch bestrebt „ […] eine Harmonie, Konsistenz oder Kongruenz zwischen seinen Meinungen, Attitüden, Kenntnissen und Wertvorstellungen herzustellen.“ (Festinger 1978) Im Fall einer Erwartungsdiskonfirmation entstehen beim Kunden kognitive Spannungen, sogenannte *Dissonanzen*. Um diesem Zustand des Ungleichgewichts entgegenzuwirken, tendieren Konsumenten dazu, wahrgenommene und erwartete Produktleistung aneinander anzupassen, was durch Senkung der Erwartungen oder Erhöhung der wahrgenommenen Produktleistung geschehen kann (Sauerwein 2000).
Dieser kurze Exkurs macht noch einmal deutlich, dass beim interaktiven IR ganz andere Probleme zu bewältigen sind als im Fall von Batchevaluierungen. Die Herausforderung besteht u.a. darin, eine vernünftige Balance zwischen pragmatisch realistischen, ebenso aber auch sinnvoll interpretierbaren Evaluierungsergebnissen zu finden.

5 Die erste Fassung der Theorie der kognitiven Dissonanz wurde im Jahr 1957 veröffentlicht. Bei der hier zitierten Fassung handelt es sich um eine Übersetzung.

6 Seit Bing, die neue Suchmaschine von Microsoft, auf dem Markt ist, wird man von der Domain http://www.live.com auf die Seite von Bing weiterverwiesen.

7 http://www.yahoo.com/

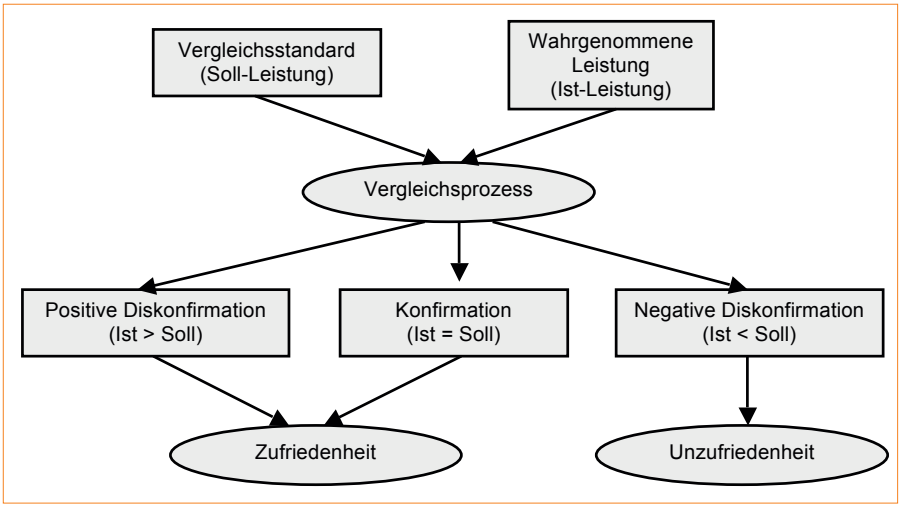

Abbildung 2: Die zentralen Bestimmungsfaktoren des Confirmation/Disconfirmation-Paradigmas der Kundenzufriedenheit (Quelle: Homburg et al. 1999).

4 Das Untersuchungsdesign im Überblick

Die Untersuchung des Zusammenhangs zwischen Systemleistung und Benutzererwartungen auf der einen und Benutzerzufriedenheit und Benutzerleistung auf der anderen Seite wurde in der vorliegenden Arbeit nicht isoliert vorgenommen. Da in einer realistischen Anwendungssituation von Suchmaschinen meistens mehrere Faktoren gleichzeitig wirken, bestand das **vorrangige** Ziel dieser Untersuchung darin, die wahrgenommene Qualität von Suchergebnissen nicht losgelöst vom eigentlichen Suchprozess, sondern integriert in den Kontext der Suche zu evaluieren. In diesem Fall bietet sich die Wahl eines mehrfaktoriellen Erhebungsdesigns an, da bei einem solchen Design auch Aussagen über Wechselwirkungen zwischen den untersuchten Faktoren gemacht werden können. Verwendet wurde ein zweifaktorielles Design, mit den jeweils zweifach gestuften Faktoren Erwartungshaltung (niedrig vs. hoch) und Systemleistung (schlecht vs. gut). Tabelle 1 zeigt den Versuchsplan mit Angabe der Zahl der Versuchspersonen pro Zelle.

Um systematische Verzerrungen durch Alter und Geschlecht der Testpersonen zu vermeiden, wurde die Stichprobe auf weibliche Testpersonen im Alter zwischen 17 und 35 Jahren beschränkt. Des Weiteren erfolgte die Zuteilung der Untersuchungsteilnehmerinnen zu den vier Versuchsgruppen zufällig. Da die Teilnehmerinnen weder wussten, in welcher Gruppe sie waren, noch dass es verschiedene Gruppen gab, handelt es sich bei diesem Experiment um einen Blindversuch.

Zur Manipulation der Erwartungshaltung bekamen Teilnehmerinnen, bei denen die niedrige Erwartungshaltung erzeugt werden sollte, mitgeteilt, dass es sich bei der zu testenden Suchmaschine um ein Studentenprojekt handle. Teilnehmerinnen aus den beiden Versuchsgruppen mit hoher Erwartungshaltung wurde dieselbe Suchmaschine hingegen als professionelles Produkt einer IT-Firma vorgestellt, dessen Kaufpreis 20.000 Euro be-

Tabelle 1: Zweifaktorielles Untersuchungsdesign mit vier Faktorstufenkombinationen.

		System	
		gut	schlecht
Erwartung	niedrig	Gruppe 1 22 Vpn.	Gruppe 2 22 Vpn.
	hoch	Gruppe 3 22 Vpn.	Gruppe 4 23 Vpn.

trage. Gewiss handelt es sich hierbei um eine stark vereinfachte Form der Operationalisierung, die einen ersten Versuch darstellt, die Erwartungshaltung der Benutzer in den Evaluierungsprozess zu integrieren. Ein wesentlicher Vorteil dieser Vorgehensweise liegt darin, dass die Erwartung innerhalb der Gruppen konstant gehalten wird.

Jede Teilnehmerin sollte drei verschiedene Suchaufgaben bearbeiten. Zur Manipulation der Systemleistung wurden im Vorfeld der Untersuchung sechs unterschiedliche Ergebnislisten erzeugt. Für jede Aufgaben wurde eine Liste für das gute und eine Liste für das schlechte System erzeugt. Ergebnislisten für das schlechte System sind durch eine Precision von 0,5 und eine Average Precision von 0,55 gekennzeichnet, für das gute System wurden eine Precision von 0,6 und eine Average Precision von 0,75 gewählt. Zur Erstellung der Ergebnislisten mit einer vorgegebenen Average Precision wurde ein von Turpin und Scholer (2006) beschriebener Algorithmus verwendet.

Für die 23 Teilnehmerinnen aus Gruppe 4 in Tabelle 1 bedeutete dies beispielsweise, dass ihnen gesagt wurde, sie würden das professionelle System verwenden, tatsächlich erhielten sie jedoch die schlechteren Suchergebnisse.

4.1 Ablauf

Um versuchsleiterbedingte Verzerrungen zu vermeiden, erfolgte die Instruktion der Testpersonen in schriftlicher Form. Nach der Begrüßung erhielten die Probandinnen je nach Versuchsgruppe entweder den Informationstext für die hohe oder die niedrige Erwartungshaltung. Diese kurze Einführung diente dazu, alle Teilnehmerinnen mit dem Thema der Untersuchung vertraut zu machen. Weiterhin erhielten sie die Information, dass die Universität Hildesheim plane, eine neue Suchmaschine für Artikel aus Fachzeitschriften in der Bibliothek einzusetzen und dass diese im Rahmen dieses Benutzertests evaluiert werden solle. Alle Teilnehmerinnen wurden dazu aufgefordert, sich vorzustellen, sie seien Journalistinnen und recherchierten nach bereits veröffentlichten Presseartikeln, die das Thema ihres nächsten Beitrags beträfen. Dieses Szenario sollte von der Künstlichkeit der Testsituation ablenken und gleichzeitig den praktischen Zugang zum Thema erleichtern.

Alle Teilnehmerinnen bearbeiteten nacheinander drei Rechercheaufgaben. Um Lerneffekte zu kontrollieren, wurde die Reihenfolge der Aufgaben zwischen den einzelnen Versuchsteilnehmerinnen variiert. Wie auch schon in einer von Kaczmirek (2003) durchgeführten Studie konnten die Testpersonen ihre Suchbegriffe nicht frei wählen. Diese Einschränkung war erforderlich, da es sich bei dem Testsystem lediglich um die Simulation einer Suchmaschine handelte. Um Irritationen bezüglich dieser Einschränkung zu vermeiden, wurde im Einführungstext darauf hingewiesen, dass diese Maßnahme dazu diene, allen Testteilnehmerinnen die gleichen Anfangsvoraussetzungen zu ermöglichen.

Hatten die Versuchspersonen die Suchbegriffe in das Suchfeld eingegeben, erhielten sie je nach Untersuchungsbedingung eine der beiden vorgefertigten Trefferlisten. Erschien ihnen eines der Ergebnisse aufgrund der Kurzbeschreibung relevant zu sein, sollten die Testpersonen diesen Presseartikel im Volltext-Fenster öffnen und anschließend als relevant beziehungsweise nicht-relevant kennzeichnen. Pro Suchaufgabe standen den Probandinnen zehn Minuten Zeit zur Verfügung. Falls sie schon früher der Meinung waren, sich einen ausreichenden Überblick über das betreffende Thema verschafft zu haben, stand es ihnen frei, schon vorher mit der nächsten Aufgabe zu beginnen. Auch dieser Aspekt sollte der Künstlichkeit der Testsituation durch die Schaffung realistischerer Rahmenbedingungen entgegenwirken. Außerdem sollte auf diese Art und Weise das Entstehen von Zeitdruck vermieden werden.

Am Ende der Untersuchung wurden die Testpersonen gebeten, einen Fragebogen zur Bewertung der Suchmaschine auszufüllen. Als Dankeschön und Belohnung für die geopferte Zeit hatten alle Teilnehmerinnen die Möglichkeit, am Ende der Untersuchung an der Verlosung von drei Geldpreisen teilzunehmen.

4.2 Aufgaben

Die in der Untersuchung verwendeten Suchaufgaben entstammen der CLEF-2001- und der CLEF-2003-Testkollektion. Die beiden Kollektionen umfassen zwischen 50 und 60 verschiedene Topics sowie ca. 750.000 und 1.500.000 Millionen nach Relevanz bewertete Presseartikel (Braschler 2001, Braschler 2003). Die deutschsprachigen Dokumente zu den drei Suchaufgaben sind der nationalen Nachrichtenagentur der Schweiz *Schweizerische Depeschenagentur* (SDA), der deutschen Tageszeitung *Frankfurter Rundschau* (FR) und der deutschen Wochenzeitschrift *Der Spiegel* aus den Jahren 1994 und 1995 entnommen. Die Kurzbeschreibungen der einzelnen Suchaufgaben sind in Tabelle 2 dargestellt. Diese Zusammenfassungen wurden auch für die Instruktionstexte zu den einzelnen Rechercheaufgaben verwendet.

4.3 Testsystem

Um den Untersuchungsteilnehmerinnen eine möglichst realitätsnahe Anwendungssituation zu bieten, wurde für den Benutzertest ein Anwendungsprogramm entwickelt, das den Suchprozess einer realen Suchmaschine möglichst gut simulierte. Gestaltung und Funktionalität der graphischen Benutzeroberfläche orientierten sich an den derzeit bekannten Internet-Suchmaschinen. Der hierdurch intendierte Wiedererkennungseffekt sowie eine einfache Benutzerführung sollten eine weitestgehend intuitive Bedienung des Anwendungssystems bewirken. Dies musste gewährleistet sein, damit eventuelle Schwierigkeiten bei der Bedienung des Systems nicht zu einer ungewollten Störvariable werden konnten, die in der Folge die Ergebnisse der Untersuchung verfälscht hätte.

Das Testsystem verhielt sich tolerant gegenüber der Reihenfolge der vorgegebenen Suchbegriffe. Um Rechtschreibfehler abzufangen, wurden auch Eingaben akzeptiert, die bis zu einer Levenshtein-Distanz von sieben mit den vorgegebe-

Tabelle 2: Themen der Testaufgaben.

Nr.	Thema	Kurzbeschreibung
1	Erneuerbare Energien	Suche Dokumente, die die Nutzung von umweltfreundlicher Energie oder eine darauf ausgerichtete Politik betreffen, d.h. von Energie, die aus erneuerbaren Energiequellen erzeugt wurde.
2	Atomtransporte in Deutschland	Finde Berichte über Proteste gegen den Transport von radioaktivem Müll in Castor-Behältern in Deutschland.
3	Kinderarbeit in Asien	Finde Dokumente, die Kinderarbeit in Asien diskutieren und Vorschläge zu deren Beseitigung oder zur Verbesserung der Arbeitsbedingungen für Kinder liefern.

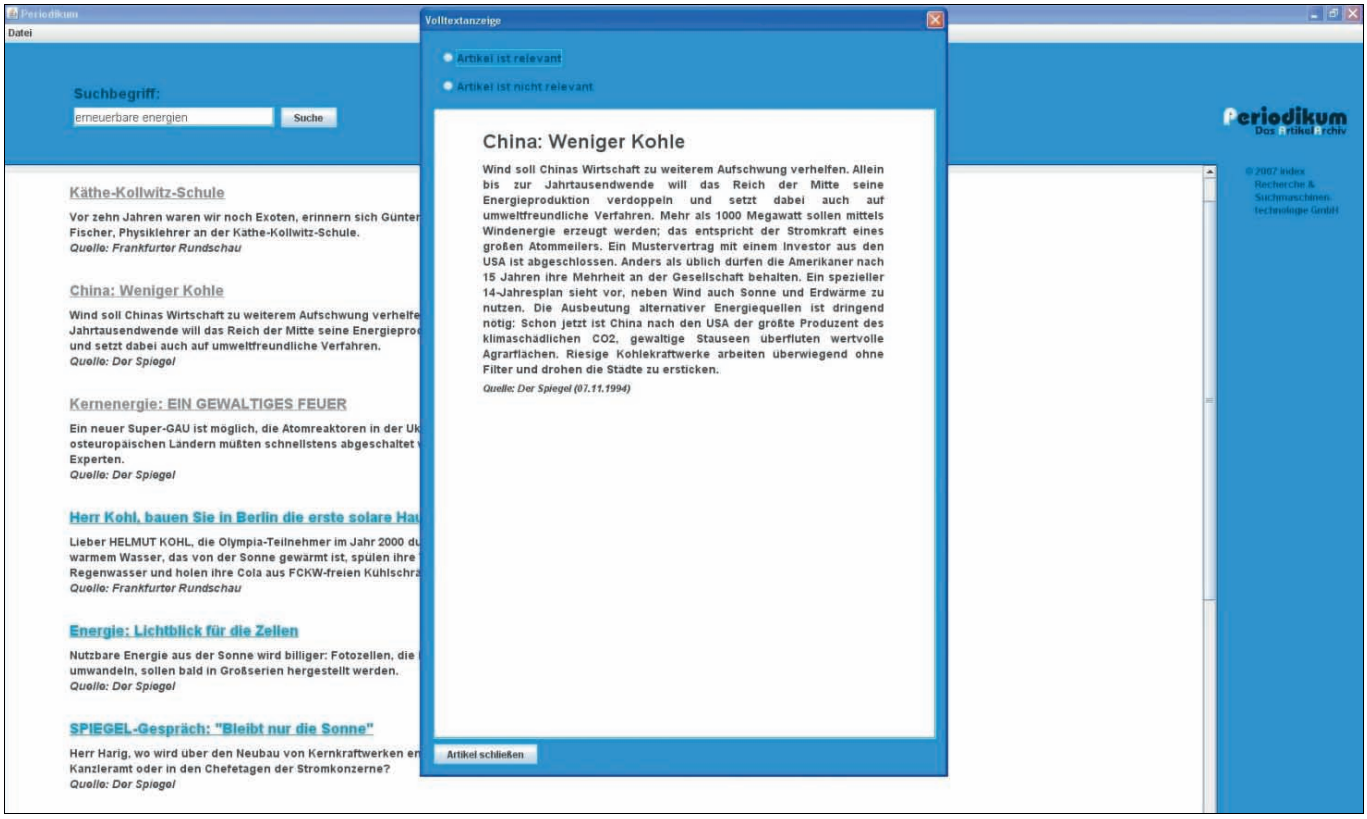

Abbildung 3: Benutzeroberfläche des Testsystems.

nen Suchbegriffen übereinstimmten. Dabei beschreibt die *Levenshtein-Distanz* zweier Wörter die minimale Anzahl der Löschungen, Einfügungen und Ersetzungen einzelner Buchstaben, die vorgenommen werden müssen, um die eine Buchstabenfolge in die andere zu überführen (Navarro 2001).
In der Ergebnisliste wurden zehn Treffer pro Seite angezeigt. Zu jedem Treffer wurden, neben dem Titel, der erste Satz und die Quelle angegeben. Die Entscheidung, den Textanfang als Kurzbeschreibung für die Dokumente zu wählen, beruhte auf den guten Ergebnissen der Testpersonen bei dieser Form der Ergebnispräsentation in der von Kaczmirek (2003) durchgeführten Studie.

4.4 Erhebungsmethoden

Die Messung der Zufriedenheit der Benutzer erfolgte über einen Fragebogen im Anschluss an den Test. Der Fragebogen besteht aus 28 Fragen zu den Oberthemen Ergebnisqualität und Gebrauchstauglichkeit der Suchmaschine sowie einem allgemeinen Teil, in dem nützliche Zusatzinformationen wie Computer- bzw. Suchmaschinenerfahrung abgefragt wurden. Die Zufriedenheitsfragen sind vergleichbar mit englischsprachigen Fragebogeninstrumenten wie z.B. dem End-User Computing Satisfaction (EUCS) Instrument von Doll und Torkzadeh (1988). Die Beantwortung der meisten Fragen erfolgte auf einer siebenstufigen Ratingskala mit Werten zwischen 1 (trifft vollkommen zu) und 7 (trifft überhaupt nicht zu). Auf diese Weise standen den Befragten je drei Abstufungen hinsichtlich Zustimmung und Ablehnung sowie eine neutrale mittlere Antwortkategorie zur Verfügung.

Die Leistung der Benutzer wurde in der vorliegenden Untersuchung mit fünf Leistungsmaßen aus anderen Benutzerstudien erfasst. Diese lassen sich in recall- und precision-orientierte Maße unterteilen. Das erste Leistungskriterium, RD, entspricht der Anzahl der korrekt relevanten Dokumente[8], die die Untersuchungsteilnehmerinnen innerhalb der vorgegebenen Bearbeitungszeit gefunden haben (Turpin & Scholer 2006). Das zweite recall-orientierte Leistungskriterium, im Weiteren als Benutzer-Recall (UR) bezeichnet, ergibt sich aus der Anzahl der von den Teilnehmerinnen gefundenen korrekt relevanten Dokumente geteilt durch die Gesamtzahl aller relevanten Treffer in der Ergebnisliste (Al-Maskari et al. 2006).
Die übrigen Kriterien messen die Benutzerleistung hinsichtlich der Genauigkeit der Relevanzbewertungen. Für das erste Maß wurde die Zeitspanne bestimmt, um das erste korrekt relevante Dokument zu finden, im Weiteren als TIME bezeichnet (Turpin & Scholer 2006). Das zweite Maß, die Benutzer-Precision (UP), ergibt sich aus der Anzahl der korrekt relevanten Dokumente geteilt durch die Gesamtzahl aller von den Teilnehmerinnen als relevant gekennzeichneten Dokumente (Al-Maskari et al. 2006).
Als weiteres Genauigkeitsmaß wurde in Anlehnung an die von Resnick und Lergier (2003) eingeführte pre-click confidence eine Pre-Click-Precision (PCP) erhoben. Zur Berechnung dieser Precision-Variante wird die Anzahl der korrekt relevanten Dokumente durch die Gesamtzahl der von den Teilnehmerinnen als möglicherweise relevant ausgewählten Treffer geteilt. Bei diesem Effektivitätsmaß wird also der erste Eindruck der Testpersonen erfasst, indem alle Dokumente, die im Volltext-Fenster geöffnet wurden, in die Berechnung der Benutzerleistung einbezogen werden.

5 Ergebnisse

Insgesamt nahmen 89 weibliche Testpersonen im Alter von 17 bis 32 Jahren an der Untersuchung teil. Bei 80 Prozent der Teilnehmerinnen handelte es sich um Studentinnen, die restlichen 20 Prozent entfielen auf andere Tätigkeiten wie gewerbliche Berufsausbildung oder Berufstätigkeit. 78 Probandinnen haben in der Woche vor dem Benutzertest an fünf bis sieben Tagen in der Woche mit dem Computer gearbeitet. Die durchschnittliche Internetnutzung der Teilnehmerin-

8 Als korrekt relevant werden im Folgenden Dokumente bezeichnet, die die Versuchsperson in Übereinstimmung mit den CLEF-Juroren als relevant bewertet hat.

nen betrug 16,7 (SD[9] ± 12,8) Stunden pro Woche.

Im Vorfeld der eigentlichen Auswertung wurden die Daten auf Störvariablen, die einen zusätzlichen Effekt auf die erhobenen Daten haben können, überprüft. Zwei Effekte sollen an dieser Stelle kurz berichtet werden. Wie auch in der Untersuchung von Turpin und Scholer hat eine einfaktorielle Varianzanalyse der Daten einen signifikanten Topic-Effekt festgestellt. Hierzu gingen die erhobenen Leistungsmaße der Benutzer zu den drei Aufgaben als Messwiederholungsfaktoren in die Analyse ein. Im Vergleich scheint Thema 2 etwas leichter zu bearbeiten gewesen zu sein als Thema 1. Das Vorhandensein von Topic-Effekten ist jedoch kein Nachteil, sondern trägt vielmehr zum Realismus der Studie bei. Schließlich weisen auch in einer realen Anwendungssituation nicht alle Suchanfragen den gleichen Schwierigkeitsgrad auf. Im Weiteren werden immer die über alle drei Suchaufgaben gemittelten Werte berichtet.

Wie bereits erwähnt, wurde die Reihenfolge der Suchaufgaben randomisiert, um auf diese Weise den Einfluss eventuell auftretender Lern- und Ermüdungseffekte zu kontrollieren. Auch für die Überprüfung eines möglichen Reihenfolgeeffekts kamen einfaktorielle Varianzanalysen mit der Bearbeitungsreihenfolge als unabhängigem Faktor zum Einsatz. Statistisch signifikante Effekte ergaben sich nur im Fall der Themen 1 und 3. Dieses Ergebnis stützt die Beobachtung, dass es sich bei Thema 2 um die am einfachsten zu bearbeitende Aufgabe handelt. Diese Überprüfung hat gezeigt, dass die Variation der Reihenfolge der Testaufgaben sinnvoll war, denn auf diese Weise wurde jedes Topic sowohl als Trainings- als auch als Abschlussaufgabe bearbeitet.

Hinsichtlich der Erwartungshaltung der Benutzer konnten keine signifikanten Unterschiede beobachtet werden. Eine mögliche Erklärung wäre, dass die Erwartung keinen Einfluss auf die Bewertung von Suchmaschinenergebnissen hat. Ebenso ist es jedoch denkbar, dass die Manipulation der Erwartungshaltung nicht ausreichend war. Diese Vermutung wurde auch durch einzelne Testpersonen im Anschluss an den Benutzertest bestätigt. Die Ergebnisse einer Folgestudie deuten jedoch vielmehr darauf hin, dass der Erwartungseffekt sich dynamisch verhält und mit der Zeit nachlässt (Lamm et al. 2010). Es scheint, als würden Benutzer ihre Erwartungen während der Suchmaschinennutzung an vorgefundene Gegebenheiten anpassen. Tatsächlich konnten in der zweiten Untersuchung nur signifikante Korrelationen zwischen Erwartungshaltung und Zufriedenheit für die zuerst bearbeitete Suchaufgabe festgestellt werden. Bei der zweiten Aufgabe hingegen schien sich die Systemleistung durchgesetzt zu haben, weshalb hier nur signifikante Korrelationen bezüglich der Systemleistung nachgewiesen werden konnten. Insgesamt legen diese Ergebnisse die Vermutung nahe, dass der Zeitpunkt der Zufriedenheitsmessung bei der ersten Untersuchung unglücklich gewählt wurde. Anstatt einer übergreifenden Messung nach der Bearbeitung aller drei Aufgaben, wäre es vermutlich besser gewesen, drei Einzelmessungen durchzuführen. Auch Szanja und Scamell (1993) konnten ein solches Abflauen des Erwartungseffekts beobachten.

Einen möglichen Erklärungsansatz für diese zeitverschobene Wahrnehmungsverzerrung liefert die in Abschnitt 3 vorgestellte Theorie der kognitiven Dissonanz. Möglicherweise führte die Diskrepanz zwischen erwarteter und wahrgenommener Systemleistung bei einigen Teilnehmerinnen zu einer Reduzierung dieses kognitiven Ungleichgewichts. Solch eine Reaktion könnte ursächlich für den fehlenden Unterschied in der Benutzerzufriedenheit sein. Darüber hinaus scheinen einige Teilnehmerinnen ihre Relevanzkriterien zu lockern bzw. zu verschärfen, wenn sie mit dem schlechteren bzw. besseren System arbeiten (siehe Abschnitt 5.2). Auch dieses Verhalten lässt sich als Maßnahme zum Ausgleich kognitiver Dissonanzen interpretieren.

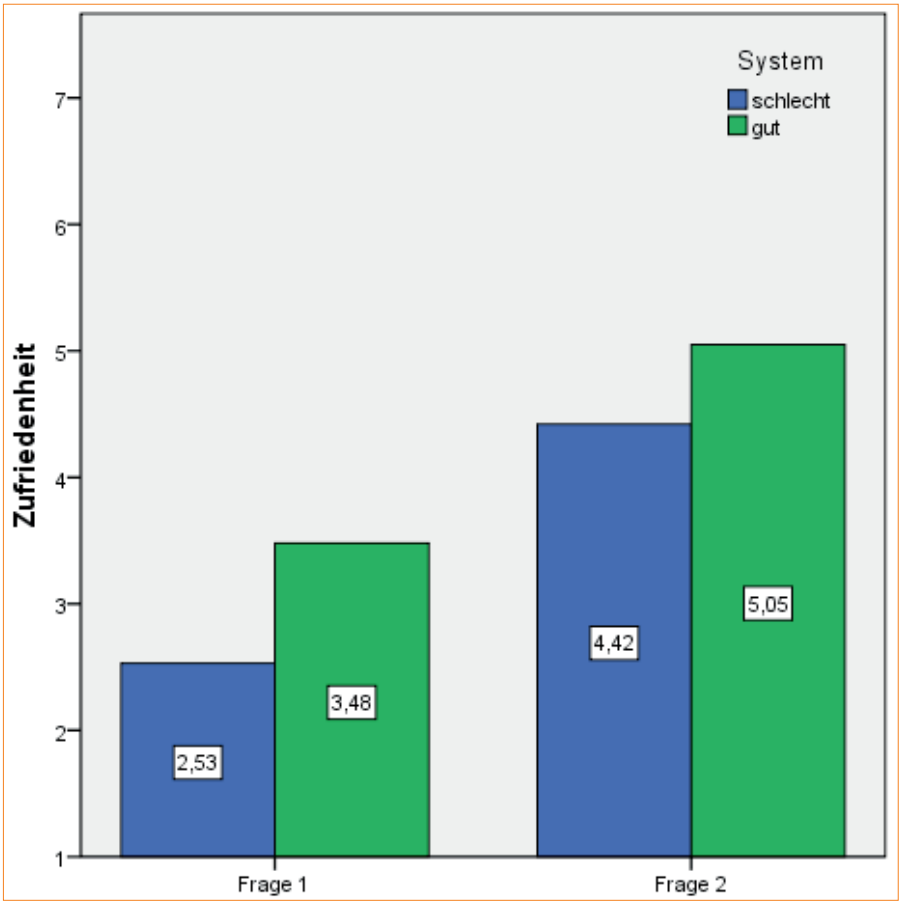

Abbildung 4: Signifikante Zufriedenheitsunterschiede für precision-orientierte Frageitems.

In Bezug auf die Systemleistung hingegen ließen sich sowohl im Fall der Benutzerzufriedenheit als auch im Fall der Benutzerleistung signifikante Unterschiede beobachten. Diese werden in den folgenden beiden Abschnitten näher dargestellt. Obgleich im Sinne der Theorie der kognitiven Dissonanz sowohl eine Anpassung der Erwartung als auch eine Anpassung der Wahrnehmung denkbar wäre, scheint die Richtung im vorliegenden Fall somit festzustehen. Angesichts der signifikanten Unterschiede hinsichtlich der Systemleistung scheinen die Teilnehmerinnen den Systemunterschied durchaus wahrzunehmen. Die nicht vorhandenen Unterschiede hinsichtlich der Erwartungshaltung hingegen deuten auf eine nachträgliche Regulierung der Ausgangserwartung hin.

5.1 Benutzerzufriedenheit

Um einen ersten Eindruck von den Daten zu gewinnen, wurden die fünfzehn Zufriedenheitsitems zunächst einzeln mit Hilfe von zweifaktoriellen Varianzanalysen ausgewertet. Dem zugrunde liegenden Untersuchungsdesign entsprechend bildeten die Systemleistung und die Erwartungshaltung die unabhängigen Variablen. Die Antworten der Probandinnen zu den einzelnen Fragen gingen jeweils als abhängige Variable in die

9 Standardabweichung

TREFFPUNKT BIBLIOTHEK

Information hat viele Gesichter

www.treffpunkt-bibliothek.de

TAUSEND VERANSTALTUNGEN

Bundesweite Bibliothekswoche

24. – 31. Oktober 2010

IN TAUSEND BIBLIOTHEKEN

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Deutschland Land der Ideen
Ausgewählter Ort 2010

Sponsoren
3pc Neue Kommunikation
borromedien
ekz bibliotheksservice
fachbuch journal
Sankt Michaelsbund Das katholische Medienhaus

Gesamtkoordination
dbv Deutscher Bibliotheksverband e.V.

Auswertung ein. Signifikante Gruppenunterschiede konnten nur bezüglich der folgenden beiden Items, bei welchen es um die Genauigkeit der Suchtreffer geht, nachgewiesen werden:

- Frage 1: Die Artikel hätten besser gefiltert werden können. ($p = 0{,}008$)
- Frage 2: Die meisten Artikel waren für die dazugehörigen Suchanfragen relevant. ($p = 0{,}025$)

In Abbildung 4 sind die Ergebnisse für diese beiden Items in einem Balkendiagramm dargestellt. Um die Antworten besser vergleichen zu können, wurde die Skala von Frage 2 invertiert, so dass nun höhere Werte für beide Items einer höheren Zufriedenheit entsprechen. Man kann erkennen, dass in beiden Fällen Teilnehmerinnen, die das bessere System verwendeten, zufriedener sind. Benutzer sind also tatsächlich in der Lage, Verbesserungen in der Systemleistung wahrzunehmen, auch wenn der absolute Unterschied nicht besonders groß ausfällt.

Obgleich in dieser ersten Analyse kaum signifikante Unterschiede nachgewiesen werden konnten, wurde in einem zweiten Analyseschritt versucht, eine gemeinsame Skala aus denjenigen Items zu bilden, die die Zufriedenheit mit den Suchergebnissen abfragten. Um die Qualität der resultierenden Skala zu testen, wurde eine Reliabilitätsanalyse mittels Cronbach's Alpha durchgeführt. Der beste Wert für Cronbach's Alpha (0,69) wird erreicht, wenn man folgende Fragen zu einer gemeinsamen Skala kombiniert:

- Frage 1: Die Artikel hätten besser gefiltert werden können.
- Frage 2: Die meisten Artikel waren für die dazugehörigen Suchanfragen relevant.
- Frage 3: Ich bin mit der Qualität der Suchergebnisse zufrieden.
- Frage 4: Die Präsentation der Ergebnisse war übersichtlich.
- Frage 5: Die Reihenfolge der Suchergebnisse spiegelte die Relevanz der Artikel wider.
- Frage 6: Die von mir aufgerufenen Artikel waren für die Recherche hilfreich.

Der signifikante Unterschied zwischen den beiden Systemlevels, der schon für die Einzelauswertungen der Fragen 1 und 2 festgestellt werden konnte, wird auch bei der Varianzanalyse für die neu gebildete Zufriedenheitsskala sichtbar ($p = 0{,}01$).

In Abbildung 5 sind die Mittelwerte für alle vier Versuchsgruppen aufgetragen. Wenn die Unterschiede in Bezug auf die Erwartungshaltung auch nicht statistisch signifikant sind, lässt sich tendenziell doch die durch das C/D-Paradigma vorausgesagte Beeinflussung der Benutzerzufriedenheit erkennen. Der Einfluss der Systemleistung ist am deutlichsten ausgeprägt und für beide Erwartungshaltungen ist die Zufriedenheit mit dem besseren System größer als mit dem schlechteren. Zusätzlich erscheinen die Angehörigen der Untersuchungsgruppen mit der niedrigen Erwartungshaltung im Durchschnitt zufriedener als die Testpersonen mit der hohen Erwartungshaltung. Gerade dies wird auch von dem C/D-Paradigma postuliert. Bei hoher Erwartungshaltung und niedrigem Systemlevel stimmen Soll- und Ist-Leistung nicht überein, was eine negative Diskonfirmation zur Folge hat. Entsprechend ist die betreffende Versuchsgruppe weniger zufrieden mit dem System als Testpersonen mit der niedrigen Erwartungshaltung. Der umgekehrte Effekt zeigt sich beim höheren Systemlevel. Hier erleben die Versuchsteilnehmerinnen mit der niedrigeren Erwartungshaltung eine positive Diskonfirmation, was sie das Suchsystem positiver beurteilen lässt als alle übrigen Probandinnen.

5.2 Benutzerleistung

Zur Auswertung der fünf in Abschnitt 4.4 beschriebenen Leistungsmaße wurden zweifaktorielle Varianzanalysen mit dem jeweiligen Maß als abhängiger und der Erwartungshaltung und der Systemleistung als unabhängiger Variablen durchgeführt. In Tabelle 4 sind deren Ergebnisse in Bezug auf die Haupt- und Wechselwirkungseffekte und in Tabelle 5 die entsprechenden Gruppenmittelwerte dargestellt. Sowohl für die Benutzer-Precision (UP) als auch für die Pre-Click-Precision (PCP) lässt sich ein signifikanter Einfluss der System- auf die Benutzerleistung nachweisen. Für die restlichen Maße zeigt keiner der Faktoren einen signifikanten Effekt ($p > 0{,}05$). In Bezug auf die recall-orientierten Maße scheinen Benutzer also in der Lage zu sein, den Unterschied in der Systemleistung zu kompensieren, was sich an den nichtsignifikanten Haupteffekten von RD und UR ablesen lässt.

Die signifikant niedrigeren Werte in der PCP bei Teilnehmerinnen, die mit dem schlechteren System gearbeitet haben, bei gleichzeitig nicht signifikant unterschiedlicher Anzahl insgesamt gefundener relevanter Dokumente (RD) (vgl. Tabelle 5), scheint im ersten Moment folgende Interpretation nahe zu legen: Die entsprechenden Versuchspersonen mussten mehr Dokumente öffnen, um die gleiche Anzahl korrekt relevanter Dokumente zu finden, wie Angehörige der Versuchsgruppe mit dem besseren System. Allerdings konnte ein derartiger Unterschied nicht gefunden werden. Stattdessen tritt hier ein Verstärkungseffekt

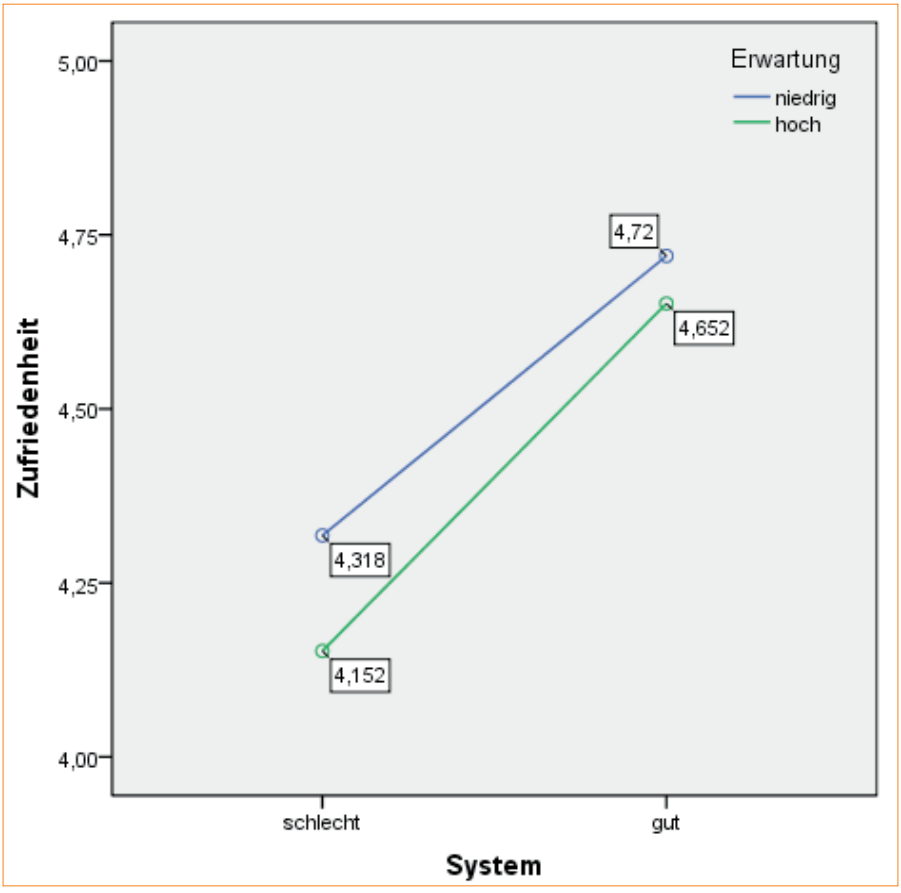

Abbildung 5: Voraussagen des C/D-Paradigmas spiegeln sich in den Daten wider.

Tabelle 4: Ergebnisse der zweifakatoriellen Varianzanalysen der Leistungsmaße.

Maß	Erwartung			System			Interaktion		
	df [1]	F [2]	p [3]	df	F	p	df	F	p
RD	1	2,87	0,09	1	0,47	0,50	1	2,73	0,10
UR	1	2,56	0,11	1	0,52	0,47	1	2,32	0,13
TIME	1	0,04	0,84	1	0,10	0,75	1	0,29	0,59
UP	1	0,49	0,49	1	13,05	**0,00**	1	4,82	0,03
PCP	1	0,94	0,34	1	4,42	**0,04**	1	1,01	0,32

[1] *Freiheitsgrade,* [2] *F-Wert,* [3] *Signifikanz*

Tabelle 5: Mittelwerte der Leistungsmaße.

	Erwartung		System		Interaktion			
	niedrig	hoch	schlecht	gut	Gruppe 1	Gruppe 2	Gruppe 3	Gruppe 4
RD	10,14	8,16	8,75	9,55	8,77	11,50	8,73	7,59
UR	0,20	0,16	0,19	0,17	0,19	0,21	0,19	0,14
TIME	440,99	443,62	440,23	444,38	435,42	446,56	445,04	442,20
UP	0,89	0,90	**0,86**	**0,93**	0,83	0,95	0,89	0,92
PCP	0,67	0,64	**0,62**	**0,68**	0,62	0,71	0,62	0,66

auf. Probandinnen mit dem schlechteren System finden (zumindest tendenziell) weniger korrekt relevante Dokumente als Angehörige der Vergleichsgruppe mit dem höheren Systemlevel. Gleichzeitig lässt sich bei den Probandinnen mit dem schlechteren System ein schwacher Trend zu einer größeren Anzahl an geöffneten Dokumenten beobachten. Für sich genommen ist keiner dieser Unterschiede statistisch signifikant. Für die Berechnung der PCP werden diese beiden Größen aber durcheinander geteilt. Die bei dem schlechteren System schon tendenziell geringere Anzahl gefundener relevanter Dokumente (RD) wird so noch durch die tendenziell größere Zahl angesehener Dokumente geteilt. Bei dem besseren System ist es genau umgekehrt. Die daraus resultierende Verstärkung der Unterschiede zwischen den beiden Systemen führt in der Konsequenz zu einer signifikanten Mittelwertdifferenz.

Das wohl interessanteste Ergebnis im Zusammenhang mit der Benutzerleistung ist der signifikante Unterschied in der Benutzer-Precision (UP). Im Folgenden soll analysiert werden, wie dieser signifikante Unterschied zu erklären ist. Die UP ist als der Quotient von RD und der Menge der als relevant markierten Dokumente (MR) definiert. Beachtet man, dass sich MR in die Summe aus RD und der Anzahl der fälschlicherweise als relevant markierten Dokumente (FR) zerlegen lässt, ergibt sich für UP:

$$UP = \frac{RD}{RD + FR} = \frac{RD}{RD\left(1 + \frac{FR}{RD}\right)} = \frac{1}{1 + \frac{FR}{RD}}$$

Man erkennt, dass die UP nur von der Größe $\frac{FR}{RD}$ abhängt. Damit wird der Unterschied in der Benutzer-Precision zwischen den beiden Systemlevels in erster Linie durch signifikante Mittelwertdifferenzen in der Menge FR verursacht. An der Benutzer-Precision ist also direkt eine restriktivere beziehungsweise weniger strenge Relevanzbewertung der beiden Versuchsgruppen abzulesen. Durch den tendenziellen Unterschied in der Zahl der korrekt relevant markierten Dokumente wird dieser Unterschied noch weiter verstärkt.

In Abbildung 6 sind diese unterschiedlichen Bewertungsstrategien noch einmal graphisch dargestellt. Wie anhand dieser Abbildung zu erkennen ist, haben Testpersonen des schlechteren Systems mehr Dokumente fälschlicherweise[10] als relevant bewertet als Testpersonen, die mit dem besseren System gearbeitet haben. Außerdem haben dieselben Testpersonen, im Vergleich zu Testpersonen mit dem besseren System, weniger Dokumente fälschlicherweise als irrelevant bewertet.

10 *Fälschlicherweise* bedeutet in diesem Fall im Gegensatz zu den CLEF-Juroren, deren Relevanzbewertung im Rahmen dieses Experiments als korrekt angenommen wurde.

Diese Ergebnisse lassen vermuten, dass Benutzer ihre Relevanzkriterien bis zu einem bestimmten Maß an die Qualität eines Systems anpassen. Je besser die Ergebnislisten, desto strenger scheinen Benutzer bei ihrer Relevanzbewertung vorzugehen. Ein ähnlicher Effekt wurde auch von Smucker und Jethani (2010) beobachtet. Da die Relevanzkriterien der Benutzer auch von ihren Vorerfahrungen und ihrer Erwartungshaltung abhängen (Jansen et al. 2007), ist es plausibel, dass auch hier Wahrnehmungsverzerrungen wie Mechanismen zum Ausgleich kognitiver Dissonanzen zum Tragen kommen können.

6 Fazit

Den Ausgangspunkt der vorliegenden Untersuchung bildete die Fragestellung, welche Wirkung die Qualität von Retrievalergebnissen auf den Sucherfolg der Benutzer einerseits und ihre Wahrnehmung des verwendeten Systems andererseits ausübt. Da in der Kundenzufriedenheitsforschung die Wahrnehmung eines Produkts eng mit der Erwartungshaltung des Konsumenten verknüpft ist, erhob sich weiterhin die Frage, ob dieser Einfluss auch in der IR-Evaluierung zu beobachten ist. Dazu wurde ein benutzerorientiertes Untersuchungsdesign entworfen, das die gleichzeitige Überprüfung beider Faktoren gestattete. Auf Benutzerseite wurden diesbezüglich die Benutzerleistung sowie die Benutzerzufriedenheit erfasst.

In Übereinstimmung mit anderen Studien konnte gezeigt werden, dass Benutzer teilweise in der Lage sind, systemseitige Leistungsunterschiede zu kompensieren und die Relevanzurteile von Benutzern nicht ohne den jeweiligen Kontext analysierbar sind. Die nicht-signifikanten Untersuchungsergebnisse hinsichtlich der Benutzererwartung sind vermutlich auf eine unglückliche Wahl des Zeitpunkts zur Messung der Zufriedenheit zurückzuführen. Gleichwohl sind die Vorhersagen des C/D-Paradigmas auch ohne signifi-

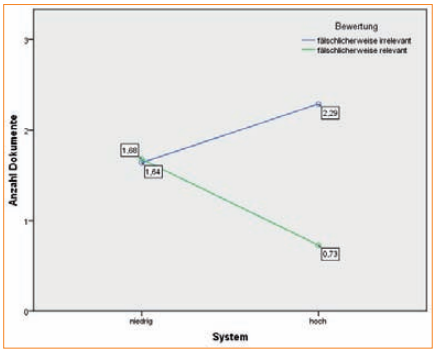

Abbildung 6: Systembedingte Anpassung der Relevanzkriterien.

kanten Erwartungseffekt in der Tendenz in den Daten erkennbar.

Zusammenfassend kann gesagt werden, dass die in der vorliegenden Arbeit gewählte interdisziplinäre Herangehensweise einen viel versprechenden Ansatz für die benutzerorientierte Evaluierung von IR-Systemen darstellt. Insbesondere die Einbeziehung der Erwartungshaltung als Determinante der Benutzerzufriedenheit sollte in diesem Zusammenhang weiter untersucht werden.

Danksagung

Für die Auszeichnung der diesem Artikel zugrunde liegenden Magisterarbeit mit einem VFI-Förderungspreis 2009 dankt die Autorin an dieser Stelle ganz herzlich dem Verein zur Förderung der Informationswissenschaft (VFI) mit Sitz in Wien.

Literatur

Al-Maskari, A.; Clough, P.; Sanderson, M. (2006): Users' Effectiveness and Satisfaction for Image Retrieval. In: Lernen - Wissensentdeckung - Adaptivität (LWA): Workshop Information Retrieval 2006 of the Special Interest Group Information Retrieval (FGIR). Hildesheim, Deutschland, 9.-11.10.2006, 84-88. http://web1.bib.uni-hildesheim.de/edocs/2007/521554985/meta/ [21.07.2008].

Al-Maskari, A.; Sanderson, M. (2006): The Effects of Topic Familiarity on User Search Behavior in Question Answering Systems. In: Lernen - Wissensentdeckung - Adaptivität (LWA): Workshop Information Retrieval 2006 of the Special Interest Group Information Retrieval (FGIR). Hildesheim, Deutschland, 9.-11.10.2006, 132-137. http://web1.bib.uni-hildesheim.de/edocs/2007/521554985/meta/ [21.07.2008].

Bollmann, P.; Cherniavsky, V. S. (1980): Probleme der Bewertung von Information-Retrieval-Systemen. In: Kuhlen, R. (Hrsg.): Datenbasen, Datenbanken, Netzwerke: Praxis des Information Retrieval. Bd. 3: Nutzung und Bewertung von Retrievalsystemen. München: Saur, S. 97-121.

Braschler, Martin (2002): CLEF 2001 - Overview of Results. In: Evaluation of Cross-Language Information Retrieval Systems: 2nd Workshop of the Cross-Language Evaluation Forum (CLEF 2001). Darmstadt, Deutschland, 03.-04.09.2001, Revised Papers. Berlin: Springer (Lecture Notes in Computer Science 2406), pp. 9-26.

Braschler, M. (2004): CLEF 2003 - Overview of Results. In: Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum (CLEF 2003). Trondheim, Norwegen, 21.-22.08.2003, Revised Papers. Berlin: Springer (Lecture Notes in Computer Science 3237), pp. 44-63.

Buckley, C.; Voorhees, E. M. (2005): Retrieval System Evaluation. In: Voorhees, E. M.; Harman, D. K. (Hrsg.): TREC: Experiment and Evaluation in Information Retrieval. Cambridge: MIT Press, pp. 53-75.

Doll, W. J.; Torkzadeh, G. (1988): The Measurement of End-User Computing Satisfaction. In: MIS Quarterly 12(2), pp. 259-274.

Festinger, L. (1978); Irle, M.; Mötmann, V. (Hrsg.): Theorie der kognitiven Dissonanz. Bern: Huber.

Harman, D. (1995): Overview of the Second Text Retrieval Conference (TREC-2). In: Information Processing & Management 13(3), pp. 271-289.

Homburg, C.; Giering, A.; Hentschel, F. (1999): Der Zusammenhang zwischen Kundenzufriedenheit und Kundenbindung. In: Bruhn, M.; Homburg, C. (Hrsg.): Handbuch Kundenbindungsmanagement: Grundlagen, Konzepte, Erfahrungen. 2., akt. u. erw. Aufl. Wiesbaden: Gabler, S. 81-112.

Ingwersen, P.; Järvelin, K. (2005): The Turn: Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer.

Jansen, B. J.; Zhang, M.; Zhang Y. (2007): The Effect of Brand Awareness on the Evaluation of Search Engine Results. In: CHI ,07 extended abstracts on Human factors in computing systems. San Jose, CA, USA, 28.04.-03.05.2007. New York: ACM, pp. 2471–2476.

Järvelin, K.; Ingwersen, P. (2004): Information Seeking Research Needs Extension toward Tasks and Technology. In: Information Research 10(1). http://informationr.net/ir/10-1/paper212.html [08.07.2010].

Kaczmirek, L. (2003): Information und Selektion: Gebrauchstauglichkeit der Ergebnisseiten von Suchmaschinen. Universität Mannheim, Fachbereich Psychologie, Dipl.-Arb.

Kowalski, G. (1997): Information Retrieval Systems: Theory and Implementation. Boston: Kluwer.

Lamm, K. (2008): Das Confirmation/Disconfirmation-Paradigma der Kundenzufriedenheit im Kontext des Information Retrieval. Universität Hildesheim, Fachbereich III - Informations- und Kommunikationswissenschaften, Institut für Angewandte Sprachwissenschaft, Mag.-Arb. http://publications.katrin-werner.com/magisterarbeit-lamm.pdf [28.07.2010].

Lamm, K.; Mandl, T.; Womser-Hacker, C.; Greve, W. (2010): User Experiments with Search Services: Methodological Challenges for Measuring the Perceived Quality. In: Proceedings of the 3rd International Workshop on Perceptual Quality of Systems (PQS) (erscheint).

Mandl, T. (2008): Recent Developments in the Evaluation of Information Retrieval Systems: Moving Toward Diversity and Practical Applications. In: Informatica - An International Journal of Computing and Informatics 32, 27-38. http://www.informatica.si/PDF/32-1/12_Mandl%20-%20Recent%20Developments%20in%20the%20Evaluation%20of...pdf [08.07.2010].

Möhr, M. (1980): Benutzerorientierte Bewertung von Information-Retrieval-Systemen. In: Kuhlen, R. (Hrsg.): Datenbasen, Datenbanken, Netzwerke: Praxis des Information Retrieval. Bd. 3: Nutzung und Bewertung von Retrievalsystemen. München: Saur, S.123-156.

Navarro, G. (2001): A Guided Tour to Approximate String Matching. In: ACM Computing Surveys 33(1), pp. 31-88.

Resnick, M. L.; Lergier, R. (2003): Task Specific User Strategies in On-line Search. In: Journal of E-Business 3(1), pp. 1-22.

Sauerwein, E. (2000): Das Kano-Modell der Kundenzufriedenheit: Reliabilität und Validität einer Methode zur Klassifzierung von Produkteigenschaften. Wiesbaden: Dt. Univ.-Verl.

Scharnbacher, K.; Kiefer, G.(1996): Kundenzufriedenheit: Analyse, Messbarkeit und Zertifizierung. München: Oldenbourg.

Smith, C. L.; Kantor, P. B. (2008): User Adaptation: Good Results from Poor Systems. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Singapore, 20.-24.07.2008, New York: ACM, pp. 147–154.

Smucker, M. D.; Jethani, C. P. (2010): Human Performance and Retrieval Precision Revisited. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Geneva, Switzerland, 19.-23.07.2010, New York: ACM, pp. 595-602.

Spiegel Online (2010). #0044CC: Microsoft findet das 80-Millionen-Dollar-Blau. http://www.spiegel.de/netzwelt/web/0,1518,684047,00.html [08.07.2010].

Szajna, B.; Scamell, R. W. (1993): The Effects of Information System User Expectations on Their Performance and Perceptions. In: MIS Quarterly 17(4), pp. 493-525.

Turpin, A. H.; Scholer, F. (2006): User Performance versus Precision Measures for Simple Search Tasks. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Seattle, Washington, USA, 06.-11.08.2006. New York: ACM, 11-18.

Womser-Hacker, C. (2004): Theorie des Information Retrieval III: Evaluierung. In: Kuhlen, R.; Seeger, T.; Strauch, D. (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Bd. 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis. 5., völlig neu gefasste Aufl. München: Saur, S. 227-235.

Benutzerforschung, Information Retrieval, Suchmaschine, empirische Untersuchung

Die Autorin

Katrin Werner, M.A.

ist seit Februar 2009 Doktorandin an der Universität Hildesheim und Absolventin des Studiengangs Internationales Informationsmanagement. Die hier vorgestellte Untersuchung basiert auf ihrer Magisterarbeit, für die sie 2009 mit dem VFI-Förderungspreis ausgezeichnet wurde. In ihrem Promotionsprojekt im Bereich des interaktiven Information Retrieval knüpft sie inhaltlich an diese Arbeit an. Katrin Werner ist daneben federführend im Projekt EFISU tätig, in dem es um den Aufbau eines Weiterbildungsangebots zum Thema Effektive Internetkommunikation und Suchmaschinenoptimierung für KMU geht.

Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Lübecker Straße 3
31141 Hildesheim
katrin.werner@uni-hildesheim.de