Masterarbeit

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmd
FACHBEREICH MEDIA
STUDIENGANG
INFORMATION SCIENCE & ENGINEERING

Messung von Relevanz in einem kontrollierten Information Seeking Experiment

im Studiengang
Information Science & Engineering / Informationswissenschaft
im Fachbereich Media
der Hochschule Darmstadt

Bearbeiter: Stefanie Reichert
Matrikel-Nr.: 710015

Referent:	Dr. Philipp Mayr
Korreferentin:	Prof. Dr. Heide Gloystein

Abgabe: 29.07.2011

Kurzfassung (deutsch)

In der vorliegenden Arbeit wird das Verhalten untersucht, das zu Relevanzentscheidungen führt. Bei der Analyse von Relevanz-Charakteristiken zeigt sich, dass es für das Konzept keine eindeutige Definition gibt, vielmehr hängt es von vielen verschiedenen Faktoren ab, was ein Nutzer unter Relevanz versteht und wie er seine Entscheidungen trifft. Im Rahmen eines explorativen Eyetracking-Experiments mit zwölf Studenten der Hochschule Darmstadt wurde untersucht, wann Relevanzentscheidungen fallen, auf Basis welcher Informationen und welche Faktoren auf die Relevanzentscheidung einwirken. Es zeigt sich, dass grundsätzlich zwei Kategorien von Bewertern zu unterscheiden sind. Anhand von vier Parametern, nämlich Länge der Gesamtbewertung, Anzahl der Fixationen, Anzahl der besuchten Datenelemente und Länge des Scanpfades können alle Nutzer der Studie entweder in die Gruppe der ökonomischen oder der gründlichen Bewerter eingeteilt werden.

Schlagwörter: Relevanz, Relevanzentscheidung, Informationssuche, Informationsverhalten, Experiment, Eyetracking

Abstract (englisch)

This study examines the behavior that leads to relevance judgements. The analysis of the nature of relevance shows that there is no single definition for the concept. How users interpret relevance and how they make decisions is influenced by a large number of factors. In an exploratory eyetracking experiment with twelve students from the University of Applied Sciences in Darmstadt, the author examines the point when decisions are made, the information that they're based on and what factors affect those judgments. Two categories of evaluators can be distinguished. Using four parameters, the overall length of the evaluation, the number of fixations, number of visited AOIs and length of the scanpath, all users of the study can be grouped into two kinds of evaluators, exhaustive and economic ones.

Keywords: relevance, relevance judgement, relevance judgement behaviour, information seeking, information seeking behaviour, user study, relevance assessment, eyetracking

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

IR	Information Retrieval
IIR	Interaktives Information Retrieval
ISB	Information Seeking Behavior
IS&R	Information Seeking and Retrieval
IT	Informationstechnik
SERP	Search Engine Result Page
AOI	Area of Interest
HMD	Head-Mounted-Display
RMV	Rhein-Main-Verkehrsverbund

1 Einführung

Relevanz ist eines der Kern-Konzepte der Informationswissenschaft. Alle Aktionen eines Nutzers laufen im Grunde genommen darauf hinaus, ein Informationsbedürfnis mit relevanten Ergebnissen zu befriedigen. Die Erwartungen der Nutzer von Informationssystemen sind dabei sehr hoch. Sie sollen mit einfachsten Suchanfragen so schnell wie möglich die relevantesten Ergebnisse - und dabei gleichzeitig möglichst wenig irrelevante Treffer – anzeigen. Die zunehmende Informationsflut erschwert dabei das Finden von relevanten Informationen spürbar. Es soll schließlich nicht irgendeine Information gefunden werden, sondern immer nur genau die Objekte, die zur Lösung eines bestimmten Problems in einem bestimmten Kontext beitragen. Die Leistung eines Informationssystems wird danach bewertet, wie gut es in der Lage ist, (potenziell) relevante Informationen bereitzustellen.

Durch die Entwicklung des World Wide Web und die enorm hohe Nutzung von Suchmaschinen im beruflichen, wissenschaftlichen sowie privaten Bereich ist jeder Internetnutzer damit konfrontiert, viele Male am Tag Relevanzurteile im Retrievalkontext fällen zu müssen. Durchschnittlich 50 Suchanfragen stellen deutsche Internetnutzer pro Woche (BITKOM 2010). Laut meiner eigenen Google Webhistorie stelle ich über 200 Anfragen pro Woche. Und dies betrifft nur Websuchmaschinen wie Google, Yahoo und Bing. Hinzu kommen die Recherchen in nicht frei zugänglichen Datenbanken.

Die Geschichte der elektronischen Informationssuche ist mit ca. 60 Jahren noch relativ jung. Relevanz und insbesondere Relevanzverhalten gehören zu den Bereichen der Informationswissenschaft, bei denen durch empirische Forschung zukünftig noch viele Wissenslücken geschlossen werden können und Informationssysteme optimal auf die Suchenden angepasst werden können. Je besser der Nutzer und sein Verhalten untersucht und verstanden werden, desto besser können auch Informationssysteme auf die unzähligen Faktoren eingestellt werden, die eine Informationssuche ausmachen. Schließlich ist es der Nutzer selbst, der die Entscheidung trifft welche ihm präsentierten Informationsobjekte in einer bestimmten Situation relevant sind oder nicht.

Für die vorliegende Arbeit wurde in einem kontrollierten Eyetracking Experiment das Relevanzverhalten von zwölf Studenten der Informationswissenschaft explorativ untersucht. Ziel war es herauszufinden, wann und auf welcher Grundlage Relevanzentscheidungen fallen, ob es bestimmte Muster gibt, die zu Relevanzentscheidungen

führen und durch welche Faktoren die Entscheidungen möglicherweise beeinflusst werden.

Im Grundlagen-Teil wird das Thema zunächst in seinen informationswissenschaftlichen Kontext eingeordnet, außerdem wird Eyetracking als Untersuchungsmethode für Interaktives Information Retrieval vorgestellt. Im Anschluss daran wird relevante Literatur ausgewertet, um den aktuellen Stand der Wissenschaft zu dem Thema Relevanz im Kontext der Informationssuche und -bewertung zu dokumentieren und einen Überblick über Konzepte, Untersuchungsmethoden und Forschungsergebnisse zu geben. Im dritten Teil werden das Eyetracking-Experiment und die daraus hergeleiteten Erkenntnisse dargestellt. Dazu werden Parameter wie die Länge der Bewertung, Anzahl besuchter Datenelemente, Zahl der Fixationen, Scanpfade und natürlich die Relevanzurteile selbst ausgewertet und interpretiert.

Ziel ist es, einen Beitrag zum Verständnis von Relevanzentscheidungen zu leisten, damit zukünftig Informationssysteme noch effizienter gemacht werden können.

2 Grundlagen

Da in der vorliegenden Arbeit die Untersuchung von Relevanz und Relevanzentscheidungen im Mittelpunkt steht, werden im folgenden Kapitel zunächst die zentralen Termini des Wissensgebietes „Information Behavior“ (Informationsverhalten) erklärt und voneinander abgegrenzt. Sie stehen zum Teil in sehr engen Verwandtschaftsbeziehungen und werden auch in der einschlägigen Literatur nicht immer mit der gleichen Bedeutung verwendet.

Da das im Rahmen der Arbeit durchgeführte Relevanz-Experiment mithilfe einer Eyetracking-Anlage durchgeführt wurde, wird zudem Eyetracking als Untersuchungsmethode für interaktives Information Retrieval sowie einige zentrale Studien aus dem Forschungsbereich vorgestellt.

2.1 Information Behavior

Information Behavior (Informationsverhalten) ist eine Teildisziplin der Informations- und Bibliothekswissenschaft. Sie ist ein weit umfassendes Wissensgebiet, welches nicht nur die aktive Informationssuche beinhaltet. Vielmehr ist Information Behavior ein Überbegriff für alle Interaktionen mit Informationen. Diese können ganz alltäglich sein und müssen sich nicht nur auf elektronische Informationssysteme beschränken. So gehören auch Aktivitäten wie Einkaufen, Urlaub buchen, Zugfahrplan lesen, zur Wahl gehen und Radio hören zum Information Behavior. Dies hebt Case (2007, S.5) in seiner Definition hervor:

> *“information behaviour […] encompasses information seeking as well as the totality of other unintentional or passive behaviors (such as glimpsing or encountering information), as well as purposive behaviors that do not involve seeking, such as actively avoiding information.”*

Auch Wilson (2000, S.49) erwähnt ausdrücklich, dass alle Arten des Umgangs mit Informationen eingeschlossen sind, unabhängig davon, ob der Nutzer diese bewusst oder unbewusst vollzieht. Laut seiner Definition spielt es auch keine Rolle ob und wie eine Person auf die Information reagiert, mit der sie konfrontiert wird:

> *“Information Behavior is the totality of human behaviour in relation to sources and channels of information, including both active and passive information seeking, and information use. Thus, it includes face-to-face*

communication with others, as well as the passive reception of information as in, for example, watching TV advertisements, without any intention to act on the information given."

Die Tatsache, dass auch Vermeidung von und unbewusster Umgang mit Information zum Informationsverhalten gehört, erinnert stark an Watzlawiks erstes Axiom der Kommunikationstheorie „Man kann nicht nicht kommunizieren". Will man dies auch für Informationsverhalten anwenden, könnte das Axiom ungefähr so lauten „man kann nicht nicht mit Informationen umgehen". Da jedes Verhalten kommunikativen Charakter hat und Kommunikation letztlich nichts anderes als der Austausch von Signalen, also Informationen im weitesten Sinne ist, hat auch jedes Verhalten informatorischen Charakter.

Da Information Behavior so ein weit gespannter Begriff ist, schließt er zudem die Konzepte des Information Seeking Behavior (Informationsrechercheverhalten) und Information Search(ing) Behavior (Informationssuchverhalten) mit ein, wie in Wilsons Nested Model of Conceptual Areas dargestellt ist (Abb. 1).

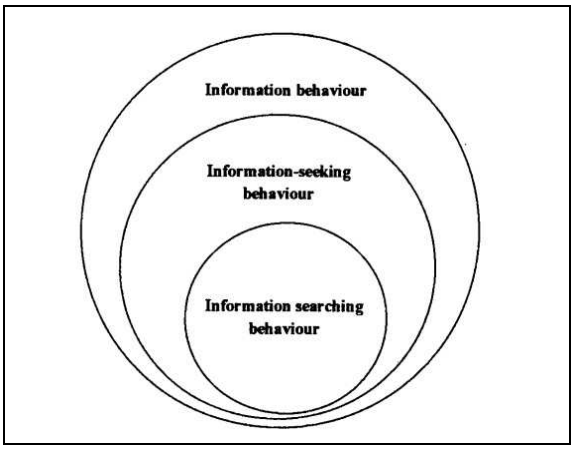


Abbildung 1: Nested Model of Conceptual Areas Quelle: (Wilson 1999a, S. 840)

2.2 Information Seeking Behavior

Information Seeking Behavior (Informationsrechercheverhalten) ist der am stärksten untersuchte Bereich des Informationsverhaltens. Er verengt den Fokus auf Verhalten, Motivation und Vorgehen des Benutzers bei der Recherche nach Informationen. Wie oben erwähnt, wird die Suchaktivität durch ein Informationsbedürfnis ausgelöst, welches ebenfalls Gegenstand der Untersuchungen ist. Außerdem wird geprüft, wodurch es ausgelöst wird und wie es befriedigt werden kann. Information Seeking Behavior ist

folglich die Informationssuche mit einem bestimmten Zweck zur Befriedigung eines Informationsbedürfnisses, wie Wilson (2000, S.49) beschreibt:

> *"Information Seeking Behavior is the purposive seeking for information as a consequence of a need to satisfy some goal. In the course of seeking, the individual may interact with manual information systems (such as a newspaper or a library), or with computer-based systems (such as the World Wide Web)."*

Schon seit Ende der 1960er Jahre versuchen Wissenschaftler das Informationsrechercheverhalten von Menschen zur Lösung eines Problems in Theorien und Modellen zu konzeptionalisieren. Drei der wichtigsten Modelle, die sich im Information Seeking Kontext bewährt haben und Basis für zahlreiche Weiterentwicklungen wurden, sind Wilsons „Model of Information Seeking Behaviour" von 1981, Ellis „Behavioural Model of Information Seeking Strategies" von 1989 sowie Kuhlthaus „Information Seeking Process" von 1991. Alle drei Modelle haben gemeinsam, dass sie den Fokus der Betrachtung auf den Nutzer und nicht auf das Informationssystem richten.

Wilson, dessen Modell von ihm selbst und durch andere mehrfach verändert und erweitert wurde, zeichnete eine grobe Karte der Information Behavior Landschaft und ihrer wichtigsten Bestandteile. Informationsrechercheverhalten wird demnach durch einen Informationsbedarf ausgelöst. Um diesen zu befriedigen, setzt der Nutzer seinen Bedarf in Suchaktivitäten um und stellt Anfragen an ein elektronisches Informationssystem oder andere informelle Informationsquellen. Diese Suche nach relevanten Informationen resultiert entweder in Erfolg oder Misserfolg. Die Anwendung von relevanten Informationen kann sowohl dazu führen, dass der Informationsbedarf gestillt wird, kann aber auch in einem veränderten Bedarf resultieren. Außerdem können durch Wissenstransfer und Informationsaustausch weitere Menschen miteinbezogen werden (Abb. 2, überarbeitete Version des 1981er-Modells).

David Ellis identifizierte auf Basis von empirischen Untersuchungen allgemeine Kategorien in der Vielfalt komplexer Verhaltensmuster im Informationsrechercheverhalten (vgl. Ellis 2009):

- Starting: Ausgangspunkt der Suche, Auswahl der Informationsquelle oder Literatursuche
- Chaining: Verfolgen von Fußnoten und Zitationen, vorwärts und rückwärts
- Browsing: halb-gezielte Suche in potenziell relevanten Gebieten durch Verfolgen von Inhaltsverzeichnissen, Titellisten, Überschriften, Namen von Organisationen und Personen, Zusammenfassungen, Weblinks etc.

- Differentiating: Filtern durch Beurteilung von Informationsquellen nach ihrer Art, Qualität, Wichtigkeit, Brauchbarkeit
- Monitoring: Sich auf dem neusten Wissensstand halten, Entwicklungen auf dem Gebiet verfolgen
- Extracting: relevantes Material identifizieren.

Carol Kuhlthau fügte den Phasen des Suchprozesses von Ellis zwei weitere Dimensionen hinzu, nämlich Gefühle und Gedanken. Ihr Ziel war es, Denkvorgänge und Empfindungen vor während und nach der Suche zu verstehen.

Sinn der Modelle ist es, Benutzerverhalten auf allen Ebenen zu jeder Zeit des Suchprozesses zu verstehen, um die Kenntnisse bei der Gestaltung von Informationssystemen wieder einsetzen zu können und so die Suche zu verbessern.

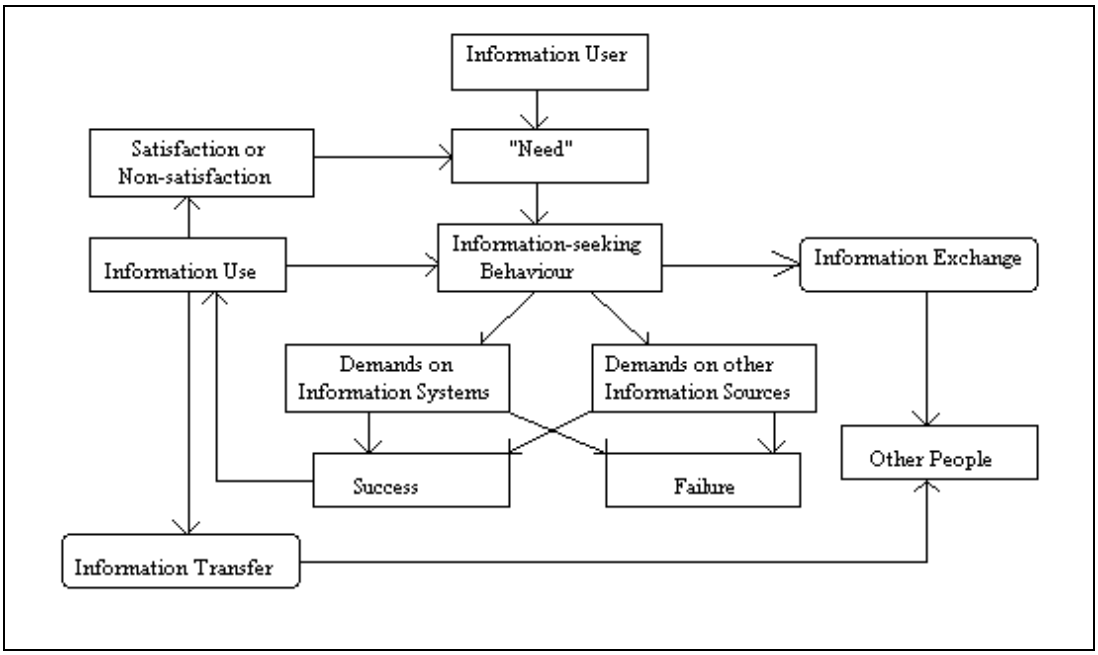


Abbildung 2: Wilsons Information Behaviour Model Quelle: (Wilson 1999b)

2.3 Information Search(ing) Behavior

Den engsten Blickwinkel auf Informationsverhalten hat das Information Search(ing) Behavior (Informationssuchverhalten). Es beschreibt das Verhalten beim Information Retrieval in elektronischen Quellen wie Datenbanken und wird daher auch als Retrieval Behavior bezeichnet. Information Searching Behavior umfasst zum Beispiel Auswahl und Eingabe der Suchterme und –Operatoren, Reformulierung von Suchanfra-

gen sowie die Anwendung verschiedener Suchstrategien. Im Zentrum des Information Searchings steht die Interaktion des Benutzers mit den Informationsquellen. Dies umfasst vor allem die Mensch-Computer-Interaktion. Laut der Definition von Wilson (2000, S.49) ist die Bewertung von Relevanz Teil des Information Searching Behaviors:

> ***"****Information Searching Behavior is the 'micro-level' of behavior employed by the searcher in interacting with information systems of all kinds. It consists of all the interactions with the system, whether at the level of human computer interaction (for example, use of the mouse and clicks on links) or at the intellectual level (for example, adopting a Boolean search strategy or determining the criteria for deciding which of two books selected from adjacent places on a library shelf is most useful), which will also involve mental acts, such as judging the relevance of data or information retrieved."*

Modelle, die die Interaktion mit IR-Systemen untersuchten und die Forschung maßgeblich beeinflusst haben, sind:

- Ingwersens Cognitive Model,
- Belkins Episode Model und
- Saracevics Stratified Interaction Model.

2.4 Interaktives Information Retrieval

Klassisches Information Retrieval (IR) ist ein Fachgebiet, welches sich allgemein mit der Repräsentation, dem Speichern, Suchen, Finden, Aufbereiten und Präsentieren von potenziell relevanten Informationen in elektronischen Ressourcen beschäftigt. Es werden objektiv messbare Funktionalitäten eines Retrievalsystems wie Schnelligkeit der Suche, Recall und Precision sowie Optimierung von Such- und Rankingalgorithmen untersucht. Das Interaktive Information Retrieval (IIR) hingegen stellt, statt des technischen Systems, den Nutzer in den Vordergrund. Untersuchungsgegenstand ist die Interaktion zwischen dem Nutzer und dem Informationssystem und seine subjektiven Wahrnehmungen während des Suchprozesses. Zu den beiden Hauptakteuren IR-System und Nutzer kommen noch weitere Einflussfaktoren hinzu wie z. B. der soziokulturelle Kontext. Ingwersen und Järwelin (2005, S.21) definieren Interaktives Information Retrieval folgendermaßen:

> *"The interactive communication processes that occur during retrieval of information by involving all major participants in IS&R, i.e., the searcher, the socio-organizational context, the IT setting, interface and information space."*

Verhalten, das zu Relevanzentscheidungen führt, ist sehr eng mit Information Seeking (und Searching) Behavior verwandt, da Relevanzentscheidungen bei der Selektion von Dokumenten essenzieller Teil des Gesamt-Recherche-Prozesses sind. Information Seeking und Searching Behavior sind wiederum eng verwandt mit interaktivem Information Retrieval. Der interaktive Faktor zeigt sich in der Einflussnahme des Users, welche auf verschiedenen Ebenen passieren kann (z. B. Reformulierung einer Suchanfrage, weil zu wenig relevante Dokumente angezeigt werden). Folglich finden sich Informationen zu Relevanzverhalten in allen genannten Teilbereichen.

2.5 Eyetracking

Eyetracking bzw. Blickrichtungsmessung, Lesezeitenanalyse und Lautes Denken zählen zu den kognitionspsychologischen Untersuchungsmethoden. Sie beschäftigen sich mit der Analyse von Prozessen, die an kognitiven Leistungen wie Gedächtnis, Denken und Sprachverstehen beteiligt sind sowie mit der Struktur von Wissensrepräsentationen, die diesen Prozessen zugrunde liegen. „Blickbewegungen und Blickrichtung sind gut messbare Verhaltensaspekte, die unmittelbar zu kognitiven Prozessen in Beziehung gesetzt werden können." (Bente 2005, S. 320)

Eyetracking ermöglicht es in Echtzeit zu verfolgen, wohin ein User schaut, während er eine bestimmte Aufgabe erfüllt und dabei zum Beispiel mit einem Computer interagiert. Die Augen- und Blickbewegungen werden von Sensoren registriert. Über eine Webcam kann außerdem Mimik und Gestik der Tester dokumentiert werden. Häufig läuft zusätzlich eine Audioaufnahme, um Kommentare aufzuzeichnen. Alle Aktionen, die die Testperson vollzieht, zum Beispiel Mausklicks, Scrollen und Texteingaben werden ebenfalls erfasst.

Es gibt drei Ansätze wie Eyetracking-Anlagen aufgebaut sein können. Beim Headset sind eine Augen- und eine Außenkamera integriert, die einen freien Blick in die Umwelt ermöglichen. Beim Head-Mounted-Display (HMD) wird das zu betrachtende Material (Stimulus) im HMD selbst angezeigt und bleibt somit immer in einer konstanten Relation zum Kopf der Testperson. Eine dritte Möglichkeit, die auch für diese Eyetracking-Studie verwendet wurde, ist ein freistehendes System bei dem die Kameras und Sensoren in der Nähe eines Monitors oder direkt darin integriert sind (Abb. 3). Auflö-

sung und Messgenauigkeit sind beim zuletzt genannten System geringer als bei den vorherigen.

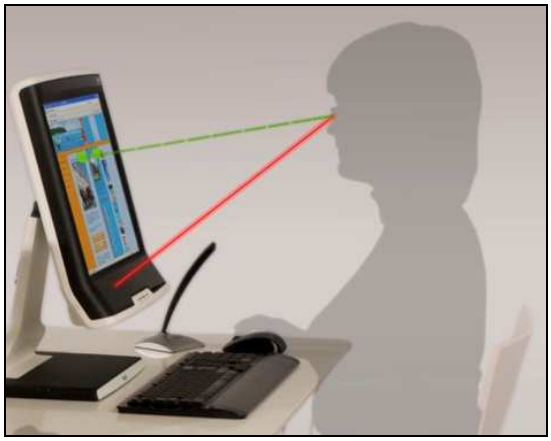


Abbildung 3: Tobii Eyetracking Quelle: (Henrici 2010)

Bei den aufgezeichneten Daten unterscheidet man Fixationen und Sakkaden. Bei den **Fixationen** ruht das Auge für etwa 200-300 Millisekunden auf einem bestimmten Bereich der Webseite. Diese Punkte weisen auf Stellen hin, denen der Benutzer Aufmerksamkeit schenkt und an denen Informationen wahrgenommen werden können. **Sakkaden** hingegen beschreiben Sequenzen, in denen die Augen schnell und ruckartig von einer Fixation zur nächsten springen. Während der Sakkaden, die 25-100 Millisekunden dauern, erfolgt keine Verarbeitung der visuellen Reize. Daher werden sie in Eyetracking-Studien meistens nicht berücksichtigt. Es gibt noch weitere Indikatoren für kognitive Prozesse, die mittels Eyetracking gemessen werden können, die aber für die vorliegende Arbeit nicht relevant sind. Dazu gehören beispielsweise Pupillendilatation, Lidschlagfrequenz und Sakkadenlänge.

Eine Sequenz von Fixationen und Sakkaden bildet den **Scanpfad** (scanpath). Mit seiner Hilfe lässt sich die Reihenfolge der Fixationen visualisieren. Beim Scanpfad werden die Fixierungen durch Kreise repräsentiert. Ihr Durchmesser beschreibt die Intensität, die im Kreis angezeigte Zahl entspricht der Position in der Blicksequenz (Abb. 4). So kann im Nachhinein analysiert werden, welche Blickfolge einem Klick oder, im konkreten Fall, einer Relevanzentscheidung vorausgegangen ist. Weitere Visualisierungsmethoden sind zum Beispiel die Heatmap (Wärmebild) bei der durch Farbverteilung der Bereich der höchsten Fixierungsintensität rot angezeigt wird, oder die Fokus Maps bei denen die fixierten Bereiche auf einem schwarzen Hintergrund „freigewischt“ werden.

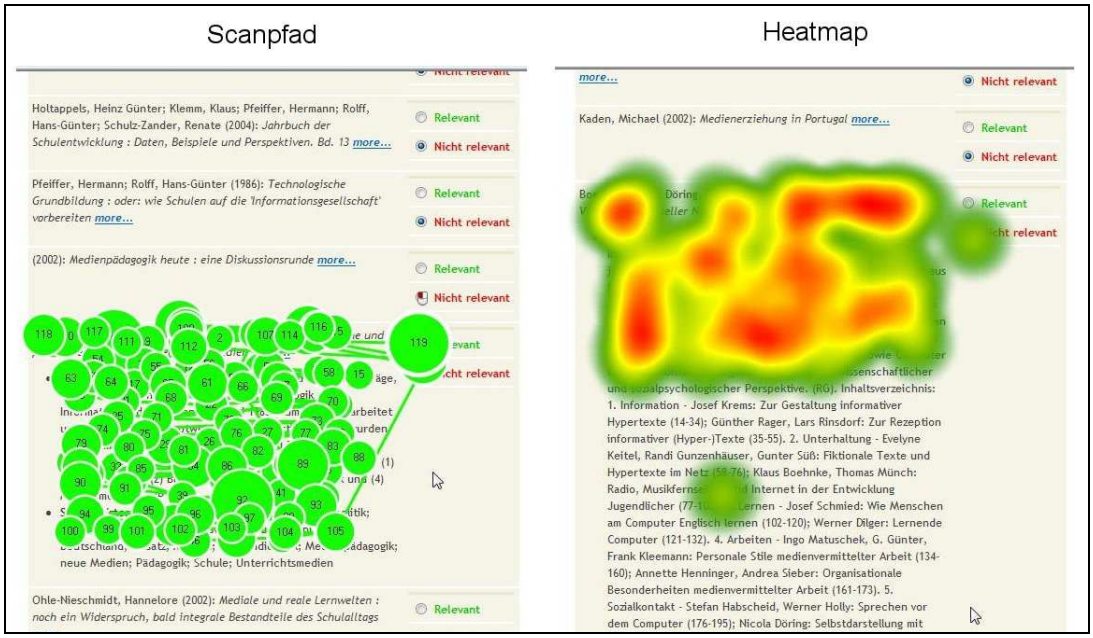


Abbildung 4: Beispiele für Scanpfad und Heatmap Quelle: Eigene Darstellung

Um bestimmte Bereiche einer Szene genauer zu untersuchen, müssen Interessensbereiche, sogenannte **Areas of Interest** bestimmt werden. Für diese definierten Sektionen, auch Lookzones genannt, können dann statistische Daten wie Anzahl und Dauer der Fixationen ermittelt werden.

Da das Zusammenspiel von Mensch und Computer genau erfasst werden kann, dienen Eyetracking-Analysen in der Wirtschaft häufig dem Testen der Benutzerfreundlichkeit von Webseiten, Software oder Spielen. Probanden bekommen zum Beispiel zwei verschiedene Versionen einer Webseite vorgelegt. Anhand der Blickerfassung, Mausklicks und nicht zuletzt durch lautes Denken während der Webseiten-Nutzung kann der Auftraggeber ermitteln lassen, welche Version den Probanden mehr zusagt und welche Version intuitiver zu bedienen ist. Die Nutzung zu Marktforschungszwecken ist aber nur eines von vielen Anwendungsbeispielen. Eyetracker werden beispielsweise auch in der Medizin, Neurowissenschaften, Psychologie und Informatik eingesetzt.

Stärken und Schwächen

Mittels einer Eyetracking-Anlage kann besonders gut die Interaktion von Mensch und Maschine beobachtet und analysiert werden. Statt sich darauf zu verlassen was User im Nachhinein nach einer Suche über ihr Verhalten berichten, erlauben Eyetracker die Erfassung von Echtzeit-Daten. Sie liefern zusätzliches Feedback über Informationsbedürfnisse, die über das hinausgehen, was die Nutzer in eine Suchzeile eingeben oder zu Logfile-Analysen und Post-Search-Interviews.

Vorteil von modernen Eyetrackern ist, dass die Testpersonen sich nicht ständig bewusst sind, dass sie aufgezeichnet werden. Die Anlagen sind, sofern freistehend, sehr unauffällig. So können sich die Testpersonen schnell an die Situation gewöhnen und sich so natürlich wie möglich verhalten. Noch vor einigen Jahren mussten die Testpersonen zum Teil Apparaturen auf dem Kopf tragen. Heute sind Eyetracker mobil, leicht und einfach zu bedienen.

Zu den Schwächen gehört, dass man nur sieht, *wohin* die Blicke gerichtet werden, aber man weiß nicht, was tatsächlich wahrgenommen wird. Wichtige Informationen in der Peripherie des Sichtfeldes können wahrgenommen werden, ohne dass sie fixiert werden. Ohne lautes Denken während des Tests oder anschließende Interviews bleibt unklar, *wieso* etwas Aufmerksamkeit erregt hat (vgl. Richter 2008 und Bente 2005).

Die meisten Eyetracking-Programme ermöglichen es zwar Scanpfade zu visualisieren, jedoch nicht sie automatisch vergleichen zu lassen. Die Analyse von Blicksequenzen ist derzeit ein noch nicht vollständig gelöstes Problem in der Eyetracking-gestützten Forschung. Es gibt die Möglichkeit Sequenzen über die Levenshtein Distanz zu vergleichen, d. h. es werden Werte errechnet für die Einfüge-, Lösch- und Ersetz-Aufwände bei der Umwandlung von einer Zeichenkette in die andere. Um überhaupt einen Blickverlauf in eine Zeichenkette zu überführen, müssen Lookzones definiert sein, welche mit Buchstaben oder Zahlen bezeichnet werden. Eine Blicksequenz beschreibt dann die Reihenfolge, in der die Lookzones besucht wurden. Eine Software, die die Levenshtein Distanz automatisch ausrechnen kann, ist eyePatterns (http://sourceforge.net/projects/eyepatterns/). Andere Ansätze arbeiten mit Übergangswahrscheinlichkeiten zwischen verschiedenen Stimulusarealen unter Zuhilfenahme von Markow-Ketten. „Die hieraus abzuleitenden Darstellungsmöglichkeiten und Analyseansätze sind vielversprechend, erfordern jedoch noch weitere Forschungsarbeit.“ (Bente 2005, S. 319).

Auch in einem jüngeren Buchkapitel über Eyetracking und Online-Suche stellen die Autoren fest, dass der Vergleich von Scanpfaden und die Erstellung einer „Durchschnittssequenz“ Probleme sind, die aktuelle Eyetracking-Programme noch nicht beherrschen:

> *„One key limitation of current commercial eye monitoring software is hat there is little support for analysing and discerning patterns in the scanpaths themselves. At present, eye monitoring data is typically visualised either as an aggregate view of what users looked at, as in heatmaps, or as individual scanpaths. […] they lack the ability to convey*

what an "average" sequence looks like, how typical a given path is compared to another, or where common subsequences lie." (Granka/Feusner/Lorigo 2008, S. 355).

2.6 Eyetracking und Interaktives Information Retrieval

Eyetracking wird seit ca. 2003 genutzt, um Suchverhalten im Internet zu analysieren. Eyetracking-Studien in diesem Kontext beziehen sich häufig auf die Analyse des Interaktionsverhaltens von Suchmaschinennutzern mit Ergebnislisten von Websuchmaschinen wie Google (Search Engine Result Page, SERP), wie gesucht wird, welche Bereiche vom User angeschaut und geklickt werden und letztlich auch wie eine Relevanzentscheidung fällt. In den letzten Jahren hat die Untersuchungsmethode an Aufmerksamkeit gewonnen. Im folgenden Abschnitt werden einige zentrale Erkenntnisse aus Eyetracking-Studien im IIR-Kontext vorgestellt.

Golden Triangle

Im Jahr 2005 beschrieben die Marketingfirmen Enquiro und Did-it in Zusammenarbeit mit der Firma Eyetools das bekannte F-Schema oder auch „Golden Triangle" (Abb. 5), welches die Bereiche, die Suchmaschinennutzer auf den Ergebnisseiten bevorzugt beachten, in einer Heatmap darstellt. Demnach schenken die User den ersten drei Ergebnissen sehr viel Aufmerksamkeit, den nachfolgenden Ergebnissen dagegen kaum. Das F-Schema wurde in späteren Studien auch für Webseiten bestätigt.

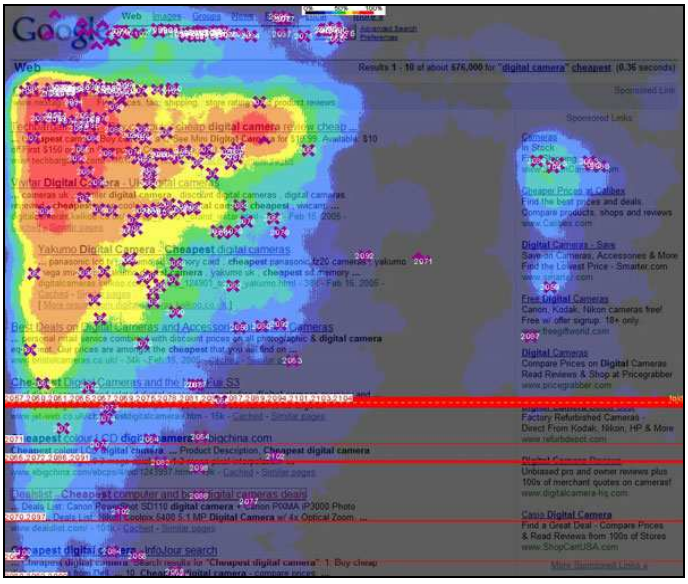


Abbildung 5: Google's Golden Triangle Quelle: (Enquiro et al. 2005, S.7)

Das Golden Triangle galt lange als Referenz, an der sich Suchmaschinenoptimierer und Webdesigner orientierten. Da sich die Ergebnisliste von Google in der Zwischenzeit durch die Anzeige von Bildern, Videos, Maps-Daten usw. gewandelt hat, ist das F-Schema aber zumindest für die Ergebnisseiten von Suchmaschinen nicht mehr ohne Einschränkungen gültig. Darüber hinaus wurde in den folgenden Jahren auch festgestellt, dass das Nutzerverhalten weitaus komplexer ist und durch mehr Faktoren beeinflusst wird, als das F-Schema berücksichtigt (vgl. Enquiro et al. 2005).

Einfluss des Aufgaben-Typs

2008 untersuchten Papaeconomou, Zijlema und Ingwersen mittels Eyetracking (und anschließenden Interviews) ob es einen Zusammenhang zwischen Lernstilen (Global and Sequential Learners) von 15 Testpersonen und deren Relevanzbewertungen von Webseiten gibt. Dazu wurden unter anderem „relevance hot spots“ untersucht, also Bereiche der Webseiten, denen die Testpersonen besonders viel Aufmerksamkeit schenkten. Dabei kamen sie zu der Erkenntnis, dass es weniger die Lernstile waren, die Einfluss auf Relevanzentscheidungen hatten, als die Art der Aufgaben. Sie schlugen vor in zukünftigen Tests mehrere verschiedene Aufgabentypen zu berücksichtigen, nicht nur zwei wie in ihrer Studie.

Bei der elektronischen Suche nach Informationen über eine Suchmaschine kann man nach Broder (2002) drei Aufgaben-Typen unterscheiden:

1. die navigatorischen Aufgaben, bei denen es das Ziel ist, eine bestimmte Webseite oder URL zu finden,
2. die informatorische Suche, bei der eine bestimmte Information gefunden werden soll, die sich aber auch auf mehreren Webseiten befinden kann und
3. die transaktionale Suche, bei der die User eine Suche ausführen mit dem Ziel ein Produkt zu kaufen.

Lorigo et al. (2006) untersuchten mit Eyetracking-Daten von 23 Testpersonen ob es Unterschiede bei der kognitiven Wahrnehmung der verschiedenen Aufgabentypen nach Broder gibt und ob sie Auslöser für unterschiedliche Suchstrategien sein können. Dabei fanden sie heraus, dass informatorische Aufgaben durchschnittlich mehr Aufwand und Zeit beanspruchen, als navigatorische Aufgaben. Jedoch hielten sich die Nutzer bei den informatorischen Aufgaben länger auf den angeklickten Webseiten an sich auf, als auf der SERP.

Für die navigatorischen Aufgaben hielten sich die Nutzer länger auf den Ergebnisseiten auf. Was die Scanpfade der Testpersonen angeht, konnten dagegen keine Unterschiede für die beiden Aufgabentypen gefunden werden. Jedoch unterscheiden sich die Scanpfade von Männern und Frauen. Männliche Testpersonen neigten demnach dazu Suchergebnisse eher linear zu betrachten (siehe „Lesetypen“ S.22). Sie schauten sich außerdem mehr Ergebnisse und Ergebnis-Seiten an als die weiblichen Testpersonen.

Saito, Terai und Egusa (2009) untersuchten ebenfalls den Einfluss des Aufgaben-Typs und der Erfahrung des Users auf Information Seeking Behavior im Web. Aufgaben in der Studie waren „Bericht schreiben“ und „Ausflug planen“. Dabei griffen sie nicht nur auf die Eyetracking-Daten zurück, sondern werteten zusätzlich Fragebögen, Logfiles, Think-Aloud-Protokolle und Post-Experiment-Protokolle aus. Die Aufgaben-Typen betreffend konnten keine Unterschiede im Verhalten festgestellt werden, stattdessen bemerkten die Autoren einen Zusammenhang zwischen der Erfahrung eines Nutzers und seinem Suchverhalten. So hielten sich die weniger erfahrenen Studenten länger auf Nicht-Ergebnisseiten auf, als die erfahreneren User und schauten sich auch eher rangtiefe Ergebnisse an. Die Autoren weisen darauf hin, dass durch die geringe Teilnehmerzahl von elf Personen zwar Zusammenhänge festgestellt werden können, aber keine verlässlichen Rückschlüsse gezogen werden können.

Ein ähnliches Problem hatten Liu et al. (2010), die in ihrer Eyetracking Studie ebenfalls Zusammenhänge zwischen Verhalten und Aufgaben-Typ erkennen konnten, diese aber aufgrund der geringen Zahl von Teilnehmern nicht verallgemeinern wollten. Ein eindeutiges Ergebnis war jedoch, dass sie erste Hinweise darauf erhalten haben, dass anhand des Nutzerverhaltens verschiedene Aufgabenfacetten wie zum Beispiel die Komplexität abgeleitet werden können.

Einfluss der dargestellten Informationen

Cutrell und Guan (2007) haben den Einfluss der Informationen in den „Snippets“ auf den Ergebnisseiten untersucht. Unter Snippets versteht man ein- bis zweizeilige Textausschnitte in den Suchergebnissen, die meistens das gesuchte Wort oder die gesuchte Phrase der Anfrage enthalten und sie im Kontext anzeigen. Die These war, dass größere Textausschnitte den Nutzern bei der Beurteilung der Relevanz einer Webseite helfen, bevor sie angeklickt wird und somit das Klicken überflüssig macht. Die Autoren haben herausgefunden, dass längere Snippets mit zusätzlichen Informationen für informationelle Suchanfragen hilfreich sind, während bei navigationalen Suchanfragen mit kurzen Snippets die beste Performance erreicht werden konnte.

Längere Snippets zogen die Aufmerksamkeit der Nutzer auf sich, während gleichzeitig die URL vernachlässigt wurde, welche zum schnellen Entscheiden bei navigatorischen Aufgaben hilfreich gewesen wäre.

Einfluss des Ranges

Die Autorengruppe rund um Lori Lorigo und Laura Granka führten von 2004 bis 2008 drei Studien zum Thema Nutzerverhalten auf Suchmaschinen-Ergebnisseiten durch. Die Autoren untersuchten mittels Eyetracking was der Nutzer tut und was er liest, bevor er tatsächlich ein Dokument auswählt. Sie verglichen unter anderem die durchschnittliche Zeit, die User damit verbringen sich einzelne Ergebnisse zu betrachten mit der Anzahl der Male, in denen diese Dokumente ausgewählt (angeklickt) wurden. Sie interessierten sich außerdem für den Einfluss des Aufgaben-Typs sowie den Einfluss weiterer Nutzercharakteristiken wie dem Geschlecht. Während Aufgabentyp und Geschlecht eher geringen Einfluss auf das Nutzerverhalten haben, erkannten die Wissenschaftler aber, dass besonders der Rang von Dokumenten eine wichtige Rolle spielt. 96 % der Testpersonen schauten sich zum Beispiel nur die erste Seite der SERP (mit 10 Ergebnissen) an und hier vorwiegend die ersten beiden Ergebnisse. Die Analyse der Blickverläufe zeigte, dass keine weiteren Ergebnisse mehr angeschaut wurden, wenn die Top drei keine relevanten Dokumente enthielten. Durchschnittlich wurden insgesamt nur drei bis fünf Abstracts überhaupt fixiert. Die ersten beiden Suchergebnisse wurden fast gleich lang betrachtet, das erste Ergebnis aber sehr viel häufiger angeklickt. Nach dem zweiten Suchergebnis nahm die Fixations-Dauer stark ab.

In einem weiteren Versuch wurden die Ergebnislisten so manipuliert, dass die Dokumente der ersten Seite in umgekehrter Reihenfolge angezeigt wurden. Trotzdem klickten die Testpersonen das Abstract auf Rang eins favorisiert an, obwohl es objektiv nicht am relevantesten war. Die Autoren bezeichneten dies als „trust bias“ (Joachims/Granka/Gay 2005, S. 154). Dem Ranking der Suchmaschine wird vertraut. In der dritten Studie wurde das Verhalten bei der Nutzung von Google und Yahoo Suchen verglichen. Es konnten hier jedoch keine Unterschiede festgestellt werden (vgl. Lorigo et al. 2008).

Bewertungstypen

Aula et al. (2005) identifizierten ebenfalls in einer Eyetracking-Studie zwei verschiedene Kategorien von Bewertungstypen: die Ökonomischen („economic evaluators“)

und die Gründlichen („exhaustive evaluators“). Die ökonomisch handelnden Nutzer trafen ihre Entscheidungen schneller und auf Basis von weniger Informationen, als die gründlichen Nutzer. Letztere wogen erst mehrere Optionen ab und benötigen mehr Informationen, bevor sie ein Resultat in der Ergebnisliste tatsächlich anklickten.

In Abbildung 6 sind exemplarisch jeweils zwei Blickpfade von beiden Nutzertypen dargestellt. Es ist deutlich zu erkennen, dass die economic evaluators viel mehr Fixationen vor einer Entscheidung hatten als die economic evaluators. Für die Studie wurden 28 Testpersonen untersucht. Da die ökonomischen User erfahrener im Umgang mit Computern waren, folgerten die Autoren, dass sich der Bewertungs-Stil mit zunehmender Erfahrung von exhaustive zu economic entwickelt.

Die economic evaluators waren außerdem effizienter bei den Suchaufgaben, woraus die Autoren schlussfolgerten, dass es von Vorteil sein könne schneller jene Resultate anzuklicken, die vielleicht relevant sind, anstatt sorgfältig das beste Resultat zu suchen.

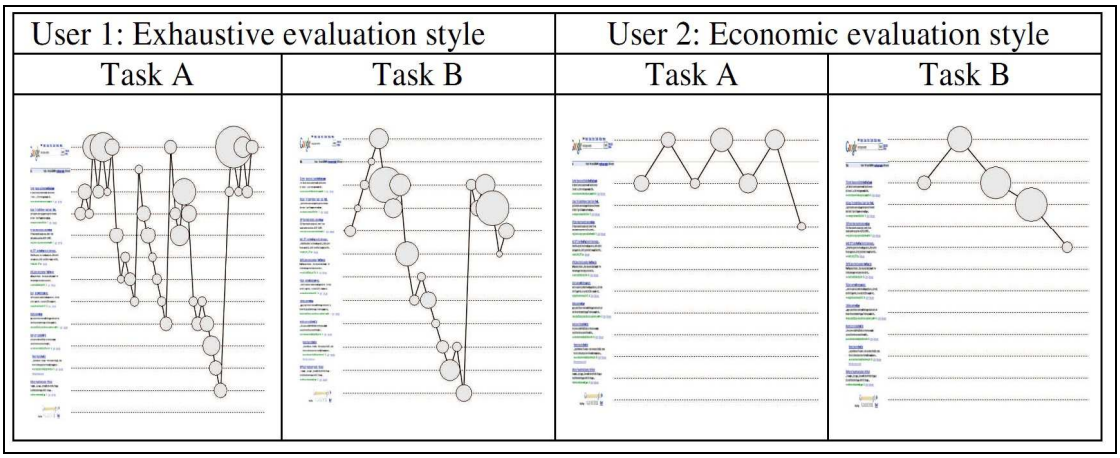


Abbildung 6: Scanpfade der Bewertungsytpen Quelle: (Aula et. al 2005, S. 1060)

Lesetypen

Eine andere Art der Klassifizierung von Benutzertypen wählten Granka et al. (2008). Sie analysierten den Blickverlauf während der Interaktion mit Ergebnislisten und konnten drei Klassen des Leseverhaltens ausmachen:

1. „Nonlinear scanning“: Beim non-linearen Lesen werden die Ergebnisse nicht von oben nach unten der Reihenfolge nach betrachtet, sondern in willkürlicher Abfolge.

2. „Linear scanning“: Beim linearen Lesen werden die Abstracts der Reihe nach betrachtet. Es ist aber auch möglich, schon vorher gelesene Ergebnisse erneut zu betrachten.
3. „Strictly linear scanning“: Das streng-lineare Lesen schließt aus, dass Ergebnisse mehrfach angeschaut werden. Auch dann nicht, wenn sie zu einem früheren Zeitpunkt bereits betrachtet wurden.

Nur ein Fünftel der Testpersonen betrachtete die SERP in linearer oder streng-linearer Weise in der Reihenfolge, wie die Ergebnisse angezeigt werden. Beim Rest konnte man Sprünge und Auslassungen („skips and jumps“) beobachten.

Nutzerverhalten als implizites Relevance Feedback

Implizite Feedback Techniken sind eine vielversprechende Möglichkeit Retrievalperformance durch Relevance Feedback zu verbessern. Sie erheben Daten indirekt vom User, indem sie die (unbewussten) Verhaltensweisen während der Suche überwachen. Die Relevanz eines Dokumentes wird passiv ermittelt, d. h. um von Relevance Feedback profitieren zu können, muss der User keinen zusätzlichen Aufwand betreiben.

Jarkko Salojärvi beschäftigt sich seit einigen Jahren mit proaktivem Information Retrieval. Das Informationssystem soll hierbei alle möglichen Informationen nutzen, sei es implizites oder explizites Feedback, um mehr relevante Dokumente zu finden und sich generell so an den individuellen User anzupassen. Salojärvi konzentriert sich dabei vor allem auf Augenbewegungs-Daten, die aus Eyetracking-Experimenten gewonnen wurden. So schafften seine Mitarbeiter und er es, Relevanzbewertungen nur anhand von Augenbewegungen zu einem gewissen Grad vorauszusagen (vgl. Salojärvi et al. 2003).

Moe, Jensen und Larsen (2005) untersuchten drei Eyetracking-Merkmale auf ihr Potenzial für implizites Relevance Feedback. Von den drei Merkmalen Gesamtzeit der Bewertung, Gründliches Lesen und Rückschritte, identifizierten sie die Zeit, die ein User damit verbringt sorgfältig und umfassend zu lesen, statt Text nur zu überfliegen oder kurz anzuschauen, als Merkmal das am ehesten geeignet sein könnte Hinweise auf Relevanz zu liefern. „The results indicate, that the feature thorough reading have the potential to identify relevant information as input for implicit relevance feedback […]” (S. 45)

3 Relevanz

Im vorherigen Kapitel ging es darum, die vorliegende Arbeit in den Information Behavior Kontext einzuordnen und Eyetracking als Untersuchungsmethode kennenzulernen. Da im anschließenden Experiment das Verhalten im Mittelpunkt steht, welches zu einer Relevanzentscheidung führt, sollen in diesem Kapitel Eigenschaften und Erscheinungsformen von Relevanz sowie Methoden zu deren Untersuchung genauer beschrieben werden.

Relevanz kann als treibende Kraft der Informationswissenschaft verstanden werden. In fast allen Artikeln zu dem Thema wird sie als „central concept“ oder „key notion“ bezeichnet. Eine der hauptsächlichen Motivationen Informationen aufzubereiten, sie formal zu erfassen und inhaltlich zu erschließen, ist es, sie irgendwann wiederfinden zu können. Das heißt in einer Situation, in der die Information von Nutzen sein kann. Dieser Nutzen wird repräsentiert durch Relevanz. Ist ein Objekt relevant, dann können wir es „gebrauchen“. Relevanz zu verstehen ist daher essenziell für das Design von Systemen, die eben solche – relevante – Dokumente so treffsicher wie möglich finden sollen. Dass das weit weniger einfach ist, als man denkt, formuliert Tefko Saracevic nach vielen Jahren in denen er sich mit dem Thema auseinandergesetzt hat, folgendermaßen: "relevance became a key notion (and key headache) in information science" (1999, S. 1058). Der Klammerzusatz lässt erahnen, dass es sich bei Relevanz um etwas handeln muss, was nicht in einem Satz erklärt werden kann.

3.1 Charakterisierung von Relevanz

Viele Wissenschaftler haben das Phänomen Relevanz untersucht und versucht es in Klassen einzuteilen, Merkmale herauszuarbeiten und Prozesse in Modellen abzubilden. Es gibt keine einheitliche, eindeutige Definition für Relevanz im Information Seeking Kontext, aber viele Versuche sich dem anzunähern. Ingwersen und Järwelin (2005, S.21) bezeichnen Relevanz in ihrem Definitionsversuch als „assessment“, d. h. sie schließen die Entscheidung und Bewertung mit ein:

> *“The assessment of the perceived topicality, pertinence, usefulness or utility, etc., of information sources, made by cognitive actor(s) or algorithmic devices, with reference to an information situation at a given point of time. It can change dynamically over time for the same actor.*

Relevance can be of a low order objective nature or of higher order, i.e., of subjective multidimensional nature."

Relevanz ist demnach nicht etwas, was einem Dokument statisch zugehörig ist, sondern entsteht erst dann, wenn durch einen Nutzer eine Bewertung stattfindet, wie nützlich das Dokument in einer bestimmten Situation ist. Dies ist nachvollziehbar, wenn man sich vorstellt, dass das ein Dokument für einen bestimmten User für eine bestimmte Fragestellung relevant sein kann, für ein anderes Problem zu einem anderen Zeitpunkt aber nicht. Ingwersen und Järwelin verdeutlichen dies durch die Einschränkung „at a given point of time".

Relevanz beschreibt Beziehungen

In seinem viel zitierten Artikel über Relevanz klassifiziert Tefko Saracevic (1975) die verschiedenen Herangehensweisen an die Konstruktion eines Relevanz-Frameworks. Er erkannte, dass es bei einem Definitionsversuch vor allem darauf ankäme, welche Faktoren mit einbezogen würden und in welchen Beziehungen diese Faktoren zueinander stünden.

Er kam zu folgender Gleichung als Zusammenfassung: „Relevance is the A of a B existing between a C and a D as determined by an E" (S. 328).

A	*B*	*C*
measure degree dimension estimate appraisal relation	corresondence utility connection satisfaction fit bearing matching	document article textual form reference information provided fact

D	*E*
query request information used point of view information requirement statement	person judge user requester information specialist

Abbildung 7: Faktoren in der Definition von Relevanz Quelle: (Saracevic 1975, S. 328)

Für die Faktoren A bis D können dabei verschiedene Begriffe (Abb. 7) verwendet werden. Je nach Kombination ergeben sich so sehr viele verschiedene Blickwinkel auf das Konzept Relevanz. Schon hier wurde der Grundstein gelegt Relevanz nicht als statische Qualität, sondern als multidimensionales System zu sehen.

Relevanz ist ubiquitär

Relevanzentscheidungen ziehen sich durch das tägliche Leben. Jeder, der schon einmal nach einem Urlaub sein E-Mail-Postfach aufräumen musste, hat Selektionen vorgenommen, ohne sich wahrscheinlich direkt bewusst gewesen zu sein, dass auch dies eine Relevanzbewertung ist (siehe Definition „Information Behavior“, Kapitel 1). Jede Interaktion mit Information läuft darauf hinaus relevante Ergebnisse zu erzielen. “Somewhere, somehow, the invisible hand of relevance, under its own or other names, enters the picture in all information activities and a great many information systems.” (Saracevic 2007a, S. 1916).

Relevanz ist dynamisch

Die Wahrnehmung von Relevanz kann sich im Laufe des Suchprozesses verändern und entwickeln (vgl. Definition Ingwersen und Järwelin oben: „It can change dynamically over time for the same actor.“). Sowohl innerhalb einer Session als auch darüber hinaus. Da sich auch das Informationsbedürfnis durch neue Erkenntnisse ständig verändern kann, können sich auch die Relevanzurteile wandeln. Dokumente, die am Anfang eines Arbeitsprozesses noch als relevant eingestuft werden, können in fortgeschrittenen Stadien der Arbeit ihren Status verlieren. Je mehr man über ein Thema lernt, desto selbstbewusster wird man bei der Entscheidung.

Tang und Solomon (2001) fanden außerdem heraus, dass sich auch die Relevanzkriterien während eines Suchprozesses verändern können. Ähnlich wie sich Informationsbedürfnisse, Suchanfragen und kognitive Gegebenheiten über die Zeit entwickeln und reifen, so können sich auch die Relevanzkriterien und deren Gewichtungen wandeln.

Dass Relevanz nicht konsistent ist, beweist auch die Tatsache, dass zwei Gruppen von Bewertern Dokumente zur selben Suchanfrage völlig unterschiedlich beurteilen können. Saracevic (2007b) schätzt, dass die Übereinstimmung zweier Nutzer bei ca. 30% liegt. Werden noch mehr Nutzer oder Nutzergruppen hinzugefügt, verringert sich die Schnittmenge noch mehr (Saracevic 2007b).

Relevanz ist subjektiv

„Individually (and not at all surpringsingly), people differ in relevance inferences, just as they differ in all other cognitive processes in general, and involving information in particular“ (Saracevic 2007b, S.2131).

Die Auffassung davon, ob etwas relevant ist, hängt maßgeblich von der Persönlichkeit des Suchenden, seinem Vorwissen (zum Beispiel Kenntnis des Wissensgebietes und des Informationssystems), seiner Motivation und auch vom Verständnis der Aufgabe ab. Menschliches Verhalten ist unendlich facettenreich und unter anderem bestimmt dadurch, wie und wodurch ein Mensch im Laufe seines Lebens geprägt worden ist. Bei einer Hausbesichtigung würde beispielsweise ein Feuerwehrmann auf andere Dinge achten als ein Koch. Diese Prägungen wirken sich auf alle Lebensbereiche aus, so auch auf Relevanzentscheidungen. Ein Urteil kann ebenso von Prioritäten, Vorlieben und aktuellen Interessen bestimmt werden, wie auch von der Herkunft des Suchenden sowohl im geografischen Sinne und dem Kulturkreis, als auch bezogen auf den wissenschaftlichen Hintergrund, z. B. mit welchen erlernten Paradigmen er sucht usw. (vgl. Socio-Cognitive Theory von Birgir Hjørland).

Entscheidungen werden außerdem beeinflusst durch Persönlichkeitstyp, Arbeitstyp, Wahrnehmungsprägung, Erinnerungen, aber auch Befinden, Abneigungen, Stress, Desinteresse, Ablenkung, schnell fertig werden Wollen, innere Haltung (Geisteshaltung), Tagesform uvm.

Relevanz ist relativ

Manche Dokumente sind für eine bestimmte Aufgabe relevanter als andere. In einem neuen Kontext kann der Grad der Relevanz derselben Dokumente genau andersherum sein. Nur der User selbst kann darüber urteilen, ob ein Dokument sein Informationsbedürfnis befriedigt oder nicht. Was für einen User relevant erscheint, kann für einen anderen User weniger oder sogar gar nicht relevant sein, z. B. wenn der derjenige das Dokument schon kennt.

Relevanz ist intuitiv

Zwar lässt sich Relevanz nicht einfach definieren, dennoch muss man Nutzern eines Informationssystems nicht sagen, was relevant ist. Sie wissen instinktiv was sie brauchen und was nicht. Im natürlichen Suchprozess benötigen Nutzer keine Anleitung,

die ihnen vorgibt, woran man Relevanz erkennen kann. Das hat Tefko Saracevic schon in seinem Artikel von 1975 festgestellt: „Intuitively, we understand quite well what relevance means. It's a primitive ‚y' know concept', as is information for which we hardly need a definition" (S. 324).

Relevanz ist messbar

Die Tatsache, dass manche Dokumente für einen Nutzer relevanter sind als andere, lässt erahnen, dass es möglich ist, Relevanz in Abstufungen zu unterteilen. Dies wurde schon früh in diversen Studien untersucht, indem verschiedene Bewertungsskalen zur Messung des Grades von Relevanz benutzt wurden. Eine Bewertung kann binär oder nicht-binär (z. B. anhand von gleitenden Skalen oder Stufen) erfolgen. Die Maße Recall und Precision basieren darauf, dass Dokumente entweder relevant oder nicht relevant sind. Studien lassen allerdings daran zweifeln, ob die binäre Bewertung tatsächlich ausreicht, einen solch komplexen kognitiven Prozess abzubilden. Binäre Beurteilungen scheinen problematisch zu sein, angesichts der vielen Ausprägungen und der Dynamik von Relevanz. Bei Maglaughlin und Sonnenwald (2002) findet man eine Auswahl von Studien, die verschiedene Bewertungsskalen, von 3-Punkt bis 11-Punkt oder auch Linien ohne Unterteilungen, erprobt haben. Bei der Messung des Grades der Relevanz sind User laut Saracevic (2007b) in der Lage mit verschiedenen Skalen umzugehen: "users are capable of using a variety of scales, from categorical to interval, to indicate their preferences." (S.2137). Jedoch gibt es nicht die eine richtige oder beste Skala.

3.2 Systemorientierte Sichtweise vs. Userorientierte Sichtweise

In den späten 40ern und frühen 50er Jahren des 20. Jahrhunderts entstanden die ersten computergestützten Information Retrieval Systeme. Aufgabe solcher Systeme war es, relevante Informationen zu einer Suchanfrage zu finden. Die Systeme basierten hauptsächlich auf Wahrscheinlichkeitsrechnung, Mathematik und Boolescher Algebra. Man nahm an, dass gefundene Dokumente relevant seien und nicht gefundene Dokumente nicht relevant. Man ging dabei von einem statischen Informationsbedürfnis des Nutzers aus. Doch schon die Pioniere der IR-Technik mussten sich eingestehen, dass nicht alle gefundenen Dokumente tatsächlich thematisch relevant oder nützlich waren. Bei der systemorientierten Sichtweise auf Relevanz wurde gemessen, wie gut die Anfrageterme mit der formalen Erfassung der Dokumente (Dokumentati-

onseinheit inklusive Klassifikation und Deskriptoren) übereinstimmten. Menschliches Suchverhalten wurde nicht berücksichtigt. Der so genannte systemorientierte Ansatz spielte zu dieser Zeit die dominante Rolle im IR-System-Design.

Aus den ersten umfangreicheren Evaluations- und Leistungs-Tests von Retrieval Systemen in den 50ern und 60er Jahren des 20. Jahrhunderts entstand eine rege Debatte darüber, wie man Relevanz feststellen solle und wer eine Beurteilung durchführen könne. In dieser Phase begannen erste Kritiker an dem Konstrukt der systembasierten Relevanz zu zweifeln. Schließlich ist es der Nutzer des Informationssystems, der aus einer gegebenen Menge von Elementen letztendlich entscheidet, was für ihn in seiner Situation relevant ist. Der Nutzer, seine Informationsbedürfnisse, seine Erwartungen und die kognitiven Prozesse während des Suchens sollten in den folgenden Jahren stärker berücksichtigt werden.

In den späten 80er Jahren bis heute hat vor allem die fortschreitende Entwicklung und Verbreitung der Informationstechnik (IT) dazu geführt, dass die Interaktion des Nutzers mit dem Informationssystem in den Mittelpunkt der Betrachtung gerückt ist. Qualitative Untersuchungsmethoden wurden von quantitativen Methoden abgelöst. Subjektive Wahrnehmungen und kognitive Erfahrungen während der Suche und während des Bewertens von Suchergebnissen standen den Systemeigenschaften und der Systemleistung im klassischen IR-Modell gegenüber. Die Entstehung des Informationsbedarfes, die Frage ob der Bedarf vollständig und korrekt durch die ausgewählten Suchbegriffe abgedeckt wird, der Kontext der Suchterme, kognitive und affektive Dimensionen spielten nun eine größere Rolle. In den letzten Jahren wurde der nutzerorientierte Ansatz auch beim Design von Informationssystemen stärker berücksichtigt, weil der systemorientierte Ansatz die Bedürfnisse der Nutzer nicht mehr ausreichend zufriedenstellen konnte.

Aus den zwei verschiedenen Sichtweisen auf IR-Systeme und Relevanz lassen sich zwei übergeordnete Klassen ableiten.

3.3 Zwei Klassen der Relevanz

Man unterscheidet heute zwischen objektiver und subjektiver Relevanz. Die **objektive Relevanz** ist dabei wie oben beschrieben im Bereich des (technischen) Information Retrievals anzusiedeln. Sie beschreibt wie sehr ein Dokument oder dessen Repräsentation in einer Datenbank mit der Suchanfrage (Terme, Operatoren) übereinstimmt, unabhängig vom Kontext. Mit ihr lässt sich die Leistung eines Retrieval Systems über technische und quantitative Methoden evaluieren. Darum bezeichnet man diese Art

der Relevanz auch als **systembasiert**. Die Auswahl relevanter Dokumente wird vom IR-System vorgenommen.

Davon zu unterscheiden ist die **subjektive Relevanz**. Sie wird auch kognitive Relevanz genannt, weil die kognitiven Strukturen des Users berücksichtigt werden. Die subjektive Relevanz beschreibt den Grad der Übereinstimmung von Informationen mit dem tatsächlichen Informationsbedürfnis des Suchenden. Im Gegensatz zur objektiven Relevanz lässt sich mit ihrer Hilfe eine Aussage darüber treffen, ob ein gefundenes Ergebnis zur Lösung eines spezifischen Problems beiträgt. Subjektive Relevanz wird auch als **Pertinenz** bezeichnet. Die Relevanzbeurteilung wird vom User durchgeführt, weswegen diese Art der Relevanz auch als **nutzerorientiert** bezeichnet wird. Der Suchende wird als essenzieller Bestandteil des IR-Systems gesehen.

3.4 Viele Arten von Relevanz

Die Einteilung in Relevanz und Pertinenz, wie oben beschrieben, ist angesichts des komplexen Konstrukts und der Multidimensionalität der Relevanz zu einfach. Saracevic (2007b) schreibt „At its base, relevance is dual, […], but from that dualism grows a pluralistic system of relevances“ (S. 1929). Da Relevanz eine Beziehung zwischen zwei Faktoren darstellt und im IR-Prozess viele verschiedene Faktoren zu berücksichtigen sind (z.B. auf menschlicher und technischer Ebene), gibt es auch mehrere Arten von Relevanz.

Eine Relevanz-Theorie, die speziell für die Informationssuche anwendbar wäre, gibt es (noch) nicht. Saracevic vergleicht die Suche nach einer allumfassenden Theorie gar mit der Suche nach dem Heiligen Gral: „Is a relevance theory that explains everything equivalent to a quest for the Holy Grail? “ (2007a, S. 1923). Statt einer guten Theorie hat die Informationswissenschaft aber Modelle, die sich dem Thema annähern.

Im folgenden Teil des Kapitels werden zwei zentrale Modelle beschrieben, die die Multidimensionalität der Relevanz veranschaulichen.

3.4.1 Relevanz-Modell von Mizzaro

Stefano Mizzaro (1996) fragte sich, wie viele Arten von Relevanz es gibt. Er arbeitete ein Modell, anhand dessen die vielfältigen wechselseitigen Beziehungen zwischen den beteiligten Objekten besser verstanden werden können.

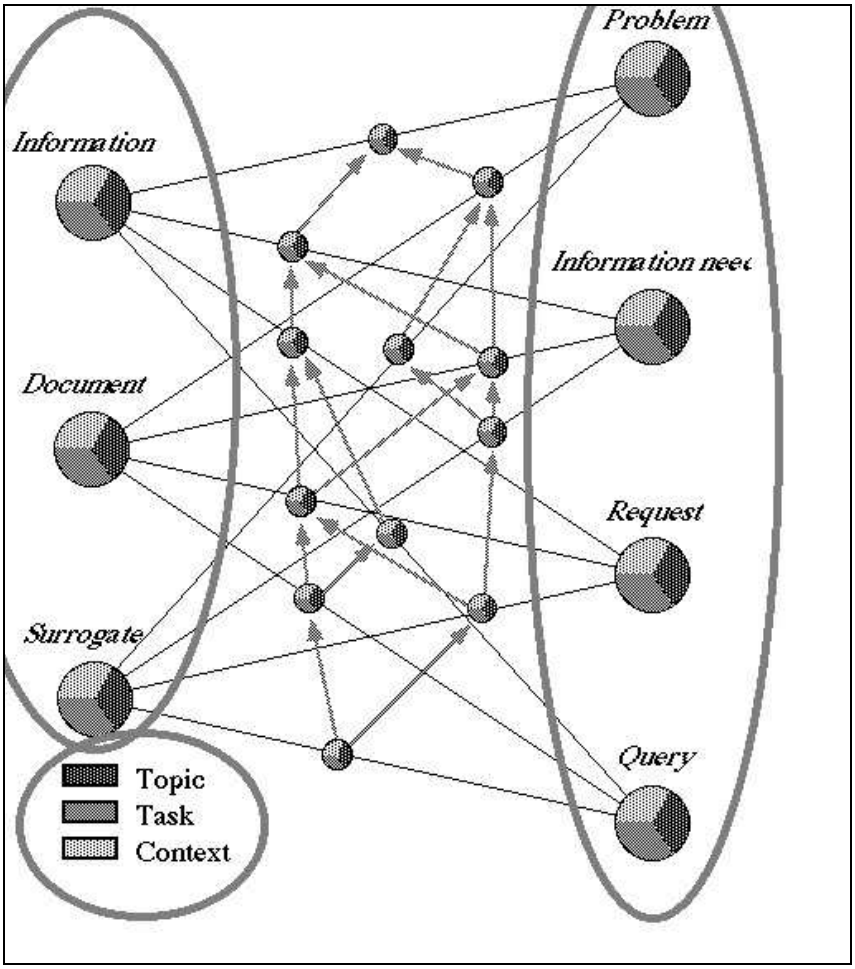


Abbildung 8: Viele Arten von Relevanz Quelle: In Anlehnung an (Mizzaro 1996)

Mizzaro unterscheidet vier Dimensionen. Drei davon sind in Abbildung 8 durch die Ellipsen dargestellt. Die erste Dimension ist die Quelle also das IR-System. Sie beinhaltet die Dokumentationseinheiten (“surrogate“), Dokumente („document“) und durch das Lesen entstandenes Wissen („information“).

Die zweite Dimension repräsentiert das Benutzerproblem. Sie besteht aus vier Elementen, nämlich der problematischen Situation („problem“), dem Informationsbedürfnis („information need“), der Anfrage in natürlicher Sprache („request“) und der in Retrievalsprache übersetzten Anfrage, der „Query“ (bestehend aus Suchtermen und Operatoren).

In der dritten Dimension in Mizzaros Modell wird das eigentliche Thema der Suche in drei Erscheinungsformen zerteilt, nämlich Topic, Aufgabe („task“) und Kontext. Diese Aspekte werden auch als „information subgoals“ bezeichnet. In einer vierten Dimension, der Zeit (die nicht in Abbildung 8 zu sehen ist) berücksichtigt Mizzaro die Tatsache, dass sich Relevanzbeziehungen im Laufe eines Suchprozesses verändern kön-

nen, was auf die Dynamik der Relevanz hinweist. Was zu einem bestimmten Zeitpunkt relevant erscheint, mag zu einem anderen Zeitpunkt nicht mehr relevant sein.

Das Modell verdeutlicht, dass genau spezifiziert werden muss, von welcher Art Relevanz gesprochen wird. Ob damit beispielsweise die Übereinstimmung von Dokument und Suchanfrage gemeint ist oder wie gut eine Information zu einem Informationsbedürfnis passt.

Ausgehend von Mizzaros Modell lässt sich Relevanz als Beziehung zwischen jeweils zwei Entitäten aus zwei verschiedenen Gruppen definieren. Relevanzbeziehungen bestehen zwischen einem Objekt der ersten Dimension und einem Objekt der zweiten Dimension. Da beide Objekte aber noch weiter zerlegt werden können, ergeben sich zusätzlich weitere mögliche Relevanzverbindungen.

3.4.2 Fünf Typen der Relevanz nach Saracevic

Saracevic (2007b) beschreibt, basierend auf seinen eigenen früheren Erkenntnissen und denen anderer Wissenschaftler fünf Ausprägungen von Relevanz, die Mizzaros Modell zum Teil ergänzen und gedanklich weiterführen:

1. Systemrelevanz oder algorithmische Relevanz: Sie beschreibt die Beziehung zwischen den Anfrage-Termen und den gefundenen Ergebnissen. Dies entspricht der Klasse der objektiven Relevanz. Konkret bedeutet dies, dass beispielsweise der eingegebene Suchbegriff im Titel des gefundenen Dokuments vorkommt. In Mizzaros Modell entspräche dies der Verbindung Query – Document.
2. Thematische Relevanz („topical/subject relevance“): Sie umfasst die Beziehung zwischen dem Thema, das in der Suchanfrage umgesetzt wird und dem Thema des gefundenen Dokuments. Das, was inhaltlich gesucht wird, das wird auch gefunden. Das Kriterium von dem die Thematische Relevanz abgeleitet werden kann, wird in der englischsprachigen Literatur auch als „aboutness“ bezeichnet.
3. Kognitive Relevanz oder Pertinenz: Sie erfasst die Beziehung zwischen dem kognitiven Wissensstand des Nutzers und der gefundenen Information.
4. Situative Relevanz: Sie drückt die Beziehung zwischen einer konkreten Situation oder einer Aufgabe und der Information aus. Sie beschreibt die Nützlichkeit („usefulness“) der Information. Will ich zum Beispiel wissen wann der Bus fährt, suche nach RMV und finde ein Dokument über die Geschichte des RMV ist dies thematisch zwar passend, aber nicht relevant in der unmittelbaren Situation.
5. Motivationale oder affektive Relevanz: Hier wird eine Verbindung zwischen Information und den Zielen bzw. der Motivation des Users hergestellt.

Abgesehen von der Systemrelevanz, sind die vier darauffolgenden Punkte schwer voneinander abzugrenzen und für den User während der Relevanzbewertung nicht bewusst zu unterscheiden.

Problematisch wird es, wenn Systemrelevanz gegeben ist, die Information aber nicht zum Informationsbedarf passt oder nicht genutzt werden kann. Auf der anderen Seite können Objekte, die nicht in der Query abgebildet werden, also keine algorithmische Relevanz aufweisen, auch nicht gefunden werden.

3.5 Untersuchung von Relevanz

In Kapitel 2.6 wurde bereits Eyetracking als eine mögliche Untersuchungsmethode für Relevanz bzw. Relevanzverhalten vorgestellt. Darüber hinaus gibt es noch eine Reihe weiterer Methoden wie Information Seeking Behavior erforscht werden kann, zum Beispiel:

- Interviews in mündlicher oder in schriftlicher Form als Fragebogen, in Gruppen (Focus groups) oder mit Einzelpersonen z. B. Experten, persönlich oder am Telefon,
- Beobachtungen,
- Video, Audio, Think Aloud (zum Dokumentieren von Sprache, non-verbalem Verhalten, kognitiver und affektiver Vorgänge während der Suche über Mimik und Gestik),
- Logfile Analysen,
- Nutzertagebücher,
- Laborexperimente,
- Feldstudien,
- Fallstudien und
- Inhaltsanalysen zum Beispiel im Anschluss an ein Interview.

Ausführlich beschriebene Beispiele findet man hierzu in Kapitel 9.1 bei Case 2007.

3.5.1 Relevanzkriterien

Wenn man Relevanz charakterisieren will, stellt man sich automatisch als Erstes die Frage, was Informationen überhaupt relevant macht. Worauf achten User, wenn sie Relevanzentscheidungen treffen? Viele Erkenntnisse hierzu lassen sich aus der For-

schung zu Relevanz-Kriterien herleiten. Die Faktoren verraten, was Menschen beachten, wenn sie die Entscheidung „brauch ich“ oder „brauch ich nicht“ treffen. Aus den Ergebnissen zahlreicher Studien konnte geschlussfolgert werden, dass es außer dem vorrangigen Merkmal der thematischen Übereinstimmung (topicality) eine Reihe von weiteren Faktoren gibt, die die Bewertung von Informationen durch User unterstützen und beeinflussen.

Betrachtet man nur die Klasse der objektiven Relevanz, dann findet lediglich ein Kriterium Anwendung, nämlich das der Übereinstimmung der Query-Terme mit den Termen des gefundenen Informationsobjekts (Systemrelevanz). Im Bereich der subjektiven Relevanz-Bewertung stehen dem Nutzer hingegen eine Vielzahl Parameter zur Verfügung.

Zentrale Veröffentlichungen, die in diesem Zusammenhang immer wieder zitiert werden, stammen von Linda Schamber, Carol Barry und Taemin Kim Park, die im folgenden Kapitel kurz vorgestellt werden. Ergänzt werden sie durch die Ergebnisse zweier neuerer Studien von Xu/Chen und Taylor/Zhang/Amadio.

Linda Schamber (1991) untersuchte Bewertungskriterien für Wetterinformationssysteme. Hierfür wurden Personen gebeten, berufliche Szenarien zu beschreiben, in denen Wetterinformationen benötigt wurden und die Quellen zu nennen, aus denen sie die Informationen zogen. Durch die Angaben der Testpersonen konnte Schamber zehn Kategorien von Kriterien bestimmen.

Barry (1994) extrahierte 23 Relevanzkategorien, die sie in sieben Oberklassen aufteilte. In ihrer Studie sollten die Teilnehmer Informationen in Datensätzen einkreisen, die sie dazu bewegen würden, auch den Volltext zu lesen. Außerdem sollten sie jene Informationen durchstreichen, die sie davon abhalten würden den Volltext zu lesen. Anschließend wurden die Personen über ihre Beweggründe befragt.

In Abbildung 9 stehen sich die beiden Kriteriensets gegenüber. Die Tatsache, dass „geografische Nähe“ in Schambers Studie als sehr wichtiges Kriterium genannt wurde zeigt, dass Kriteriensets selbst zunächst nicht allgemeingültig sind, sondern ebenfalls durch eine Vielzahl von Faktoren wie dem beruflichen Umfeld beeinflusst werden.

Table 1. Frequency of criterion category mentions

The Barry study: 448 mentions of criterion categories by 18 respondents		The Schamber study: 811 mentions of criterion categories by 30 respondents	
Category	Number of Mentions	Category	Number of Mentions
Depth/Scope	64	Presentation Quality	115
Accuracy/Validity (Obj, Subj.)	60	Currency	114
Content Novelty	53	Reliability	107
Tangibility	29	Verifiability	103
Affectiveness	25	Geographic Proximity	96
Recency	25	Specificity	84
Availability Environment	21	Dynamism	63
Consensus	20	Accessibility	52
External Verification	19	Accuracy	43
Background/Experience	19	Clarity	34
Source Reputation	18		
Effectiveness	16		
Access (Obtain., Cost)	14		
Source Quality	14		
Source Novelty	10		
Clarity	9		
Ability to Understand	9		
Relationship with Author	7		
Time Constraints	6		
Personal Availability	5		
Document Novelty	5		

Abbildung 9: Relevanzkriterien von Schamber '91 und Barry '94
Quelle: (Barry/Schamber 1998, S. 223)

Barry und Schamber verglichen 1998 ihre zwei Kriteriensets, die aus den beiden oben genannten Studien gewonnen wurden. Sie konnten Überlappungen finden und klassifizierten diese in zehn Gruppen (Abb. 10). Die Tatsache, dass es so viele Überlappungen zwischen den Kriterien zweier verschiedener Gruppen mit unterschiedlichen Hintergründen gab, führte zur Schlussfolgerung, dass es eine bestimmte Menge von Relevanzkriterien geben müsse, die über Benutzertypen, situative Kontexte und Informationsquellen hinaus gelten müsse.

Wie in Abbildung 10 zu sehen ist, wurden die 10 Kriterien-Gruppen zum Teil noch einmal bestätigt durch eine weitere Studie von Schamber und Bateman 1996.

TABLE 2. Overlapping relevance categories identified in studies comparing relevance criteria literature.

Barry and Schamber (1995, 1998)	Schamber and Bateman (1996)
Depth scope/specificity	
Accuracy/validity	
Clarity	Clarity
Currency	Currency
Tangibility	
Quality of source	Credibility
Accessibility	
Availability of information	Availability
Verification	
Affectiveness	
Topical appropriateness[a]	Aboutness

[a] Category assumed but not studied directly.

Abbildung 10: Überlappungen von Relevanz-Kriterien-Sets
Quelle: (Maglaughlin/Sonnenwald 2002, S. 330)

Xu und Chen veröffentlichten 2006 die Ergebnisse ihrer Studie zum Thema Relevanzkriterien. Ziel war es, fünf Faktoren zu bestimmen und diese auf ihre Wichtigkeit hin zu untersuchen. Abgesehen von thematischer Relevanz („topicality“) gab es bis zu diesem Zeitpunkt keine Einigkeit unter den Forschern über ein gemeinsames Kern-Set von Kriterien. Zwar erwähnen Xu und Chen in ihrer Studie auch Barry und Schamber, sie kritisieren aber, dass es keinen Konsens über ein klar abgegrenztes Set gäbe. Sie bemängelten an vorangegangen Studien die meist enorm hohe Anzahl an Merkmalen sowie eine uneinheitliche Terminologie. Viele inhaltlich gleiche Faktoren hätten bei verschiedenen Autoren unterschiedliche Bezeichnungen.

Die fünf Kriterien, die Xu und Chen untersuchen wollten, leiteten sie aus Grics's Maximen der menschlichen Kommunikation her. Die Kriterien waren:

1. Abdeckung (Breite und Tiefe der Behandlung),
2. Neuheit (die Information ist dem Nutzer nicht bekannt),
3. Verlässlichkeit/Validität (der Inhalt der Informationen ist glaubhaft und richtig),
4. Inhaltliche Relevanz (der Inhalt bedient ein aktuelles Informationsbedürfnis) und

5. Verständlichkeit (der Nutzer kann den Text verstehen, es gibt keine übertriebene Fachterminologie, oder Grafiken ohne Erklärung etc.).

In der Studie fanden Xu und Chen heraus, dass **inhaltliche Relevanz** und **Neuheit** die beiden Haupt-Kriterien zur Beurteilung der Relevanz waren. Verständlichkeit und Verlässlichkeit waren zwar auch von Bedeutung, aber nicht so vordergründig wie die anderen beiden Faktoren.

Taylor, Zhang, Amadio 2009 untersuchen Kriterien im Zusammenhang mit der Suchphase. Verständlichkeit, Klarheit, Tiefe, Genauigkeit, Spezifität sind die bevorzugten Kriterien der User über alle Suchphasen hinweg. Je weiter die Suchenden in ihrer Suche fortschreiten, desto höher wurden die Kriterien gewichtet.

3.5.2 Faktoren, die die Relevanz-Bewertung beeinflussen

Im vorangegangenen Teil standen die Dokument-Eigenschaften im Mittelpunkt, die Relevanz bestimmen, sowohl positiv als auch negativ. Im folgenden Abschnitt werden zwei Kriteriensets gezeigt, die Faktoren enthalten, die den User in seiner Bewertung beeinflussen und nicht direkt dem Informationsobjekt zugehörig sind, sondern zum Beispiel dem Nutzer, dem Informationssystem oder verschiedenen Kontexten.

Die erste Sammlung stammt von Linda Schamber (1994) und umfasst 80 Faktoren, von denen einige auch in die obere Kategorie einsortiert werden können. Neu hinzu kommen aber Einflüsse, die den Nutzer betreffen, wie Voreingenommenheit, geistige Fähigkeiten, Ausbildung und Forschungserfahrung. Zugangsmöglichkeiten zum Informationssystem, Kosten und Usability sind ebenfalls Faktoren, die davor kaum berücksichtigt wurden (Abb. 11).

TABLE 1. Selected factors affecting relevance judgments, as determined in experimental studies.*

Judges	Information Systems
Biases	Access
Cognitive style	Browsability
Formal education	Comprehensiveness
Intelligence	Cost savings
Knowledge/experience	Ease of detection of relevance
Research stage	Effort expended
Requests	Precision of subject output
Difficulty level	Judgment conditions
Subject matter	Breadth of document set
Textual attributes	Definition of relevance
Documents	Order of presentation
Aboutness	Size of document set
Accuracy (truth)	Time for judging
Aesthetic value	Choice of scale
Difficulty level	Ease of use
Importance	Number of rating categories
Novelty	Type of scale
Recency	
Style	

* Factors are selected from 80 factors reported in Table 1 of Schamber, 1994, p. 11.

Abbildung 11: Zusammenfassung der Faktoren, die die Relevanzbewertungen beeinflussen nach Schamber 1994 Quelle: (Harter 1996, S.39)

Park (1993) stellt drei stark handlungsorientierte, nutzerbezogene Kategorien vor, die bei einer Relevanzbeurteilung Einfluss nehmen. Sie stammen, wie die Kriterien von oben, ebenfalls aus einer empirischen Studie. Die Daten wurden hauptsächlich durch Interviews gewonnen. Folgende Kategorien, die die Interpretation bibliografischer Daten (Titel, Autor, Journal, Dokumenttyp, Abstract) beeinflussen, hat Park herausgearbeitet:

1. Interner Kontext aus der Erfahrung des Nutzers: dazu gehören zum Beispiel Fachkenntnisse im Wissensgebiet, Kenntnis der Literatur und relevanten Autoren, Forschungserfahrung, Bildungshintergrund usw.
2. Externer Kontext der Informationssuche: z. B. Eindruck von Informationsqualität, Ziel der Suche, Verfügbarkeit, Prioritäten in Bezug auf das Informationsbedürfnis, in welcher Phase der Forschung sich der User befindet, usw.

3. Problemkontext: Er charakterisiert das Informationsproblem des Users genauer. Dazu gehört zum Beispiel wie umfassend der Nutzer das Problem erfasst und verstanden hat oder ob eine gefundene Zitation dazu beiträgt das Problem besser zu verstehen.

Zusammenfassung

Saracevic (2007b) rezensierte 16 Studien aus den Jahren 1991 bis 2005 in denen Kriterien herausgearbeitet und klassifiziert wurden. Er hat folgende Gemeinsamkeiten zusammenfasst: Verschiedene User benutzen ähnliche Kriterien, vergeben aber unterschiedliche Gewichtungen. Die Kriterien stehen nicht gesondert und unabhängig voneinander, zwischen den Gruppen findet Interaktion statt. Inhaltsorientierte Kriterien werden am wichtigsten angesehen. Die Wichtigkeit verändert sich mit der Dokumentrepräsentation, d. h. in Dokumentationseinheiten (Datensätze mit bibliografischen Angaben) und Volltexten können dieselben Kriterien unterschiedliche Gewichtungen haben. Ein klares Muster, welche Art der Repräsentation relevanter bewertet wird, gibt es nicht. Von den bibliografischen Angaben produzieren Titel und Abstract die meisten Kriterien.

Relevanzfaktoren können laut Saracevic in folgende Gruppen eingeteilt werden:

- „Information characteristics“:
  - Inhalt (Thema, Qualität, Tiefe, Aktualität, Klarheit),
  - Objekt-Eigenschaften (Typ, Format, Repräsentation, Verfügbarkeit, Kosten),
  - Validität (Richtigkeit der Information, Vertrauenswürdigkeit der Quelle, Nachprüfbarkeit) und
- „Human characteristics“:
  - Nutzen in der jeweiligen Situation (Angemessenheit bezogen auf das Problem oder die Aufgabe, Nützlichkeit, Wert),
  - Kognitive Übereinstimmung (Verständlichkeit, Neuheit),
  - Affektive Übereinstimmung (emotionale Reaktion auf die Information, Spaß, Frust, Unsicherheit),
  - Glaubwürdigkeitsgrad im Ermessen des Nutzers (Überzeugung, Vertrauen)

Es wird deutlich, dass es zur Relevanzbeurteilung viele mögliche Wege gibt, die individuell verschieden ausfallen können und somit keine einheitliche Definition erlauben.

4 Eyetracking-Experiment

Für diese Arbeit wurden im Rahmen eines kontrollierten Eyetracking-Experiments die Relevanzentscheidungen von zwölf Testpersonen untersucht. An vier Terminen wurden die Studierenden der Informationswissenschaft an der Hochschule Darmstadt bei der binären Relevanzentscheidung über eine Eyetracking-Anlage beobachtet. Die gewonnen Daten wurden im Anschluss zum Teil über die zugehörige Software und zum Teil durch Videoexporte am heimischen PC ausgewertet. Untersucht wurde dabei vor allem, wann die Entscheidungen fallen, wie viel Zeit bis zu einer Entscheidung benötigt wird, auf Basis welcher Informationen sie getroffen wird und ob es Hinweise darauf gibt, dass anhand von charakteristischen Verhaltensweisen Relevanz abgeleitet werden kann.

4.1 Hardware und Software

Bei der verwendeten Hardware handelt es sich um das System T60 von der Firma Tobii. Bei dem freistehenden Eyetracker sind die Sensoren, die die Augenbewegungen aufzeichnen, unauffällig in eine schmale Leiste an der Unterseite eines Monitors eingefasst (Abb. 12). Die Testperson wird dadurch nicht von aufwendigen Apparaturen abgelenkt und kann sich frei bewegen.

Abbildung 12: Arbeitsplatz mit Tobii T60 Eyetracker Quelle: Eigene Darstellung

Mithilfe der dazugehörigen Software „Tobii Studio“ wurde das Versuchsdesign realisiert. Der Test bestand aus einer Einführungsseite und dem sogenannten Webelement, welches die Bewertungsoberfläche im Internet Explorer aufrief. Die Software ermöglichte außerdem die statistische Analyse und visuelle Darstellung der aufgezeichneten Blickbewegungsdaten zum Beispiel in Form von Scanpfaden und Heatmaps.

4.2 Testpersonen

Die zwölf Teilnehmer der Studie waren alle Studenten der Informationswissenschaft an der Hochschule Darmstadt und hatten daher Erfahrung im Umgang mit Suchergebnislisten und Relevanzentscheidungen. Keiner von ihnen hatte spezielles Vorwissen im Wissensgebiet des Topics, nämlich Medienpädagogik und Erziehungswissenschaft. Acht Personen studierten bereits im Masterstudiengang, drei befanden sich im vierten Fachsemester, ein Student im zweiten Fachsemester des Bachelorstudiengangs. Die Master-Studierenden waren alle Teilnehmer des Seminars „Information Seeking Behavior“ (Sommersemester 2011). Die Geschlechterverteilung war ausgeglichen mit sechs weiblichen und sechs männlichen Teilnehmern. Das Alter von elf Teilnehmern bewegte sich zwischen 22 und 28, ein Nutzer war 42 Jahre alt.

Tabelle 1: Die Testpersonen

	Studienbereich	Fachsemester	Geschlecht	Alter
Nutzer 1	Master	2	weiblich	28
Nutzer 2	Master	2	weiblich	22
Nutzer 3	Master	2	weiblich	23
Nutzer 4	Master	3	weiblich	25
Nutzer 5	Master	2	weiblich	23
Nutzer 6	Master	2	männlich	26
Nutzer 7	Master	2	männlich	25
Nutzer 8	Master	2	männlich	27
Nutzer 9	Bachelor	4	männlich	25
Nutzer 10	Bachelor	4	weiblich	24
Nutzer 11	Bachelor	2	männlich	42
Nutzer 12	Bachelor	4	männlich	22

Ein Experte führte darüber hinaus ebenfalls eine Bewertung ohne Eyetracking durch. Seine Ergebnisse wurden als Benchmark für die Relevanz verwendet.

4.3 Testumgebung

Als Relevanzbewertungs-Tool diente eine Webseite, die ursprünglich im Rahmen des DFG-geförderten Projektes „Value-Added Services for Information Retrieval“ des Leibnitz Institutes für Sozialwissenschaften (gesis) realisiert wurde (vgl. Mayr et al. 2011).

Die Bestandteile der Bewertungsoberfläche sind in Abbildung 15 dargestellt: auf der Webseite kann über ein Dropdown-Menü ein Topic ausgewählt werden (1). Topic und Aufgabenstellung können jederzeit nachgelesen werden (2). Nach Auswahl des Themas werden dem Tester eine Reihe von Dokumentrepräsentationen angezeigt, die aus Autor(en), Publikationsjahr, Titel, Abstract und Deskriptoren bestehen (3). Der Bewertende hat die Möglichkeit eine binäre Relevanzentscheidung zu treffen und dies via Klick auf einen Radiobutton (relevant / nicht relevant) auszuführen (4).

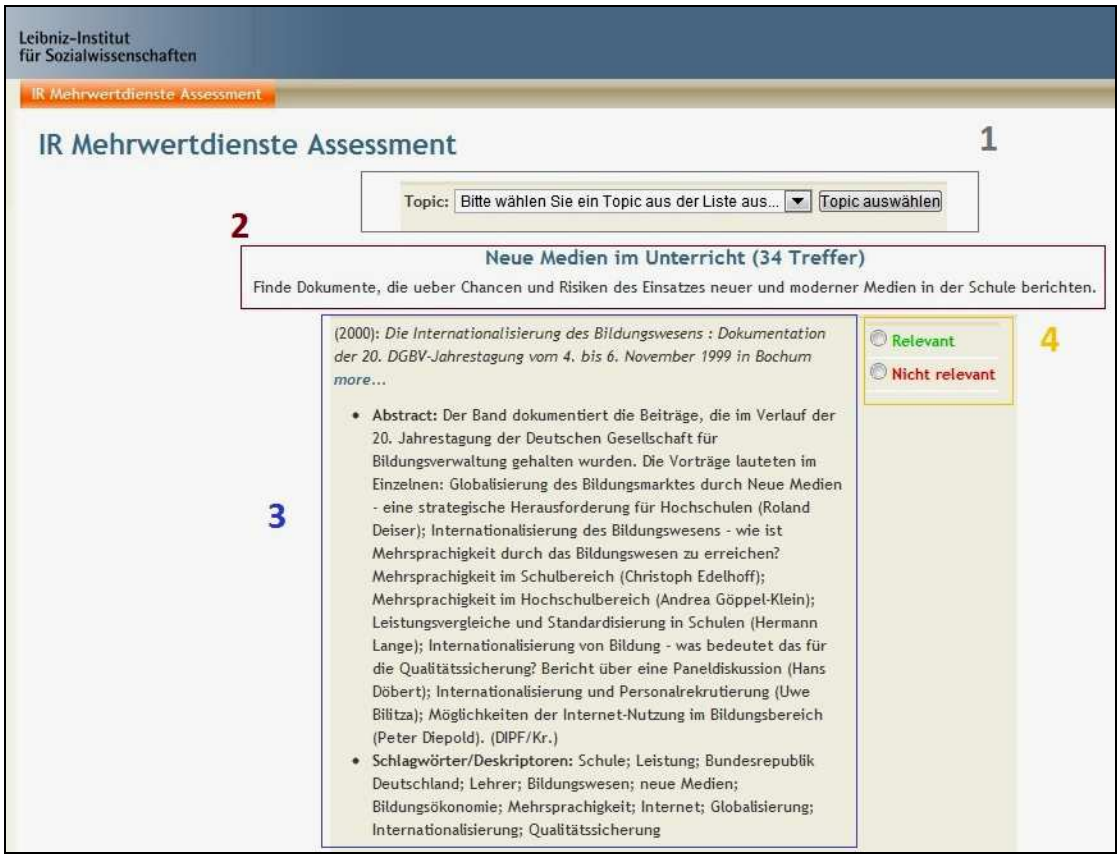


Abbildung 13: Screenshot der Testumgebung Quelle: Eigene Darstellung

Die Webseite wurde für das Experiment verwendet, weil sie eine optimale Bewertungsoberfläche bot, keine zusätzlichen Programmierarbeiten nötig waren und weil sie sofort zur Verfügung stand.

4.4 Testaufgabe

Als Thema für das Experiment wurde „Neue Medien im Unterricht“ gewählt. Die Arbeitsaufgabe lautete: „Finde Dokumente, die über Chancen und Risiken des Einsatzes neuer und moderner Medien in der Schule berichten“. Das Thema konnte über ein Dropdown-Menü direkt auf der Test-Webseite ausgewählt werden. Die Testpersonen mussten keine Suchanfrage formulieren. Es wurden darüber hinaus auch keine weiteren Vorgaben gegeben, welche Kriterien zur Bewertung herangezogen werden sollten. Alle Tester konnten den Begriff „relevant“ frei interpretieren und nach eigenen Kriterien bewerten.

Zusammen mit dem Topic und der Arbeitsanweisung beinhaltete die Einführungsseite des Tests noch zwei Bearbeitungshinweise. So sollten die Testpersonen die Bewertung der Reihe nach durchführen, wie die Dokumente auf dem Bildschirm erscheinen. Dies hatte den Zweck Scrolling so gering wie möglich zu halten, um die Auswertung nicht zu erschweren. Dadurch, dass jeder Mausklick eine Bewertungsenscheidung darstellt, bildet die Zeit von einem linken Mausklick zum nächsten eine Bewertunssequenz für ein Dokument. Beim zweiten Hinweis wurde verdeutlicht, dass die angezeigten Dokumente nicht nach Relevanz geordnet sind, sondern willkürlich gemischt angezeigt werden. Da Nutzer in ihrer Entscheidung sehr stark vom Rang eines Dokumentes in einer geordneten Ergebnisliste beeinflusst werden („trust bias“, siehe Seite 21), wurden jeder Testperson die Dokumente in unterschiedlicher Reihenfolge angezeigt. Der Rang lieferte somit keine Hinweise auf die Relevanz eines Dokumentes.

4.5 Testablauf

Jeder Testdurchlauf wurde nach demselben Schema durchgeführt. Es untergliedert sich in folgende Schritte:

1. Begrüßung und Vorgespräch

In einem kurzen Briefing wurden die Testpersonen über die Studie, die Eyetracking-Anlage und die kommenden Schritte aufgeklärt. Außerdem fand eine Einweisung in die Bedienung der Bewertungswebseite statt. Dies bezog sich vor allem auf das au-

tomatische Ausklappen des Abstracts des nachfolgenden Dokuments, sobald eine Bewertung per Klick auf den Radiobutton stattgefunden hat.

2. Einrichten des Monitors

Als Nächstes musste die Eyetracking Anlage auf die jeweiligen Tester eingestellt werden. Der Monitor wurde dazu horizontal und vertikal so ausgerichtet, dass sich zwei Referenzpunkte am Bildschirm möglichst in der Mitte eines Quadrates befinden. Anschließend fand eine Kalibrierung statt, bei der die Testperson mit ihren Augen einen Punkt auf dem Bildschirm verfolgen mussten. Wenn alle Parameter im grünen Bereich waren, konnte die Aufnahme gestartet werden. Gegebenenfalls musste der Stuhl des Testers noch angepasst werden. Die Körperhaltung während der Kalibrierung sollte sich nicht von der Haltung während der Bearbeitung der Aufgabe unterscheiden.

3. Durchführung der Bewertung

Nach Aufruf des Instruction-Elementes mit der Aufgabenstellung und Bearbeitungshinweisen konnten die Tester die Bewertungswebseite durch Drücken einer beliebigen Taste öffnen lassen und die Bewertung nach Eingabe der Login-Daten starten. Jeder Nutzer bewertete die selben 34 Dokumente in unterschiedlichen Reihenfolgen.

4. Feedback-Gespräch

Nachdem die Bewertung abgeschlossen und die Aufnahme beendet war, fand ein kurzes informelles Feedbackgespräch statt, bei dem die Tester ihre Eindrücke schildern konnten. Darüber hinaus vergaben die Tester an dieser Stelle Punkte für die Wichtigkeit der Datenelemente für den aktuellen Test. Es konnten Werte von 0 für „gar nicht wichtig“ bis 10 für „sehr wichtig“ vergeben werden.

Je nach Interesse bekamen die Tester außerdem eine Einführung in die Funktionalitäten der Eyetracking-Software.

4.6 Auswertung

Die Auswertung der gewonnen Daten fand zum Teil am heimischen PC und zum Teil im Uselab an der Hochschule statt. Da für detaillierte Statistiken (Auswertung der Fixations-Zahlen) die Software Tobii Studio benötigt wurde, mussten für diese Analysen Termine an der Eyetracking-Anlage vereinbart werden.

Durch den Export der aufgezeichneten Bewertungs-Videos war es möglich, eine Zahl von Analysen ohne Tobii Studio zu machen. Dazu gehörte:

- Dauer der Bewertung jedes Dokuments von jedem User inklusive Zeitstempel und Listenposition, um die Bewertungen später bei Bedarf schneller wiederzufinden,

- die Relevanzentscheidung (relevant / nicht relevant) und
- die Beschreibung des Blickverlaufs für jede Bewertungssequenz und Darstellung der Sequenzen in verkürzter Schreibweise (siehe Anlage A1 und A2).

Detailliertere statistische Auswertungen konnten nur im Uselab erfolgen. Diese waren sehr aufwendig und konnten nur für 17 von 34 Dokumenten durchgeführt werden. Folgende Handlungsschritte konnten über die Software ausgeführt werden:

- Definition von jeweils drei Lookzones für jede Bewertungssequenz (Abb. 14):
  T (Autoren/Publikationsjahr/Titel),
  A (Abstract),
  D (Deskriptoren),
- Bestimmen der Anzahl der Fixationen für jede AOI
- Visualisierung des Scanpfades für jede Szene

Die Lookzone „Titel“ wurde nicht mehr in Autoren, Publikationsjahr und Titel unterglie-dert, weil diese Angaben sehr nah zusammenstehen und durch die Ungenauigkeiten beim Eyetracking nicht immer so scharf voneinander abgegrenzt werden konnten.

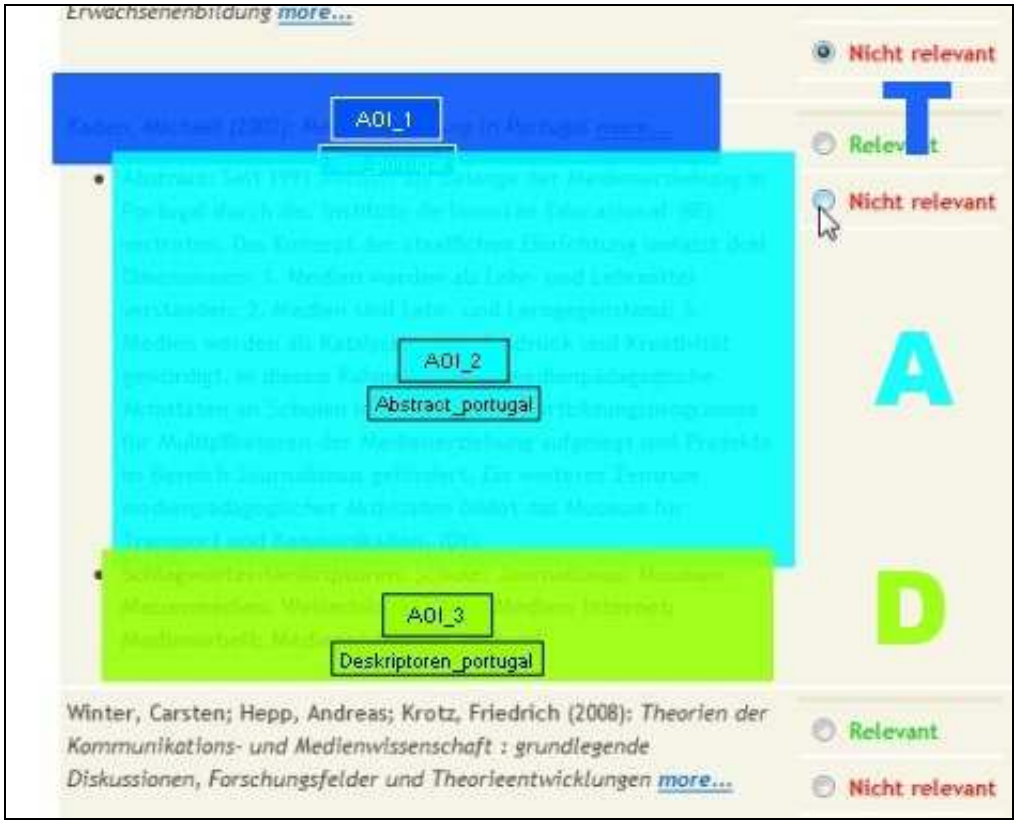


Abbildung 14: Definition der AOIs Quelle: Eigene Darstellung

5 Ergebnisse

In diesem Kapitel werden die Ergebnisse der Eyetracking-Studie dargestellt. Die Unterkapitel sind so aufgebaut, dass zunächst Thesen und Fragen formuliert werden, die dann anhand der gewonnenen Daten bestätigt oder verneint werden.

5.1 Bewertungsdauer

Wie lange benötigen die Nutzer, um Dokumente zu bewerten?

Das erste Datum, welches einfach zu messen war, ist die Länge der Bewertungssequenzen, wobei man die Länge der Gesamtbewertung pro Nutzer und die Längen der einzelnen Bewertungssequenzen voneinander unterscheiden kann. Da jeder linke Mausklick eine Bewertungsentscheidung darstellt, bildet die Zeit von einem linken Mausklick zum nächsten eine Bewertungssequenz. Die Zeit von der ersten Fixation im ersten Dokument bis zur letzten Relevanzentscheidung bildet die Gesamtlänge.

Die Spanne der Gesamt-Bewertungslängen variiert stark und reicht vom schnellsten Benutzer mit 10,22 Min. bis zum langsamsten Benutzer mit 23,50 Min. Durchschnittliche Bewertungsdauern umfassen von 18,03 Sekunden bis zu 41,47 Sekunden.

Die schnellste Bewertung eines Dokumentes dauerte nur drei Sekunden. Sie wurde von Nutzer 5 durchgeführt. In diesem Fall führte der Titel: „'Neue Medien' in die Schulen?: Probleme und Risiken der Medienpädagogik an der Schwelle zum 'Informationszeitalter'“ durch die enthaltenen Stichwörter „Neue Medien“ und „Schulen“ zu der sehr schnellen Entscheidung das Dokument relevant zu bewerten. Nutzer 5 hat nach dem Titel keine weiteren Datenelemente betrachtet.

Die längste Bewertung dauerte 98 Sekunden und wurde von Nutzer 3 ausgeführt. Zwar befand sich auch hier das Stichwort „Neue Medien“ im Titel („Kids und neue Medien: Netz- oder Pixelgesellschaft?“), der Nutzer war sich aber scheinbar unsicher und musste das Abstract zwei Mal komplett lesen, bevor eine Entscheidung getroffen werden konnte. Auch dieses Dokument wurde relevant bewertet.

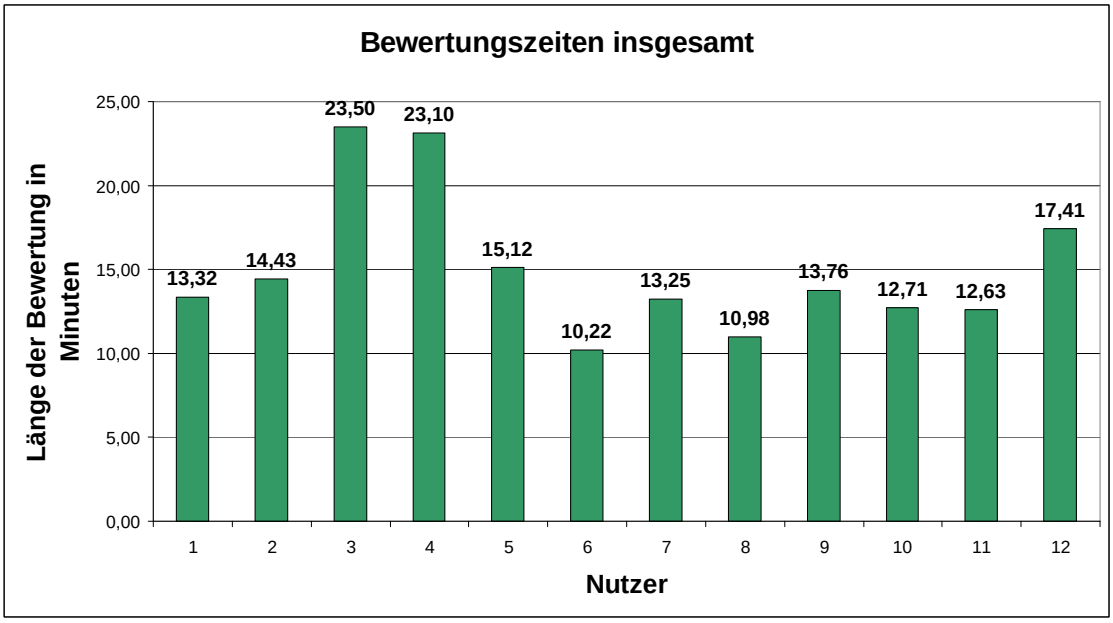


Abbildung 15: Bewertungszeiten

*Besteht ein Zusammenhang zwischen der **Gesamtlänge** der Bewertung eines Nutzers und der Relevanzbewertung?*

Es ist auffällig, dass die schnelleren, entscheidungsfreudigeren Bewerter eher mehr Dokumente nicht-relevant als relevant bewertet haben. Bei den langsameren Bewertern überwiegt hingegen die Zahl der als relevant beurteilten Dokumente.

In Abbildung 16 wird die Gesamtlänge der Bewertung für einen Nutzer und die Anzahl seiner relevanten und nicht-relevanten Urteile dargestellt. Die Nutzer sind zur besseren Übersicht so sortiert, dass der schnellste Bewerter ganz links und der langsamste Bewerter ganz rechts steht. Es ist deutlich zu erkennen, dass die nicht-relevanten Bewertungen auf der linken Seite überwiegen, wohingegen auf der rechten Seite mehr relevante Urteile gefällt wurden. Die Gesamtlänge der Bewertung ist also ein Hinweiskriterium dafür, ob die Bewerter mehr Dokumente relevant oder nicht-relevant bewerten.

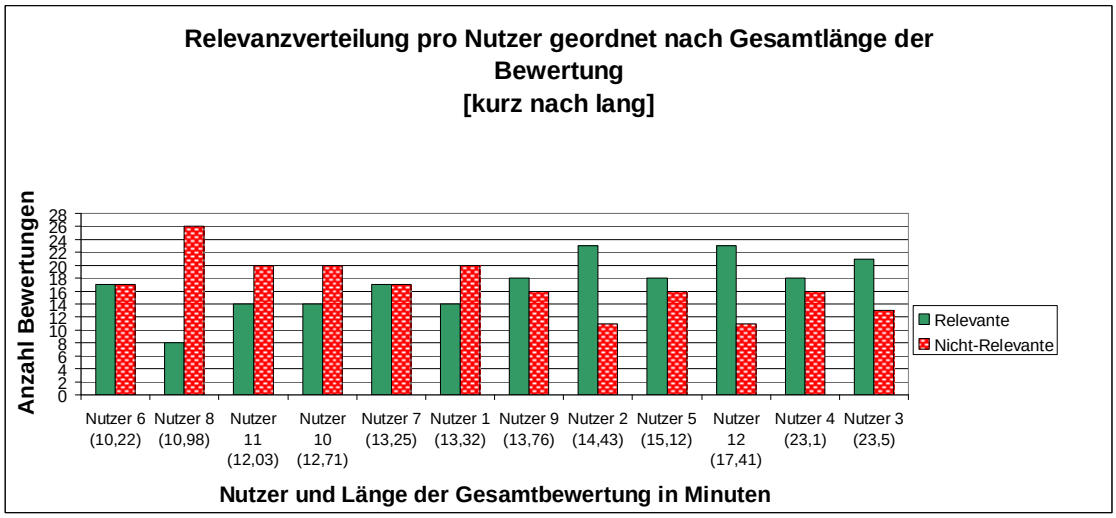


Abbildung 16: Relevanzverteilung nach Gesamtlänge der Bewertungen

*Besteht ein Zusammenhang zwischen Länge einer **Einzelbewertung** und der Relevanzbewertung? Werden relevante Dokumente schneller identifiziert als nicht-relevante?*

Wie in Tabelle 2 zu sehen ist, brauchten die Tester in fünf Fällen für jene Dokumente länger, die sie relevant bewerteten. In sieben Fällen dauerte die Bewertung der nicht-relevanten Dokumente länger, wobei der Unterschied in einem dieser Fälle nur 0,91 Sek. betrug, also nicht bedeutend gewichtet werden kann. So ist das Verhältnis trotzdem ausgeglichen. Es besteht demnach kein Zusammenhang zwischen durchschnittlicher Lesedauer und Relevanzbewertung eines Users.

Zwar tendieren die „Gesamt-Schnellbewerter“ eher dazu nicht-relevant anzuklicken, schnelle Entscheidungen resultieren aber nicht in jedem Fall in nicht-relevanten Bewertungen. In Abbildung 17 sind die Relevanzverteilungen dargestellt, aufgeteilt nach Lesedauer. Die Dauer ist zur besseren Darstellung in Intervalle von jeweils zehn Sekunden Dauer aufgeteilt. Wie man sieht, liefert die Dauer der Einzelbewertung keinen direkten Hinweis darauf, wie das Dokument am Ende beurteilt wird. Die Verteilung von relevanten und nicht-relevanten Dokumenten in jedem Intervall ist relativ ausgeglichen, d. h. relevante Dokumente werden nicht länger gelesen als nicht-relevante. Eine Ausnahme bilden die beiden Dokumente, die jeweils am längsten gelesen wurden. Sie wurden beide als relevant eingestuft.

Tabelle 2: Länge der Bewertungen

	Durchschnitt relevante Bewertungen [Sek.]	Durchschnitt nicht-relevante Bewertungen [Sek.]
Nutzer 1	**25,64**	22,00
Nutzer 2	20,87	**29,64**
Nutzer 3	**44,19**	37,08
Nutzer 4	**41,78**	38,50
Nutzer 5	22,61	**31,25**
Nutzer 6	15,35	**20,71**
Nutzer 7	21,24	**25,53**
Nutzer 8	**20,75**	18,96
Nutzer 9	23,28	**25,44**
Nutzer 10	19,43	**24,55**
Nutzer 11	22,00	**22,91**
Nutzer 12	**32,29**	29,65

Ein Beispiel für die enorme Spanne in den Einzel-Lesedauern ist das Dokument „Wissenserwerb durch 'interaktive' neue Medien: aus Sicht der Erziehungswissenschaft“, welches alle als relevant eingestuft haben. Die User benötigten zwischen 8 und 68 Sekunden um eine Entscheidung zu treffen.

Die Erkenntnis, dass die Lesedauer eines Dokumentes keine Rückschlüsse auf die Relevanz erlaubt, erkannten auch Kelly und Belkin in einer Studie 2001. Die Zeit, die für das Lesen der relevanten Dokumente sowie für das Lesen der nicht-relevanten Dokumente verwendet wurde, war jeweils ähnlich und die Unterschiede wurden als nicht signifikant eingestuft.

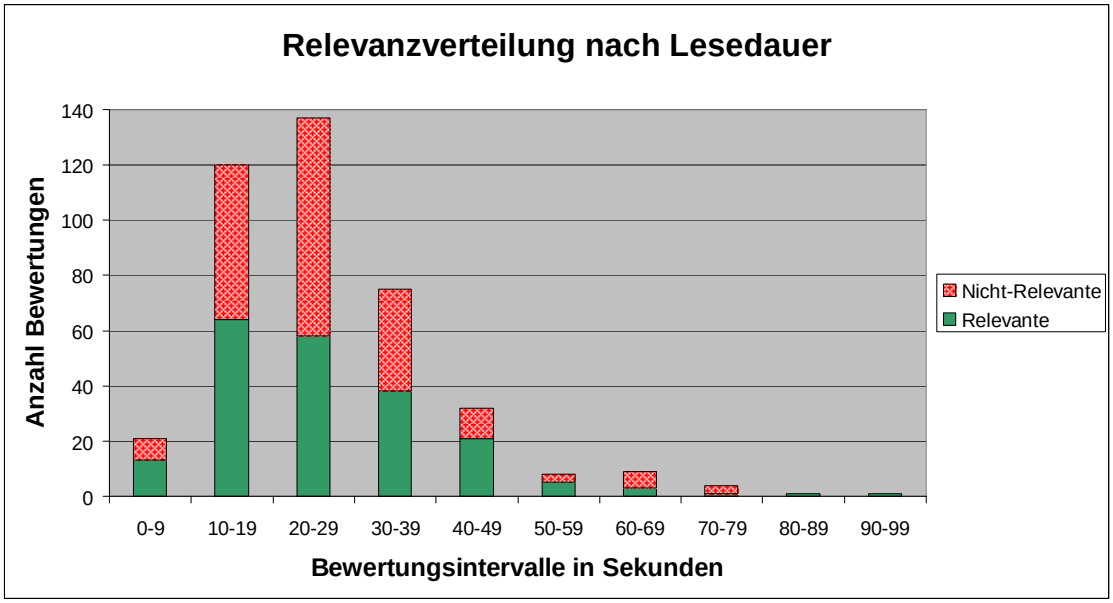


Abbildung 17: Relevanzverteilung nach Lesedauer in Intervallen

5.2 Lesegeschwindigkeit

Werden die Längen der Einzelbewertungssequenzen zum Ende hin kürzer?

Durch Visualisierung der Bewertungszeiten eines Users in einem Balken-Diagramm kann durch Einfügen einer Trendlinie eine Tendenz in der Lesegeschwindigkeit ermittelt werden. In acht Fällen wurden die Tester zum Ende hin schneller, d. h. die Bewertungsdauer für jedes einzelne Dokument nimmt in der Tendenz ab. Da die Dokumente jedes Mal neu gemischt wurden, konnte gewährleistet werden, dass sich die längeren Abstracts nicht immer an derselben Stelle befinden. Dass die Bewertungszeiten in acht Fällen dennoch schneller werden, spricht für ein Nachlassen der Konzentration oder Motivation je länger die Bewertung dauert. Die Ergebnisliste war mit 34 Dokumenten jedoch auch sehr lang. Auf SERPs werden in der Regel nur 3-5 Abstracts betrachtet (vgl. „Grundlagen“ S. 21).

5.3 Übereinstimmung in den Relevanzbewertungen

Angesichts der Tatsache, dass die Studienteilnehmer eine recht homogene Gruppe bildeten (Altersklasse, Studiengang, Herkunft, Vorwissen), könnte man annehmen, dass die Übereinstimmung in den Relevanzbewertungen vergleichsweise hoch sein müsste.

Wie viele Dokumente werden von allen gleich bewertet?

Nur vier von 34 Dokumenten wurden von allen Usern übereinstimmend bewertet, d. h. bei 30 Dokumenten waren sich die Nutzer nicht einig und bewerteten zu unterschiedlichen Anzahlen relevant und nicht relevant. Das entspricht einem Interrater Agreement von 11,76 %. Nimmt man die Expertenbewertung dazu, verringert sich die Zahl sogar auf zwei Dokumente, was 5,9 % Überlappung entspricht.

Wie sehr weichen die einzelnen Beurteilungen von der Expertenmeinung ab?

Saracevic (2007b) schätzte die Übereinstimmung zweier Bewertenden auf ca. 30 %, was in dem aktuellen Experiment zehn gleich bewerteten Dokumenten entsprechen würde. Mit durchschnittlich 17,8 Übereinstimmungen liegen die Werte in diesem Experiment deutlich darüber. Im folgenden Diagramm (Abbildung 18) ist der Grad der Übereinstimmung der Bewertungen mit der Expertenmeinung in Prozent dargestellt. Wie man sieht, bewegen sich alle, trotz insgesamt unterschiedlicher Bewertungen, in einem relativ engen Rahmen mit nur einem Ausreißer nach unten (Nutzer 5).

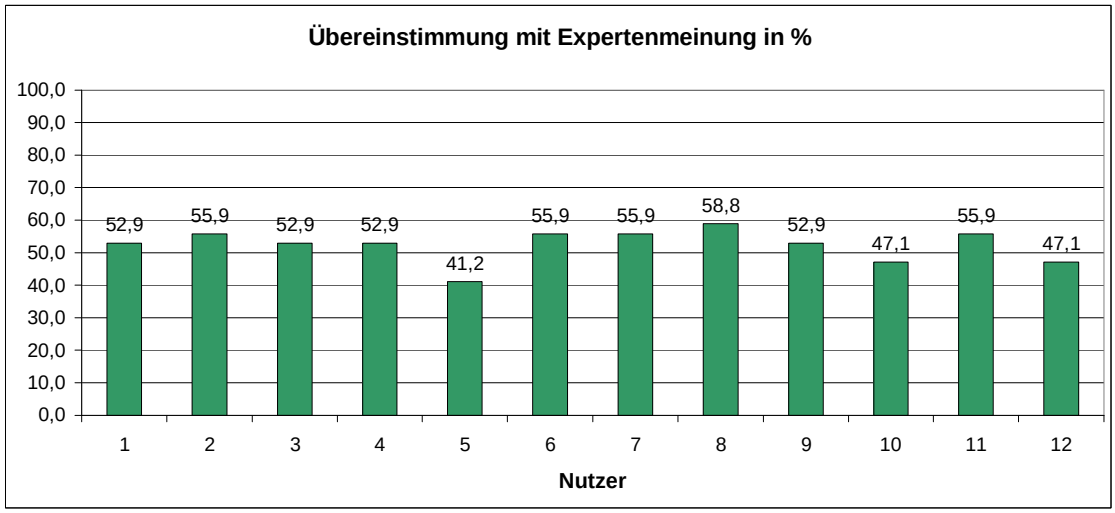


Abbildung 18: Grad der Übereinstimmung mit Benchmark-Bewertung

5.4 Einfluss von Dokumentlänge und –alter auf die Bewertung

Die Dokumente in der Ergebnisliste hatten nicht alle die selbe Länge. Manche Abstracts bestanden aus einem deutschen Autorenreferat, einem Inhaltsverzeichnis und einem englischen Abstract. Einige Nutzer äußerten im Nachhinein, dass die Darstellung des Textes ohne Formatierungen anstrengend und abschreckend sei.

Beeinflusst die Dokumentlänge die Relevanzbewertung? Werden Dokumente mit langen Abstracts eher nicht-relevant bewertet?

In Abbildung 19 ist dargestellt wie sich die Anzahl der relevanten Bewertungen pro Dokument verändert, wenn diese nach Anzahl der Wörter von wenig (72 Wörter) nach viel (471 Wörter) geordnet sind. Die hinzugefügte Trendlinie zeigt die Tendenz, dass mit zunehmender Wortzahl die Anzahl der positiven Bewertungen abnimmt. Die beiden wortreichsten Dokumente erhielten gar keine relevanten Bewertungen. Sie wurden allerdings auch vom Experten als nicht relevant beurteilt, sich so dass das Ergebnis nicht nur auf die Länge des Abstracts zurückführen lassen kann. Aber auch ohne die beiden Dokumente bleibt die Tendenz erhalten, dass Dokumente mit sehr langen Abstracts eher nicht-relevant bewertet werden.

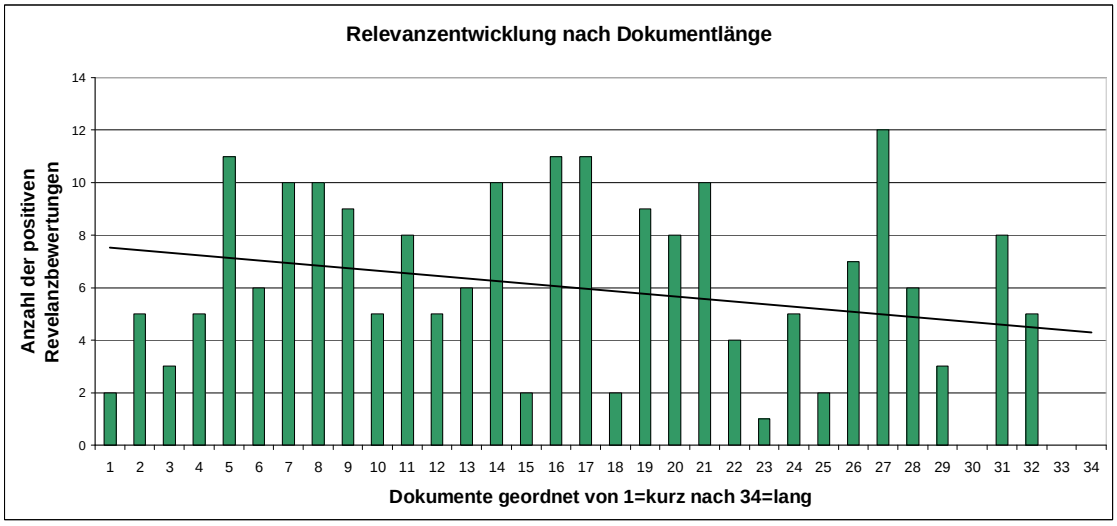


Abbildung 19: Entwicklung der positiven Relevanzurteile nach Länge des Dokuments von kurz nach lang

Beeinflusst das Dokument-Alter die Relevanzbewertung? Werden ältere Dokumente eher nicht-relevant bewertet?

Eines der beiden wichtigsten Relevanzkriterien nach Xu und Chen (siehe Kapitel 3.5.1) ist die Neuheit eines Dokuments. Im Gespräch nach der Bewertung haben einige Nutzer angegeben, dass sie ältere Dokumente aus dem 1980er Jahren eher als nicht-relevant eingestuft hätten, weil sie den Begriff „Neue Medien“ aus der Aufgabenstellung eher auf aktuelle Technik, Computer und Internet bezogen haben.

Die Anzahl der positiven Relevanzurteile nimmt jedoch in der Tendenz ganz leicht ab, je jünger die Dokumente werden (Abb. 20). Dies kommt vor allem deshalb zustande, weil die beiden neusten (und längsten) Dokumente von 2005 und 2008 von allen als nicht-relevant beurteilt wurden, auch vom Experten.

Betrachtet man die Verteilung der vorhergehenden 32 Dokumente gesondert, dann ist die Tendenz umgekehrt und würde somit eher der These entsprechen, dass Neuheit ein wichtiges Relevanzkriterium ist. Aber auch in diesem Fall wäre der Anstieg eher gering (Abb. 21). Das Alter hatte also trotz gegenteiliger Aussagen keinen Einfluss auf die Relevanzbewertung.

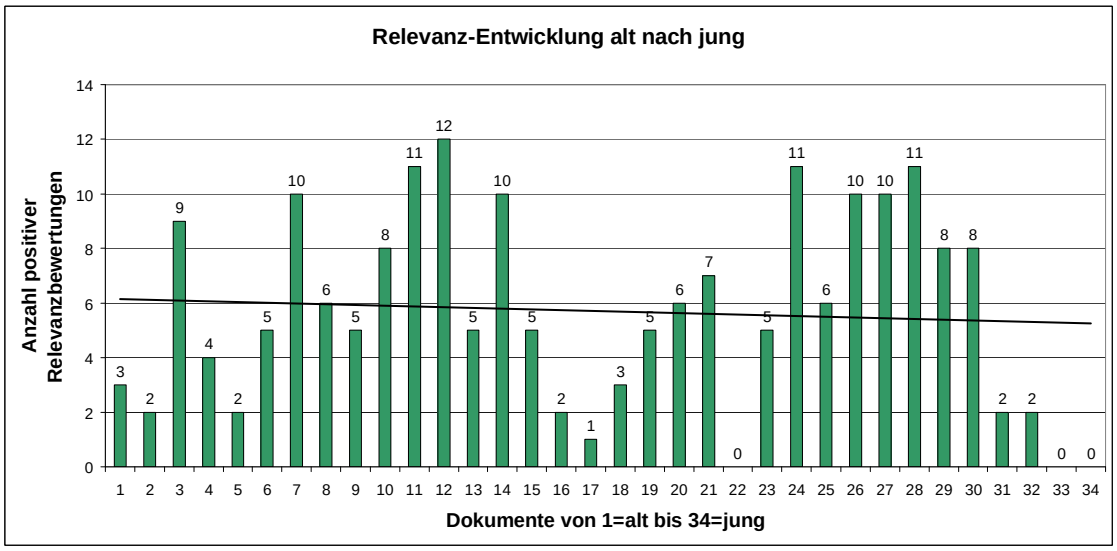


Abbildung 20: Entwicklung der positiven Relevanzurteile geordnet nach Alter der Dokumente von alt nach jung

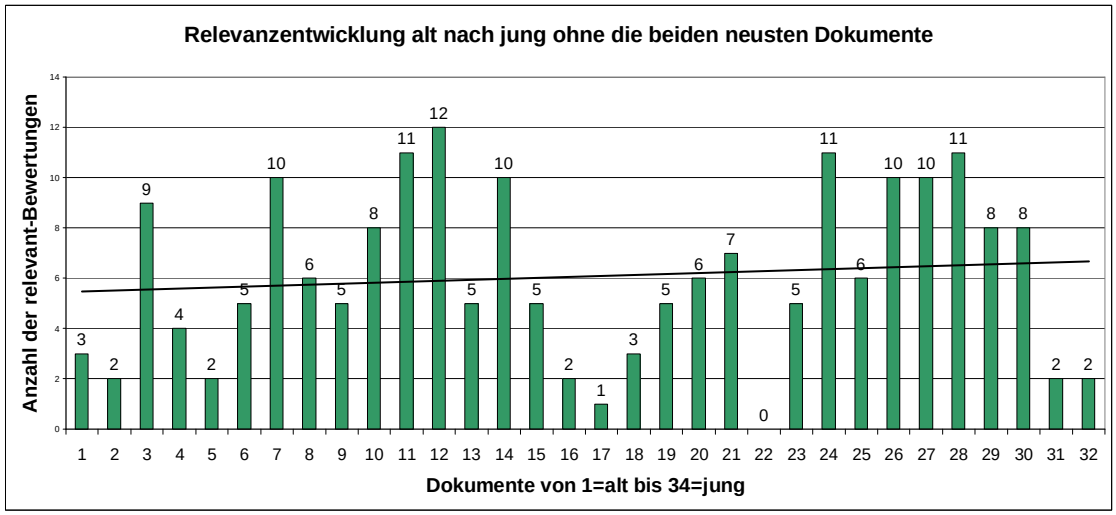


Abbildung 21: Entwicklung der positiven Relevanzurteile geordnet nach Alter der Dokumente von alt nach jung, ohne die beiden jüngsten Dokumente

5.5 Wichtigkeit der Datenelemente

Auf welchen Teil der Dokumentrepräsentationen wird am meisten vertraut, wenn Relevanzentscheidungen getroffen werden?

Es würde keinen Sinn machen die Anzahl der Fixationen in den AOIs über alle Dokumente hinweg zu vergleichen, da die Abstracts ohnehin aus viel mehr Text bestehen als Titel und Deskriptoren und schon allein dadurch viel mehr Fixationen enthalten.

Die Wichtigkeit der Datenelemente in dem vorliegenden Experiment wird stattdessen auf drei andere Arten festgestellt: Vergabe von Punkten im Interview, Beobachtung welcher Bereich zuletzt vor einer Entscheidung betrachtet wurde (Absprungmarken) und die Anzahl der Besuche in den AOIs (Visit Count).

1. Vergabe von Punkten im Interview

Da keiner der Tester im Bereich Medienpädagogik oder Erziehungswissenschaft spezielles Vorwissen hatte, waren allen die Autoren unbekannt. Von daher verwundert es nicht, dass diesem Datenelement nur vier Punkte zugesprochen wurden. Dies entspricht einem durchschnittlichen Wert von 0,3 Punkten pro User.

Beim Publikationsjahr gingen die Meinungen am stärksten auseinander. Es wurden sowohl 0 als auch 10 Punkte vergeben. Dies lag vor allem daran, dass der Begriff „Neue Medien“ aus der Aufgabenstellung unterschiedlich interpretiert wurde und für einige ausschließlich Computer und Internet umfasste. Diese Nutzer haben angegeben ältere Dokumente (z. B. aus den 80ern) eher als nicht-relevant eingestuft zu haben. Bei einer genaueren Betrachtung der Bewertungen dieser User stellt sich dies

aber nicht als zutreffend heraus. Der Gesamteindruck aus Kapitel 5.4 bestätigt, dass das Alter der Dokumente kein bemerkenswert wichtiger Faktor bei der Bewertung gewesen sein kann.

Auch bei den Deskriptoren gingen die Meinungen auseinander, es wurden Werte von 2 bis 10 vergeben. Sie lagen mit insgesamt 80 Punkten auf Rang 3.

Das Abstract wurde mit Abstand als wichtigstes Datenelement bewertet. Dicht gefolgt vom Titel, der mit 3 mal 10 Punkten am häufigsten die Höchstwertung erhalten hat. Der Titel wurde jedoch von einigen zurückhaltender bewertet, da er in manchen Fällen allein nicht ausreichend genug den tatsächlichen Inhalt beschreibt.

Tabelle 3: Wichtigkeit der Datenelemente: Punktevergabe

	Publikationsjahr	Autor	Titel	Abstract	Deskriptoren
Nutzer 1	10	0	8	10	2
Nutzer 2	4	1	7	8	4
Nutzer 3	8	0	10	6	9
Nutzer 4	1	0	8	9	8
Nutzer 5	2	0	10	10	5
Nutzer 6	6	0	7	8	8
Nutzer 7	4	0	7	9	5
Nutzer 8	5	0	7	9	6
Nutzer 9	3	0	10	8	10
Nutzer 10	0	0	8	7	10
Nutzer 11	2	0	4	8	7
Nutzer 12	9	3	6	9	6
Ø	**4,50**	**0,33**	**7,67**	**8,42**	**6,67**
Sum.	54	4	92	101	80

2. Absprungmarken

Um herauszufinden, zu welchem Zeitpunkt die Relevanzentscheidungen fallen, wurde der zuletzt vor der Entscheidung betrachtete Bereich im Dokument isoliert betrachtet. Viele Testpersonen haben ausgesagt, dass sie ein Dokument so lange lesen, bis sie sichere Hinweise auf die Relevanz erhalten. Demnach müsste der zuletzt betrachtete Punkt die Stelle sein, an der die Tester ausreichend Informationen für eine fundierte Entscheidung gefunden haben. Dies bedeutet aber nicht, dass dies auch der Bereich des Dokuments ist, der tatsächlich am wichtigsten für die Entscheidung war, welche schon vorher getroffen worden sein kann. Häufig wurde im Interview gesagt, dass die letzten Blicke der Entscheidungs-„Absicherung“ dienen. Die Absprungmarke kann also sowohl entscheidend hinweisgebender Punkt, als auch bekräftigend für eine

schon getroffene Entscheidung sein. Er führt auf jeden Fall zu einer finalen Entscheidung und Ausführung des Klicks und hat von daher eine herausragende Bedeutung für den Bewertenden. Was für einen Nutzer in einem speziellen Dokument entscheidend ist, muss aber nicht auch für einen anderen User entscheidend sein.

Tabelle 4: Auszählung der Absprungmarken

	Nutzer 1	Nutzer 3	Nutzer 4	Nutzer 5	Nutzer 8	Nutzer 9	Nutzer 10	Nutzer 11	Nutzer 12	Sum.
T	7	11	5	10	10	2	7	0	9	61
A	26	6	10	13	18	10	9	26	11	129
D	1	17	19	11	6	22	18	8	14	116

Das Abstract war 129-mal Absprungmarke und bestätigt damit seine Stellung als Datenelement, das die meisten Relevanz-Hinweise liefert.

Für vier Nutzer war das Abstract häufigster Absprung-Punkt. Auffällig ist, dass das Abstract zwar insgesamt die häufigste Absprungmarke ist, dass fünf Nutzer ihre Entscheidungen aber häufiger nach Lektüre der Deskriptoren getroffen haben. Beide Datenelemente sind also wichtige Hinweisgeber auf Relevanz.

Der Titel war nur halb so oft Absprungmarke wie das Abstract und liegt damit in diesem Ranking auf dem letzten Platz.

3. Visit Count

Die Lookzones „Titel“ und „Abstract“ wurden beide gleich oft besucht. Deskriptoren wurden deutlich weniger häufig besucht. Es ist aber problematisch nur deshalb davon auszugehen, dass die Deskriptoren weniger Hinweise auf Relevanz liefern, weil sie weniger oft gelesen werden. Mehrfaches Lesen einer AOI kann auch Verständnisprobleme und Unsicherheiten bedeuten, was ebenfalls Einflussfaktoren für die Entscheidung sind. Es kann hier lediglich mit Sicherheit festgehalten werden, dass das Schlagwortfeld nicht so häufig beachtet wird und deshalb kein Datenfeld ist, welches maßgeblich die Relevanzentscheidung beeinflusst hat. Ob es das aber getan hätte, wenn die Tester das Feld häufiger gelesen hätten, kann man hier nicht ableiten.

Tabelle 5: Anzahl der Besuche in den AOIs

AOI	**Anzahl Besuche**
T	372
A	373
D	204

Zusammengefasst zeigt sich, dass das Abstract in allen Punkten auf Platz eins und somit wichtigstes Datenelement für das Erkennen von Relevanz ist. Dies ist übereinstimmend mit den Ergebnissen einer Studie von Joseph Janes (1991) „Abstracts are by far the most important field and have the greatest impact, followed by titles, bibliographic information and indexing." (S. 629)

5.6 Bewertungstypen

Lassen sich die Benutzer in das Klassifikationsschema von Aula et al. einordnen?

Wie in Kapitel 2.6 beschrieben, konnten Aula et al. die Teilnehmer ihrer Eyetracking-Studie in zwei Klassen von Bewertungstypen einordnen. Zum einen die besonders gründlichen Bewerter, die exhaustive evaluators. Zum anderen die economic evaluators, die sich mit wenig Informationen zufriedengeben bevor sie eine Entscheidung treffen.

Die Teilnehmer diese Studie lassen sich ebenfalls in die beiden Gruppen einsortieren. Es wurden nur die Nutzer berücksichtigt, für die brauchbare Eyetracking-Daten vorlagen. Die Zugehörigkeit zu einer der beiden Klassen wurde durch folgende Kriterien festgestellt: Gesamtlänge der Bewertung, durchschnittliche Anzahl der Fixationen pro Dokumentbewertung, Anzahl der besuchten AOIs insgesamt (zur Erinnerung: 3 AOIs pro Dokument, 34 Dokumente in der Ergebnisliste) sowie durchschnittliche Länge des Scanpfades pro Nutzer über alle Bewertungen. Für alle Kriterien wurden Tabellen erstellt, die dann von geringen zu hohen Werten sortiert wurden.

Teilt man alle vier Tabellen (Tab. 6, Tab. 7, Tab 8 und Tab. 9) in der Mitte in zwei Gruppen, so beinhalten diese jeweils dieselben Nutzer, wenn auch nicht immer in derselben Reihenfolge. Die oberen vier Nutzer (1, 11, 8, 9) bewerteten folglich am schnellsten, fixierten dabei am wenigsten Punkte innerhalb der Dokumente, trafen ihre Entscheidungen auf Basis der wenigsten Datenelemente insgesamt und hatten außerdem die kürzesten Scanpfade. Dies qualifiziert sie für die Einordnung als „**economic evaluators**".

Die unteren vier Nutzer hingegen (3, 4, 12, 5) brauchten deutlich länger, fixierten mehr Punkte, besuchten mehr Datenelemente insgesamt und bewerteten die Dokumente auf Basis von mehr Informationen als die economic evaluators. Daher kann man diese Nutzer zur Gruppe der „**exhaustive evaluators**" zählen.

Tabelle 6: Gesamtlänge der Bewertung in Minuten

Nutzer 8	10,98
Nutzer 11	12,63
Nutzer 1	13,32
Nutzer 9	13,76
Nutzer 5	15,12
Nutzer 12	17,41
Nutzer 4	23,10
Nutzer 3	23,50

Tabelle 7: Durchschnittliche Anzahl der Fixationen pro Dokumentbewertung

Nutzer 1	52,3
Nutzer 8	62,7
Nutzer 11	66,8
Nutzer 9	68,0
Nutzer 5	88,2
Nutzer 12	95,1
Nutzer 4	108,1
Nutzer 3	111,4

Tabelle 8: Anzahl der besuchten AOIs insgesamt

Nutzer 1	70
Nutzer 11	85
Nutzer 8	95
Nutzer 9	107
Nutzer 5	117
Nutzer 3	119
Nutzer 4	122
Nutzer 12	124

Tabelle 9: Durchschnittliche Länge des Scanpfades über alle Bewertungen

Nutzer 1	2,06
Nutzer 11	2,5
Nutzer 8	2,79
Nutzer 9	3,15
Nutzer 5	3,44
Nutzer 3	3,5
Nutzer 4	3,59
Nutzer 12	3,65

Nutzer 9 und 5 finden sich bei allen Kriterien im Mittelfeld wieder. Sie haben zwar eindeutige Tendenzen, ihre Werte unterscheiden sich aber nicht immer so deutlich wie die der darüber oder darunter befindlichen Nutzer.

Gibt es einen Zusammenhang zwischen Bewertungstyp und Relevanzbewertung?

Es gibt zwei Aspekte, bei denen sich Beziehungen zwischen Bewertungstyp und Relevanz erkennen lassen. Zum einen sind die beiden Nutzer, die am meisten Dokumente relevant bewerten, exhaustive evaluators. Die drei Nutzer, die am wenigsten Dokumente relevant bewerteten, sind economic evaluators (Tab. 10). Zugehörigkeit zu einer Bewertungsgruppe kann also darauf hinweisen ob man mehr oder weniger Dokumente relevant bewertet.

Tabelle 10: Verhältnis relevant / nicht-relevant

	Verhältnis relevant / nicht relevant
Nutzer 12	23/11
Nutzer 3	21/13
Nutzer 4	18/16
Nutzer 5	18/16
Nutzer 9	18/16
Nutzer 1	14/20
Nutzer 11	14/20
Nutzer 8	8/26

Zum Anderen sind die economic evaluators näher an der Expertenbewertung. Vergleicht man die Übereinstimmung der Relevanzbewertungen mit der Expertenbewertung, dann ist die Verteilung der Nutzer zwar etwas gemischter, als bei den anderen Kriterien, eine Tendenz ist dennoch erkennbar. So sind zwei der economic evaluators mit 19 und 20 gleich bewerteten Dokumenten an der Spitze, zwei exhaustive evaluators haben mit 16 und 14 Übereinstimmungen die größte Entfernung zum Experten.

Die economic evaluators waren somit effizienter. Dies bestätigt die Erkenntnisse von Aula et al. 2005. Auch da waren die economic evaluators effizienter.

Tabelle 11: Übereinstimmungen mit Expertenbewertung

	Übereinstimmungen
Nutzer 8	20
Nutzer 11	19
Nutzer 1	18
Nutzer 3	18
Nutzer 4	18
Nutzer 9	18
Nutzer 12	16
Nutzer 5	14

5.7 Unterschiede nach Geschlecht

Weibliche Tester bewerten eher langsamer, fixieren mehr Punkte, besuchen mehr AOIs und haben längere Scanpfade als die männlichen Tester. Dies bestätigt sich, wenn man die Aufteilung der economic und exhaustive evaluators betrachtet. Die

Gruppe der gründlicheren Bewerter besteht aus drei weiblichen und einem männlichen Tester, die Gruppe der ökonomischen Bewerter besteht zu ¾ aus männlichen Testern.

Ein Punkt, der dabei besonders auffällt, ist die Verteilung der Fixationen im Schlagwortfeld. Während Titel und Abstract ähnlich hohe Werte haben, fixierten weibliche Tester fast doppelt so viele Punkte im Deskriptorenfeld wie männliche Tester.

Tabelle 12: Einzelergebnisse weibliche Tester – männliche Tester

	Gesamtdauer	Anzahl relevante Dokumente	Fixationen pro Bewertungssequenz	Anzahl besuchter AOIs	Länge Scanpfad	Fixationen nach AOI		
						T	A	D
Ø weibl.	17,0	19,0	90,0	107,0	3,1	12,3	66,9	10,8
Ø männl.	13,0	18,5	73,2	102,8	3,0	11,3	56,1	5,8

5.8 Unterschiede zwischen Bachelorstudenten und Masterstudenten

Gibt es Unterschiede zwischen Bachelor- und Masterstudenten?

Die Werte von Bachelor- und Masterstudenten unterscheiden sich kaum. Den einzigen deutlicheren Unterschied gibt es bei der Anzahl der Fixationen insgesamt und speziell im Deskriptorenfeld. Das heißt die Bachelorstudenten haben das Deskriptorenfeld weniger oft oder weniger genau gelesen. Dies steht zunächst im Gegensatz zu den Punktebewertungen für die Wichtigkeit der Datenelemente (siehe S. 55). Hier gaben die Bachelor dem Deskriptorenfeld viel mehr Punkte für die Wichtigkeit als die Master, nämlich durchschnittlich 8,3 im Gegensatz zu 5,9 Punkten der Master. Aber allein die Tatsache, dass ein Bereich häufig oder genau gelesen wird, bedeutet nicht automatisch, dass er auch bedeutsam für die Relevanzentscheidung ist.

Tabelle 13: Einzelergebnisse Master - Bachelor

	Gesamtdauer	Anzahl relevante Dokumente	Fixationen pro Bewertungssequenz	Anzahl besuchter AOIs	Länge Scanpfad	Fixationen nach AOI		
						T	A	D
Ø Master	15,5	19,0	84,5	104,6	3,1	11,5	63,4	9,6
Ø Bachelor	14,1	18,5	76,6	105,3	3,1	12,3	58,3	6,1

Sind Masterstudenten eher economic evaluators?

Aula et al. (2005) vermuten, dass es einen Zusammenhang zwischen Erfahrung und Bewertungstyp gibt. In ihrer Studie haben sie festgestellt, dass die economic evaluators mehr Routine im Umgang mit Computern hatten, als die exhaustive evaluators. „Thus, the result evaluation style seems to evolve towards a more economic style as the users gain more experience." (S. 1058). Für dieses Experiment lässt sich diese These aber nicht bestätigen, wenn man davon ausgeht, dass die Masterstudenten die erfahreneren Nutzer sein sollen. Tatsächlich besteht die Gruppe der economic evaluators aus jeweils zwei Master- und Bachelorstudenten, in der Gruppe der exhaustive evaluators befinden sich sogar drei Masterstudenten.

6 Zusammenfassung und Diskussion

Im vorherigen Kapitel wurden die Einzelergebnisse der Eyetracking-Studie detailliert vorgestellt. In diesem Kapitel werden die Haupt-Erkenntnisse noch einmal kurz zusammengefasst. Außerdem werden Bedingungen beschrieben, die zur Einordnung der Studie dienen. Schließlich werden Gestaltungshinweise für die Darstellung von Ergebnislisten und zukünftigen Studien gegeben.

Es gibt zwei Bewertungstypen, economic und exhaustive evaluators. Die economic evaluators, überwiegend männliche Nutzer, treffen ihre Entscheidungen schnell, fixieren dabei wenig Punkte, besuchen die wenigsten Datenelemente und haben die kürzesten Scanpfade. Sie sind außerdem die effektiveren Bewerter, denn ihre Übereinstimmungen mit der Expertenbewertung sind um einiges höher als die der exhaustive evaluators. Der gründliche Bewertungstyp hingegen, größtenteils weiblich, braucht insgesamt viel länger, liest mehr und genauer, hat längere Scanpfade und bewertet auch mehr Dokumente relevant als nicht relevant.

Die Aussagen der Nutzer im Anschluss an die Bewertung stützen diese Resultate. Die langsameren exhaustive evaluators betonten beispielsweise, dass sie lieber auch allgemeinere, thematisch nicht so spezifische Dokumente relevant bewertet hätten, für den Fall dass etwas Interessantes enthalten sein könnte. Solche Dokumente könnten auch als Überblicks-Informationen dienen, vielleicht statistische Daten oder weiterführende Literaturhinweise enthalten. Die exhaustive evaluators beurteilten eher auch dann relevant, wenn sie sich nicht ganz sicher waren, damit keine potenziell relevante Information verloren geht. Sie waren eher bereit auch Sammelwerke relevant zu bewerten, in denen vielleicht nur ein Beitrag thematisch passend sein könnte. Die economic evaluators beurteilten indessen bei Unsicherheit Dokumente eher nichtrelevant. Wichtiges Kriterium für diesen Typ der Bewerter war die hohe topikalische Übereinstimmung der Dokumente. Zu allgemeine Werke oder Sammelbände hätten zu viel Arbeitsaufwand für zu wenig Information bedeutet, so die Aussagen der economic evaluators.

Nur vier von 34 Dokumenten wurden von allen gleich bewertet. Das spricht dafür, dass die Nutzer sehr unterschiedliche Relevanzkriterien angelegt haben bzw. ähnliche Kriterien mit unterschiedlichen Gewichtungen belegt haben.

Grundsätzlich ist festzuhalten, dass die Ergebnisse nicht verallgemeinerbar sind, sondern nur in dem aktuellen Zusammenhang gültig sind. Die Studie wurde mit nur zwölf Testpersonen durchgeführt. Um die Erkenntnisse zu verifizieren, müsste man sie mit größeren Personenmengen und mehreren Aufgabentypen, Schwierigkeitsgraden, Wissensgebieten usw. erneut überprüfen. Es lag keine natürliche Recherche-Situation vor, die Tester hatten kein echtes eigenes Informationsbedürfnis und wenig Kontext für die Aufgabenstellung.

Obwohl 12 User teilgenommen haben, konnten nur die Eyetracking-Daten von 8 tatsächlich ausgewertet werden. Bei einem User gab es bereits bei der Kalibrierung Unregelmäßigkeiten. Bei ihm war die Eyetracking-Aufnahme gar nicht brauchbar. Bei den anderen drei Usern wurde die Entscheidung getroffen die Daten nicht zu benutzen, da sie insgesamt weniger genau waren als die acht, die letztlich verwendet wurden. Dadurch dass die Datenelemente in der Testoberfläche so nah zusammen stehen, dass in einigen Fällen nicht ganz klar ist, ob der User zum Beispiel noch die letzte Zeile des Abstracts, oder schon die Deskriptoren liest, wurden nur die besten Aufnahmen verwendet. Dennoch können auch bei diesen Aufnahmen Fixationen außerhalb der definierten AOIs vorkommen. Es kann sein, dass zwar Inhalt wahrgenommen wurde, aber die Fixationen nicht mitgezählt wurden. Daher können die Zahlen nur als ungefähre Anhaltspunkte dienen.

Die Darstellung von aggregierten Fixationsschwerpunkten (Cluster) war nicht möglich, da die Bewertungsszenen, durch die immer wieder unterschiedlichen Reihenfolgen der Anzeige, nicht bei allen Testpersonen im gleichen Bildausschnitt waren. Eine automatische Analyse aller Bewertungen für ein spezifisches Dokument, bei dem die Software Tobii Studio Cluster mit Schwerpunkten der Blickfixierungen errechnet hätte, war somit leider nicht möglich.

Die Eigenschaften der Relevanz wie sie in Kapitel 3.1 beschrieben wurden (dynamisch, subjektiv, relativ, intuitiv, messbar), können auch in diesem Experiment bestätigt werden. So weisen die Unterschiede in Gesamt- und Einzelbewertungslängen, Fixations-Zahlen, Blicksequenzen und das niedrige Interrater Agreement auf die hohe Subjektivität und Relativität der Relevanz hin. Die Testpersonen waren zwar fortgeschrittene User von Suchmaschinen, aber Laien im Wissensgebiet des Topics. Dass nur vier von 34 Dokumenten gleich bewertet wurden zeigt, dass jeder Nutzer ganz individuelle Relevanzkriterien und Gewichtungen angewendet haben muss. Keinem Nutzer wurden Richtlinien vorgegeben, worauf besonders geachtet werden soll. Alle wussten instinktiv was mit Relevanz gemeint ist. Durch die binäre Relevanzentscheidung konnte gemessen werden wie oft ein Dokument relevant bewertet wird. Die Dy-

namik der Relevanz spiegelt sich in den Aussagen der Tester wieder, die nach der Bewertung die Ergebnisse gerne noch einmal re-evaluiert hätten, da sie im Laufe der Bewertung einen besseren Überblick über das Thema und dessen Vokabular erhalten haben. Das spricht dafür, dass während der Bewertung ein Lernprozess stattgefunden hat, der sich auf die Relevanzentscheidungen auswirkt.

Für die Gestaltung von Ergebnislisten ergibt sich aus der Studie, dass weniger oft mehr ist. Sehr lange Abstracts, bei denen viel Text auf engem Raum steht, schrecken eher ab. Sehr häufig wurde nur die obere Hälfte der Abstracts oder weniger gelesen (für ein Beispiel siehe Anhang B). Sogar kurze Abstracts wurden häufig nicht komplett gelesen. Das Abstract ist aber nach Auswertung dreier verschiedener Parameter das wichtigste Datenelement für die Ableitung von Relevanz. Für Dokumentare heißt das, die Kerninformationen möglichst kurz und treffend am Anfang des Abstracts zusammenzufassen. Man könnte auch in den Dokumentrepräsentationen auf den Ergebnislisten zunächst neben Titel und Deskriptoren nur ein indikatives Abstract anzeigen. Bei Bedarf können die Nutzer dann ein längeres Referat anklicken.

Eine weitere Überlegung wäre es nach dem Titel gleich die Deskriptoren anzuzeigen und dann erst das Abstract. Die Schlagwörter wurden mit 80 Punkten noch als eines der drei wichtigsten Datenelemente zum Ablesen der Relevanz eingestuft, die Fixations-Zahlen zeigen aber, dass das Feld weit weniger oft gelesen wird. In dem Test befand sich das Deskriptorenfeld unter dem (zum Teil sehr langen) Abstract. Durch eine geschicktere Platzierung könnte es den Nutzern erleichtert werden schneller die wichtigsten Informationen überblicken zu können.

Fast alle Testpersonen gaben an, dass sie Schwierigkeiten mit der sehr engen Darstellung, dem geringen Zeilenabstand und den fehlenden Formatierungen hatten. In einigen Abstracts wurden Inhaltsverzeichnisse aufgeführt, was alle Befragten grundsätzlich für sinnvoll hielten. Sie fühlten sich aber durch die schlichte Aneinanderreihung des Textes ohne Absätze, Hervorhebungen oder Nummerierungen eher vom Lesen abgeschreckt.

In zukünftigen Studien könnte daher auch mit verschiedenen Versionen von Ergebnislisten mit unterschiedlichen Darstellungsformen gearbeitet werden. Es sollte dann aber darauf geachtet werden, dass die Präsentation der Dokumente technisch so realisiert wird, dass alle als Webseite mit eigener URL angezeigt werden oder so, dass jede Bewertungssequenz bei jedem Nutzer den gleichen Bildausschnitt hat. So könnten von der Software automatisch Bewertungen mehrerer Nutzer gleichzeitig ausgewertet werden. Außerdem sollte Scrollen möglichst vermieden werden, da auch das die Analyse und den Vergleich erheblich erschwert.

7 Fazit

Obwohl durch den Testaufbau die Rahmenbedingungen für jede Bewertung sehr ähnlich waren (homogene Gruppe von Testern, gleicher Ablauf), zeigt sich doch, dass Menschen keine Roboter sind und ein höchst individuelles Verständnis von Relevanz, individuelle Lese-Gewohnheiten, eigene Auffassungen der Themenstellung usw. haben.

Dennoch lassen sich auch in der großen Vielfalt individueller Persönlichkeiten und Verhaltensweisen bestimmte Gemeinsamkeiten und Muster erkennen. Anhand von vier Parametern, nämlich Länge der Gesamtbewertung, Anzahl der Fixationen, Anzahl der besuchten AOIs und Länge des Scanpfades, konnten alle Nutzer einer von zwei Bewertungsgruppen zugeordnet werden, den ökonomisch handelnden „economic evaluators“ und den gründlicheren „exhaustive evaluators“. Die Idee für die Bewertungsgruppen stammte aus einer Arbeit von Aula et al. 2005, das Modell wurde in dieser Studie aber ergänzt und mit mehr Kriterien spezifiziert.

In wie weit die Erkenntnisse aus diesem kleinen Versuchsaufbau zu verallgemeinern sind, muss in weiteren Tests mit größeren Mengen von Testpersonen überprüft werden.

Anhang A: Scanpfade als Sequenzen

A.1 Scanpfade Teil 1

Dokument	Nutzer 1	Nutzer 3	Nutzer 4	Nutzer 5	Nutzer 8	Nutzer 9	Nutzer 10	Nutzer 11	Nutzer 12
Auchter, Roman (2004): Alle Macht den Bild-Medien? : Bildkommunikation von Kindern und Jugendlichen	TA	TADA	TAD	TDA	TA	TADD	TATD	ADA	TADA
Voullieme, Helmut (1986): Medientechnologie gegen Lebenswirklichkeit?	A	TAD	TABDTA	TATA	TA	TDA	TAT	A	TA
Weißeno, Georg (2002): Politikunterricht im informationszeitalter	A	A	TADA	TADTA	TA	TAD	TADT	AD	TATAD
Holtappels, Heinz Günter; Klemm, Klaus; Pfeiffer, Hermann; Rolff, Hans-Günter; Schulz-Zander, Renate (2004): Jahrbuch der Schulentwicklung : Daten, Beispiele und Perspektiven. Bd. 13	A	ATAD	TADADA	TADT	TA	ATD	TAD	TADA	TAD
Schnoor, Detlev (1998): Schulentwicklung durch neue Medien	AT	ADTA	TATADT	TDATA	TAT	TAD	TAD	A	TATAD
Rüden, Peter von (1987): Thesen zur Medienpädagogik in der Erwachsenenbildung	AT	TADT	TAD	TADT	TAD	TATD	TAD	AD	TATAT
Bachmair, Ben; Diepold, Peter; Witt, Claudia de (2003): Jahrbuch Medienpädagogik 3	TA	TAD	TADA	TASA	TAT	ATAD	TADT	TADADA	TATATAD
Ohle-Nieschmidt, Hannelore (2002): Mediale und reale Lernwelten : noch ein Widerspruch, bald integrale Bestandteile des Schulalltags	TTA	A	TADTA	T AD	TA	TAD	TAD	A	TTAT
Eckert, Andreas; Hofer, Manfred (1999): Wissenserwerb durch 'interaktive' neue Medien : aus Sicht der Erziehungswissenschaft	TTTA	TAD	TAD	TT	T AD	TDA	TAD	ATA	TADT
Ernst, Annette; Pullich, Leif (2001): Fernstudium (FESTUM) - ein medienpädagogisches Zusatzstudium	TA	ATADTT	TADT	TA	ATADA	TAD	TAD	A	TATAD
Krotz, Friedrich (2000): Kids und neue Medien : Netz- oder Pixelgesellschaft?	TA	TADTTA	TAD	TAD	TA	TAD	TAD	TADA	TADTT
Seeber, Franziska (2002): Chancen und Möglichkeiten Neuer Medien in der Schule : können andere Lehr- und Lernformen durch Neue Medien entstehen?	TA	TAD	TADT	TDA	TA	TAD	TATA	TA	TA
Schachtner, Christina (2002): Entdecken und Erfinden : neue Medien - neues Lernen?	TAD	TADT	TADT	TADA	TAT	TADT	TADA	A	TA
Laurien, Hanna-Renate (1987): Bildungspolitische Aufgaben der Medienpädagogik	TA	ATAT	TAD	TAT	ADA	TAD	TADT	ADA	TAT
Rössler, Patrick; Krotz, Friedrich (2005): Mythen der Mediengesellschaft	A	TATD	TAD	TTA	AT	TDA	TADA	A	TAD
Oberliesen, Rolf; Stiebeling, Anneliese (1988): Neue Medien, neue Technoligien : Bildung und Erziehung in der Krise?	T A	TAD	TADTA	T AD	TAD	TAD	TADT	ADA	TAD
Boehnke, Klaus; Münch, Thomas (2001): Radio, Musikfernsehen und Internet in der Entwicklung Jugendlicher	TA	T A D	TAD	TADT	TTA	TDA	TAD	ADA	TATAD
Kübler, Hans-Dieter (1984): 'Neue Medien' in die Schulen? : Probleme und Risiken der Medienpädagogik an der Schwelle zum 'Informationszeitalter'	TA	TADTAT	TDA	T	TTADAT	TAD	TAD	A	TADA
Ortner, Gerhard E. (1984): Bildschirm - Bildung? : pädagogische und politische Perspektiven der Neuen Medien	TTA	AT	TAD	TADTADT	TADAT	TA	TA	A	TA

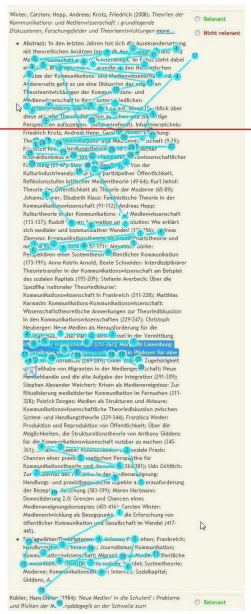
A.2 Scanpfade Teil 2

Dokument	Nutzer 1	Nutzer 3	Nutzer 4	Nutzer 5	Nutzer 8	Nutzer 9	Nutzer 10	Nutzer 11	Nutzer 12
Ortner, Gerhard E. (1984): Bildschirm - Bildung? : pädagogische und politische Perspektiven der Neuen Medien	TTA	AT	TAD	TADTADT	TADAT	TA	TA	A	TA
Lauffer, Jürgen; Volkmer, Ingrid (1995): Kommunikative Kompetenz in einer sich verändernden Medienwelt	TA	ADT	TAD	TTADT	TA	TDA	TAD	ADD	TAD
Pfeiffer, Hermann; Rolff, Hans-Günter (1986): Technologische Grundbildung : oder: wie Schulen auf die 'Informationsgesellschaft' vorbereiten	T	TATD	TAD	TAD	ATATA	TATAD	TAD	A	TA
(2000): Die Internationalisierung des Bildungswesens : Dokumentation der 20. DGBV-Jahrestagung vom 4. bis 6. November 1999 in Bochum	TA	T AT	TA	TADTA	TA	TAD	TAD	ATD	TADT
Kaden, Michael (2002): Medienerziehung in Portugal	AT	TAD	TAD	TAD	A	TAD	TA	ADADA	TATA
Boehnke, Klaus; Döring, Nicola (2001): Neue Medien im Alltag : die Vielfalt individueller Nutzungsweisen	T AT	TDAT	T AD	TA	AAT	TATAD	ATADT	A	TAD
Schorb, Bernd; Faulstich-Wieland, Hannelore; Fauser, Richard; Haan, Oswald; Jourdan, Manfred; Kupser, Paul; Rolff, Hans-Günter; Schell, Alfred; Spanhel, Dieter; Vogelgesang, Waldemar (1989): Bildung trotz Computer? : eine Zwischenbilanz des informationstechnischen Unterrichts	T	TAD	TAD	TAD	T AD	TATD	TA	TA	TATAD
Krotz, Friedrich; Hasebrink, Uwe (2002): Medienkompetenz von Kindern und Jugendlichen für die Informationsgesellschaft und ihre Bedingungen in Japan und Deutschland : Kurzbericht über ein international vergleichendes Projekt	TA	TAD	TAD	TT	T	TDA	TAD	TADATA	TAD
Krotz, Friedrich (2002): And the Winner is - BMW : James Bond, die Medien und die Märkte	T	TADA	TAD	T AD	TAD	TD	TAD	AD	TAD
Winter, Carsten; Hepp, Andreas; Krotz, Friedrich (2008): Theorien der Kommunikations- und Medienwissenschaft : grundlegende Diskussionen, Forschungsfelder und Theorieentwicklungen	TA	T AD	TTAD	TA	TA	DTAT	TAD	ADA	TADT
Ziemer, Maike (2004): Schule auf den Kopf gestellt : SchülerInnen unterrichten LehrerInnen	TA	TAD	TDA	TATAD	AT	TA	TAD	AD	TA
Kübler, Hans-Dieter (2000): Mediale Universalität : Medientheorie zwischen phänomenologischer Entgrenzung und gegenständlicher Identität	TA	TAD	TA	TAD	TAD	TAD	TADA	ATADA	TATAD
Marsden, Nicola (2002): Vorurteile über virtuelle Welten an Schulen : soziale Stereotypen als Barrieren auf dem Weg in den medialen Lernraum	TA	TADT	TAD	TTADA	TA	TAD	TAD	TA	TADT
Sander, Wolfgang (2000): Neue Medien und Politikunterricht : Probleme, Aufgaben und Chancen am Beispiel der CD-ROM 'Forschen mit GrafStat	TTA	TAD	TTADT	TD	TAAT	TA	TADT	ADA	TADT
(2002): Medienpädagogik heute : eine Diskussionsrunde	TA	ADT	TTAD	TDAT	TAD	TAD	TADATA	AD	TA
(1997): Medienkompetenz im Informationszeitalter	TADA	TADAD	TAD	TAD	TADA	TA	TA	AD	TTA

Anhang B: Scanpfade visualisiert

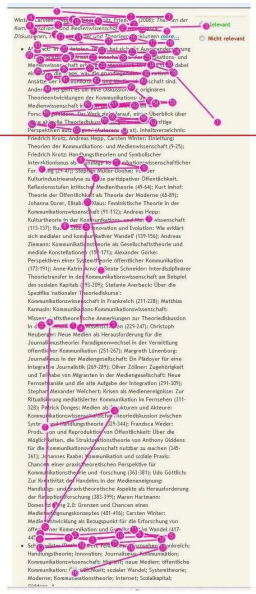
Rec 1

Rec 11



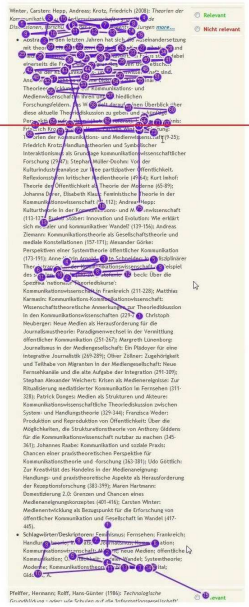
Rec 8

Rec 9



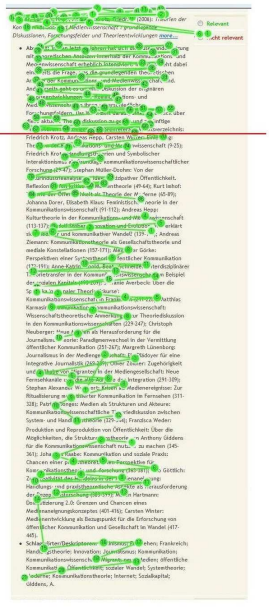
Rec 5

Rec 3



Rec 4

Rec 12



Quellenverzeichnis

Aula, Anne; Majaranta, Päivi; Räihä, Kari-Jouko (2005): Eye-Tracking Reveals the Personal Styles for Search Result Evaluation. In: *Proceeding of INTERACT 2005*, LNCS 3585, S. 1058 –1061, September 2005, Rom, Italien

Barry, Carol L. (1994): User-Defined Relevance Criteria: An Exploratory Study. In: *Journal of the American Society for Information Science*, Vol. 45, Nr. 3, S. 149-159, 1994

Barry, Carol L.; Schamber, Linda (1998): Users' Criteria For Relevance Evaluation: A Cross-Situational Comparison. In: *Information Processing & Management,* Vol. 34, Nr. 2/3, S. 219 – 236, 1998

Bateman, Judy (1998): Changes in Relevance Criteria: A Longitudinal Study. In: *ASIS '98. Proceedings of the 61st ASIS Annual Meeting*, Vol. 35, S. 23-32, 1998

Bente, Gary (2005*):* Erfassung und Analyse des Blickverhaltens, In: Mangold, Roland; Vorderer, Peter; Bente, Gary (Hrsg.): *Lehrbuch der Medienpsychologie*, Göttingen: Hogrefe-Verlag, 2005, S. 297 - 324

BITKOM (2010): Suchmaschinen im Boom.
URL: http://www.bitkom.org/de/presse/66442_65444.aspx, Stand: 7.10.2010

Broder, Andrei (2002): A taxonomy of web search. In: *Newsletter ACM SIGIR Forum*, Vol. 36, Nr. 2, Herbst 2002, S. 3-10

Case, Donald (2007*):* Looking for Information. A Survey of Research on Information Seeking, Needs, and Behavior. 2. Aufl. Bingley: Emerald Group Publishing Limited, 2007

Cosijn, Erica; Ingwersen, Peter (2000*):* Dimensions of relevance. In: *Information Processing and Management*, Vol. 36, Nr. 4, S. 533-550, 2000

Cutrell, Edward; Guan, Zhiwei (2007): What Are You Looking For? An Eye-tracking Study of Information Usage in Web Search. In*: SIGCHI 2007*. *Proceedings of the SIG-CHI conference on Human factors in computing systems,* April/Mai 2007, San José, USA

Ellis, David (2009): Ellis's Model of Information-Seeking Behavior. In: Fisher, Karen; Erdelez, Sanda; McKechnie, Lynne (Hrsg.): *Theories of Information Behavior*. 3. Aufl. Medford: Information Today, Inc., 2009

Enquiro, EyeTools, Did-It (Hrsg.) (2005): Eye Tracking Study.
URL: http://www.enquiroresearch.com/images/eyetracking2-sample.pdf, Stand: Juni 2005

Fisher, Karen; Erdelez, Sanda; McKechnie, Lynne (Hrsg.) (2009): Theories of Information Behavior. 3. Aufl. Medford: Information Today, Inc., 2009

Granka, Laura; Feusner, Matthew; Lorigo, Lori (2008): Eye Monitoring in Online Search, IN: Hammoud, R.I. (Hrsg.): *Passive Eye Monitoring. Signals and Communication Technology*, Berlin: Springer, 2008, S. 347-372

Henrici, Matthias (2010): EyeTracking – Analyseverfahren zur Usability- und Konversionsoptimierung;
URL: http://www.konversionskraft.de/hintergrunde/eyetracking-analyseverfahren-zur-usability-und-konversionsoptimierung.html, Stand: 22 April 2010

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

Janes, Joseph W. (1991): Relevance Judgements and the Incremental Presentation of Document Representations. In: *Information Processing & Management* Vol.27, Nr. 6, S. 629 – 646, 1991

Joachims, Thorsten; Granka, Laura; Pan, Bing (2005): Accurately Interpreting Clickthrough Data as Implicit Feedback. In: *SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval,* August 2005, Bahia, Brasilien

Kelly, Diane; Belkin, Nicholas (2001): Reading Time, Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevance Feedback. In: *SIGIR '01. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, September 2001, New Orleans, USA

Liu, Jingjing; Cole, Michael J.; Liu, Chang; Bierig, Ralf; Gwizdka, Jacek; Belkin, Nicholas J.; Zhang, Jun; Zhang, Xiangmin (2010): Search Behaviors in Different Task Types. In: JCDL'10*. Proceedings of the 10th annual joint conference on Digital libraries*, Juni 2010, Gold Coast, Australien

Lorigo, Lori; Pan, Bing; Hembrooke, Helena; Joachims, Thorsten; Granka, Laura; Gay, Geri (2006): The influence of task and gender on search and evaluation behavior using Google. IN: *Information Processing and Management* , Vol. 42), S. 1123 – 1131, 2006

Lorigo, Lori; Haridasan, Maya; Brynjarsdottir, Hrönn; Xia, Ling; Joachims, Thorsten; Gay, Geri (2008): Eye Tracking and Online Search: Lessons Learned and Challanges Ahead. IN: *Journal of the American Society for Information Science and Technology,* Vol. 59, Nr. 7, S. 1041 – 1052, 2008

Maglaughlin, Kelly L.; Sonnenwald, Diane H. (2002): User Perspectives on Relevance Criteria: A Comparison among Relevant, Partially Relevant, and Not-Relevant Judgements. IN: *Journal of the American Society for Information Science and Technology*, Vol. 53, Nr. 5, S. 327 – 342, 2002

Mayr, Philipp; Mutschke, Peter; Petras, Vivien; Schaer, Philipp; Sure, York (2011): Applying Science Models for Search. In: *ISI 2011. Internationales Symposium für Informationswissenschaft*, Hildesheim, 2011

Moe, Kirsten Kirkegaard; Jensen, Jeanette M., Larsen, Birger (2005): A Qualitative Look at Eye-tracking for Implicit Relevance Feedback. In: *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval*. Roskilde, Dänemark, 2005

Mizzaro, Stefano (1996) How many Relevances in IR? In: *Proceedings of the Workshop `Information Retrieval and Human Computer Interaction'*, GIST Technical Report GR96-2, Glasgow University, S. 57-60, Glasgow, UK, 1996

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

Richter, Tobias (2008): Forschungsmethoden der Medienpsychologie. In: Batinic, Bernard; Appel, Markus (Hrsg.): *Medienpsychologie*, Heidelberg: Springer Medizin Verlag, 2008, S. 3-44

Saito, Hitomi; Terai, Hitoshi; Egusa, Yuka; Takaku, Masao; Miwa, Makiko; Kando, Noriko (2009): How Task Types and User Experiences Affect Information-Seeking Behavior on the Web: Using Eye-tracking and Client-side Search Logs. In: *Understanding the User SIGIR 2009 Workshop*, Boston, USA, 2009

Salojärvi, Jarkko; Kojo, Ilpo; Simola, Jaana; Kaski, Samuel (2003): Can relevance be inferred from eye movements in information retrieval? In: *WSOM 2003. Proceedings of the 4th Workshop on Self-Organizing Maps,* Hibikino, Japan. September 2003, S. 261-266

Saracevic, Tefko (1975): Relevance: A Review of and a Framework fort he Thinking on the Notion in Information Science. In: *Journal of the American Society for Information Science and Technology*, Vol. 26, Nr. 6, S. 321–343, 1975

Saracevic, Tefko (1997): The stratified model of information retrieval interaction: Extension and applications. In: *Proceedings of the American Society for Information Science*, Washington DC, Vol. 34, S. 313-327, 1997

Saracevic, Tefko (2007a): Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. In: *Journal of the American Society for Information Science and Technology*, Vol. 58, Nr. 13, S. 1915-1933, 2007

Saracevic, Tefko (2007b): Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance. In: *Journal of the American Society for Information Science and Technology*, Vol. 58, Nr.13, S. 2126-2144, 2007

Schamber, Linda (1991): Users' criteria for evaluation in multimedia information seeking and use situations. In: *ASIS '91. Proceedings of the 54th Annual Meeting of the American Society for Information Science*, Washington DC, 28, S. 126-133.

Schamber, Linda (1994): Relevance and Information Behavior. IN: *Annual Review of Information Science and Technology (ARIST)*, Vol. 29, S. 3 – 48, 1994

Tang, Rong; Solomon, Paul (2001): Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies. In: *Journal of the American Society for Information Science and Technology,* Vol. 52, Nr. 8, S. 676–685, 2001

Wilson, T.D. (1999a): Exploring models of information behaviour: the ‚uncertainty' project. In: *Information Processing and Management*, Vol. 35, Nr.6, S. 839-849,1999

Wilson, T.D. (1999b).: Models in Information Behaviour Research. In: *Journal of Documentation*, Vol. 55, Nr. 3, S. 249-270, 1999

Wilson, T.D. (2000): Human Information Behavior. In: *Informing Science*, Vol. 3, Nr. 2, S. 49-55, 2000

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig erstellt und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Soweit ich auf fremde Materialien, Texte oder Gedankengänge zurückgegriffen habe, enthalten meine Ausführungen vollständige und eindeutige Verweise auf die Urheber und Quellen.

Alle weiteren Inhalte der vorgelegten Arbeit stammen im urheberrechtlichen Sinn von mir, soweit keine Verweise und Zitate erfolgen.

Mir ist bekannt, dass ein Täuschungsversuch vorliegt, wenn die vorstehende Erklärung sich als unrichtig erweist.

Ort, Datum	Unterschrift

Ausleihebestimmungen

Erklärung

Bitte ankreuzen:

☐ Mit der Ausleihe der Bachelorarbeit bin ich einverstanden.

☐ Mit der Ausleihe bin ich nicht einverstanden, die Bachelorarbeit ist gesperrt.

Studentin / Student _____ _____

Datum Unterschrift

Betreuende Professorin /

Betreuender Professor _____ _____

Datum Unterschrift