Untersuchung von Relevanzeigenschaften in einem kontrollierten Eyetracking-Experiment[1]

Stefanie Reichert[2], Philipp Mayr

Abstract

In diesem Artikel wird ein Eyetracking-Experiment beschrieben, bei dem untersucht wurde, wann und auf Basis welcher Informationen Relevanzentscheidungen bei der themenbezogenen Dokumentenbewertung fallen und welche Faktoren auf die Relevanzentscheidung einwirken. Nach einer kurzen Einführung werden relevante Studien aufgeführt, in denen Eyetracking als Untersuchungsmethode für Interaktionsverhalten mit Ergebnislisten (Information Seeking Behavior) verwendet wurde. Nutzerverhalten wird hierbei vor allem durch unterschiedliche Aufgaben-Typen, dargestellte Informationen und durch das Ranking eines Ergebnisses beeinflusst. Durch Eyetracking-Untersuchungen lassen sich Nutzer außerdem in verschiedene Klassen von Bewertungs- und Lesetypen einordnen. Diese Informationen können als implizites Feedback genutzt werden, um so die Suche zu personalisieren und um die Relevanz von Suchergebnissen ohne aktives Zutun des Users zu erhöhen. In einem explorativen Eyetracking-Experiment mit 12 Studenten der Hochschule Darmstadt werden anhand der Länge der Gesamtbewertung, Anzahl der Fixationen, Anzahl der besuchten Metadatenelemente und Länge des Scanpfades zwei typische Bewertungstypen identifiziert. Das Metadatenfeld Abstract wird im Experiment zuverlässig als wichtigste Dokumenteigenschaft für die Zuordnung von Relevanz ermittelt.

The article describes an eyetracking experiment which examines relevance judgements within the context of subject-related document assessments. We analyze in the study on what information the judgements of our test persons are based on and which document specific properties influence the relevance decisions. In the state of the art the authors present relevant studies that use eyetracking methodology as a research method to investigate information seeking behaviour models. The three factors that particularly influence user behaviour are: different task types, search results presentation, and document ranking. Furthermore, the results of these eyetracking studies help us, to classify users into typical evaluation and reading types. This information can then be used as implicit feedback to personalize the search. Relevance of search results could thus be improved without any further involvement by the users. In an exploratory eyetracking experiment with twelve students from the University of Applied Sciences in Darmstadt, we were able to identify two typical evaluation types, based on total length of the evaluation, number of fixations, number of visited metadata elements and length of the scan path. This experiment shows that the metadata field abstract is clearly the most important document property to assign topical relevance to scientific articles.

Keywords: user study, evaluation, information retrieval, search results, empirical study, eyetracking

---

[1] To appear in IWP: DOI 10.1515/iwp-2012-0029

[2] Email: reichert.stefanie@arcor.de

1 Einführung

Relevanz ist eines der Kern-Konzepte der Informationswissenschaft (vgl. Saracevic 2007a). Alle Aktionen eines Nutzers laufen im Grunde genommen darauf hinaus, ein Informationsbedürfnis mit relevanten Ergebnissen über wenige Interaktionen zu befriedigen. Die Erwartungen der Nutzer von Informationssystemen sind dabei i.d.R. sehr hoch. Nutzer erwarten trotz einfachster Suchanfragen hochrelevanteste Ergebnisse mit einem möglichst geringen Anteil an irrelevanten Treffern. Auf der anderen Seite erschwert die zunehmende Überflutung mit Informationen das Finden von relevanten Informationen im Alltag aber auch bei der Suche nach wissenschaftlichen 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 Effektivität 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 Internet-Suchmaschinen im beruflichen, wissenschaftlichen sowie privaten Bereich ist jeder Internetnutzer damit konfrontiert, viele Male am Tag Relevanzurteile im Kontext der Dokumentensuche fällen zu müssen. Durchschnittlich 50 Suchanfragen stellen deutsche Internetnutzer pro Woche (BITKOM 2010).

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 (Saracevic 2007a). 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 (siehe dazu Mutschke et al. 2011). Schließlich ist es der Nutzer selbst, der die Entscheidung trifft welche ihm präsentierten Informationsobjekte in einer bestimmten Situation relevant sind oder nicht. Die Disziplin innerhalb der Informationswissenschaft, zu der diese Art der Forschung zugeordnet werden kann, wird u.a. Interaktive Information Retrieval (IIR) (Ingwersen 1992) oder auch Information Behaviour genannt (Fisher et al. 2005). Die IIR- bzw. Information Behaviour-Forschung stellt, statt des technischen Systems beim klassischen IR, den Nutzer bzw. die „menschlichen Aspekte“ (human aspects) bei der Suche in den Vordergrund. Untersuchungsgegenstand ist i.d.R. die aufgabenbezogene Interaktion mit einem Informationssystem und die subjektiven Wahrnehmungen während des Suchprozesses. Bei der Teildisziplin Information Seeking Behavior (ISB) wird der Fokus auf Verhalten, Motivation und Vorgehen des Benutzers bei der Recherche nach Informationen verengt.

Im vorliegenden Artikel werden die zentralen Ergebnisse eines Eyetracking-Experiments im Rahmen einer Master-Thesis im Sommer 2011 vorgestellt (Reichert 2011). Dafür wurde in einer explorativen Eyetracking-Studie das Verhalten von Studenten bei Relevanzentscheidungen bei einer klassischen IR-Evaluation untersucht. Zu diesem Zweck bewerteten Studenten anhand eines vorgegebenen Retrieval-Topics eine Auswahl von Dokumenten relevant oder nicht relevant. Ziel war es herauszufinden, wann und auf welcher Grundlage die individuellen Relevanzentscheidungen fallen, ob es bestimmte Muster gibt, die zu Relevanzentscheidungen führen, durch welche Faktoren die Entscheidungen möglicherweise beeinflusst werden und ob es Hinweise darauf gibt, dass anhand von charakteristischen Verhaltensweisen Relevanz abgeleitet werden kann.

2 State of the Art

Das Verfahren Eyetracking wird seit ca. 2003 genutzt, um Suchverhalten im Internet zu analysieren. Die Methode ermöglicht es, mit Sensoren und Kameras Blickbewegungen einer Person in Echtzeit zu verfolgen, aufzuzeichnen und zu untersuchen (Abbildung 1). Neben Blickrichtung und -intensität 3können auch Mimik und Gestik der Nutzer sowie Klicken, Scrollen, Texteingaben und Kommentare („Think Aloud“) dokumentiert werden. Eine Kerngröße sind die Fixationen, bei denen das Auge eine gewisse Zeit lang auf einem Punkt des Bildschirms ruht, der besondere Aufmerksamkeit erregt hat. Eyetracking-Studien im Kontext der Suche beziehen sich häufig auf die Analyse des Interaktionsverhaltens von Nutzern mit Ergebnislisten von Websuchmaschinen wie Google (Search Engine Result Page, SERP). Sie analysieren Fragestellungen wie z.B. gesucht wird, welche Bereiche vom User angeschaut und geklickt werden und letztlich auch wie eine Relevanzentscheidung fällt.

Im Jahr 2005 beschrieben die Marketingfirmen Enquiro und Did-it in Zusammenarbeit mit der Firma Eyetools das bekannte F-Schema oder auch „Golden Triangle“, welches die Bereiche, die Suchmaschinennutzer auf den Ergebnisseiten bevorzugt beachten, in einer Heatmap darstellte. Demnach schenkten die User den ersten drei Ergebnissen sehr viel Aufmerksamkeit, den nachfolgenden Ergebnissen dagegen kaum. Da sich die Ergebnisliste von Google in der Zwischenzeit durch die Anzeige von Bildern, Videos, Karten usw. gewandelt hat, ist das F-Schema aber zumindest für die Ergebnisseiten von Suchmaschinen nicht mehr ohne Einschränkungen gültig. Nutzerverhalten ist außerdem weitaus komplexer und wird durch viel mehr Faktoren beeinflusst, als das F-Schema berücksichtigt.

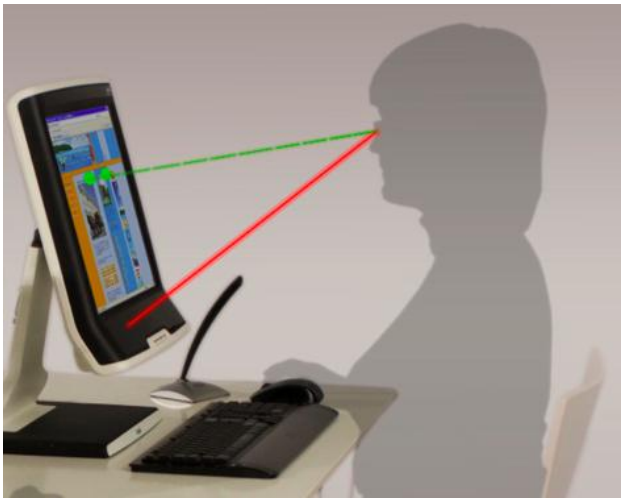

Abbildung 1: Eyetracking schematisch (Quelle: http://www.konversionskraft.de/hintergrunde/eyetracking-analyseverfahren-zur-usability-und-konversionsoptimierung.html)

Im folgenden Abschnitt werden einige zentrale Eyetracking-Studien und Erkenntnisse vorgestellt. Dabei geht es insbesondere um verschiedene Faktoren, die das Nutzerverhalten beeinflussen und den Versuch Nutzertypen zu klassifizieren.

Einfluss des Aufgabentyps

2008 untersuchten Papaeconomou, Zijlema und Ingwersen mittels Eyetracking und anschließenden Interviews ob es einen Zusammenhang zwischen Lernstilen (Global and Sequential Learners) und deren Relevanzbewertungen von Webseiten gibt. Dazu wurden unter anderem „relevance hot spots“

untersucht, also Bereiche der Webseiten, denen die 15 Testpersonen besonders viel Aufmerksamkeit schenkten. Dabei kamen die Autoren zu der Erkenntnis, dass es weniger die Lernstile waren, die Einfluss auf Relevanzentscheidungen hatten, als die Art der Aufgaben.

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

1. Navigatorische Aufgaben, bei denen es das Ziel ist, eine bestimmte Webseite oder URL zu finden.
2. Informatorische Aufgaben, bei der eine bestimmte Information gefunden werden soll, die sich aber auch auf mehreren Webseiten befinden kann.
3. Transaktionale Aufgaben, 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 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 unterschieden sich die Scanpfade von Männern und Frauen. Männliche Testpersonen neigten demnach dazu Suchergebnisse eher linear zu betrachten (siehe „Lesetypen“). 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 im Hinblick auf Information Seeking Behavior Strategien im Web. Die beiden Aufgaben in der Studie waren „Bericht schreiben“ und „Ausflug planen“. Dabei griffen sie nicht nur auf 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 Suchverhalten und task type erkennen konnten, diese aber aufgrund der geringen Zahl von Teilnehmern nicht verallgemeinern wollten. Die Autoren konnten zu einem gewissen Grad Facetten von Aufgaben, z.B. die Komplexität, anhand des Suchverhaltens vorhersagen. Dieses Wissen wollen sie zukünftig nutzen um die Personalisierung der Suche durch implizites Feedback noch genauer und konkret anwendbar zu machen.

Einfluss der dargestellten Information

Cutrell und Guan (2007) haben den Einfluss der Informationen in „Snippets“ auf den Ergebnisseiten untersucht. 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 machen. 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 wie der Nutzer am Bildschirm agiert und was er liest, bevor er tatsächlich ein Dokument auswählt. Sie verglichen unter anderem die durchschnittliche Zeit, die User damit verbringen 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 in den Studien eher geringen Einfluss auf das Nutzerverhalten hatten, 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 beiden ersten Abstracts. 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. Dem Ranking der Suchmaschine wird großes Vertrauen entgegen gebracht. Die Autoren bezeichneten dies als „trust bias“ (Joachims et al. 2005, S. 154). 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).

Klassifizierung von Nutzertypen

Aula et al. (2005) identifizierten 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 (vgl. Scanpfade in Abbildung 2). 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.

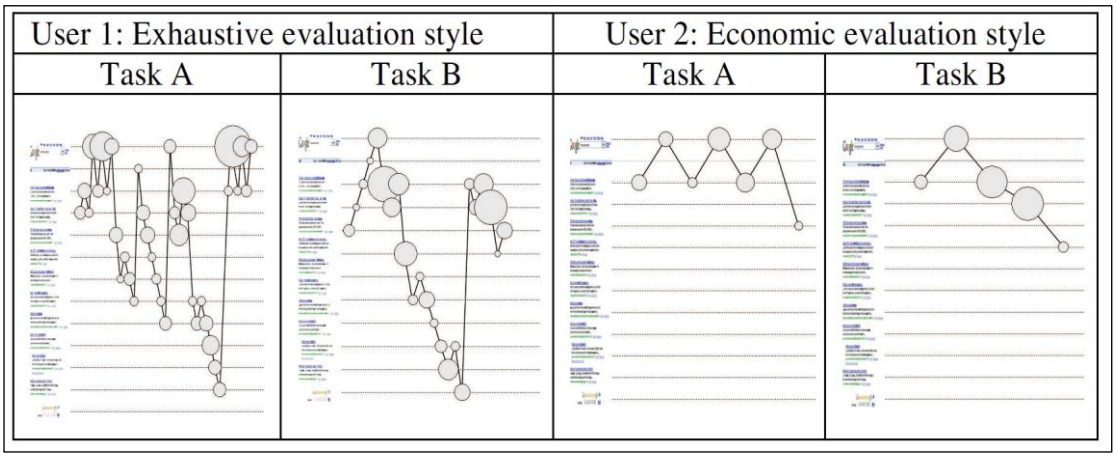

Abbildung 2: Scanpfade der Bewertungsytpen nach Aula et. al 2005 (S. 1060)

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 scannen.
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 im Leseverhalten („skips and jumps“) beobachten.

Nutzerverhalten als implizites Relevance Feedback

Implizite Feedbacktechniken 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 (1), sorgfältiges Lesen (2) und Zurückspringen (3), identifizierten sie die Zeit, die ein User damit verbringt eingehend 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 Eyetracking-Experiment

Für das in der Masterarbeit durchgeführte Experiment wurden zwölf Studierende der Informationswissenschaft an der Hochschule Darmstadt über eine Eyetracking-Anlage bei der binären Relevanzentscheidung in einem Retrievalexperiment beobachtet (vgl. Mutschke et al. 2011). Den Testpersonen wurde dafür eine ungeordnete Liste von 34 Dokumenten zu dem Topic „Neue Medien im Unterricht" angezeigt, die über eine Weboberfläche relevant oder nicht relevant bewertet werden sollten. Die Arbeitsaufgabe lautete: „Finde Dokumente, die über Chancen und Risiken des Einsatzes neuer und moderner Medien in der Schule berichten". Da das Thema über ein Dropdown-Menü direkt auf der Test-Webseite ausgewählt werden konnte, mussten die Testpersonen keine eigene Suchanfrage formulieren (Abbildung 3). Es gab keine Vorgaben welche Kriterien zur Relevanzbewertung herangezogen werden sollten. Alle Tester konnten den Begriff „relevant" frei interpretieren und nach eigenen Kriterien bewerten. Die Aufgabe und die Liste der Dokumente waren bei jedem Teilnehmer gleich, nur die Reihenfolge der Dokumente änderte sich jedes Mal beliebig. Da Nutzer in ihrer Entscheidung sehr stark vom Rang eines Dokumentes in einer geordneten Ergebnisliste beeinflusst werden („trust bias", vgl. Joachims et al. 2005, S. 154), wurden jeder Testperson die Dokumente in zufälliger Reihenfolge angezeigt. Die Nutzer wurden über dieses Verfahren vor dem Test informiert. Im Anschluss an die Bewertung fand ein informelles Feedback-Gespräch statt, in dem die Nutzer unter anderem eine formalisierte Frage über die Wichtigkeit der Dokumentbereiche für die Relevanzbewertung beantworten konnten. Dabei konnten 0-10 Punkte vergeben werden, 10 für den Dokumentbereich, der für den aktuellen Test die größte Bedeutung für die Relevanzentscheidung hatte. 0 für den Bereich, der für die Relevanzentscheidung keine Rolle gespielt hat.

Fragestellung

Ziel der Masterarbeit war es herauszufinden, wann und auf welcher Grundlage individuelle Relevanzentscheidungen fallen, ob es bestimmte Muster gibt, die zu Relevanzentscheidungen führen, durch welche Faktoren die Entscheidungen möglicherweise beeinflusst werden und ob es Hinweise darauf gibt, dass anhand von charakteristischen Verhaltensweisen Relevanz abgeleitet werden kann.

Folgender Versuchsaufbau wurde innerhalb der Masterarbeit umgesetzt.

Hardware und Software

Bei der verwendeten Eye Monitoring Hardware handelte 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. Die Testperson wird dadurch nicht von aufwendigen Apparaturen abgelenkt und hat während der Bewertung einen gewissen Bewegungsspielraum (Abbildung 1).

Mithilfe der dazugehörigen Software „Tobii Studio" wurde das Versuchsdesign realisiert. Der Test bestand aus einer Einführungsseite mit Bearbeitungshinweisen und dem Link, welches die Bewertungsoberfläche im Internet Explorer aufrief. Die Software ermöglichte außerdem die statistische Analyse und visuelle Darstellung der aufgezeichneten Blickbewegungsdaten in Form von Scanpfaden (Abbildung 5).

Testpersonen

Die zwölf Teilnehmer der Studie waren Studenten der Informationswissenschaft an der Hochschule Darmstadt und hatten daher Erfahrung im Umgang mit Suchergebnislisten und Relevanzentscheidungen. Keiner von ihnen hatte jedoch spezielles Vorwissen in der Topic-Domäne Erziehungswissenschaft („Neue Medien im Unterricht"). Acht Personen studierten bereits im Masterstudiengang, vier Teilnehmer kamen aus dem Bachelorstudiengang. Die Master-Studierenden waren Teilnehmer des Seminars „Information Seeking Behavior" von Dr. Philipp Mayr im Sommersemester 2011. Die Geschlechterverteilung war ausgeglichen mit sechs weiblichen und sechs männlichen Teilnehmern. Die Nutzer waren im Alter von 22 und 28, ein Nutzer war 42 Jahre alt. Ein Experte führte darüber hinaus eine Bewertung ohne Eyetracking durch. Seine Ergebnisse wurden als Richtwert für die Relevanz verwendet.

Die Master-Studierenden wurden im Rahmen des Seminars „Information Seeking Behavior" persönlich angesprochen, die Studenten im Bachelor-Studiengang wurden durch eine Rundmail im Fachbereich über das Experiment informiert. Im Anschluss an die Bewertungen erhielten die Tester eine kurze Einführung in die Eyetracking-Anlage als Dankeschön für die Teilnahme.

Testumgebung

Als Relevanzbewertungs-Tool diente eine Webseite, die ursprünglich im Rahmen des DFG-geförderten Projektes IRM „Value-Added Services for Information Retrieval" des Leibniz Institutes für Sozialwissenschaften (GESIS) realisiert wurde (vgl. Mayr et al. 2011, Mutschke et al. 2011). [3]

Die Bestandteile der Bewertungsoberfläche sind in Abbildung 3 dargestellt: auf der Webseite können die Tester über ein Dropdown-Menü das Topic auswählen. Danach werden dem Nutzer eine Reihe von Dokumentrepräsentationen angezeigt, die aus Autor(en), Publikationsjahr, Titel, Abstract und Deskriptoren bestehen. Der Bewertende hat die Möglichkeit eine binäre Relevanzentscheidung zu treffen und diese via Klick auf einen Radiobutton (relevant / nicht relevant) auszuführen.

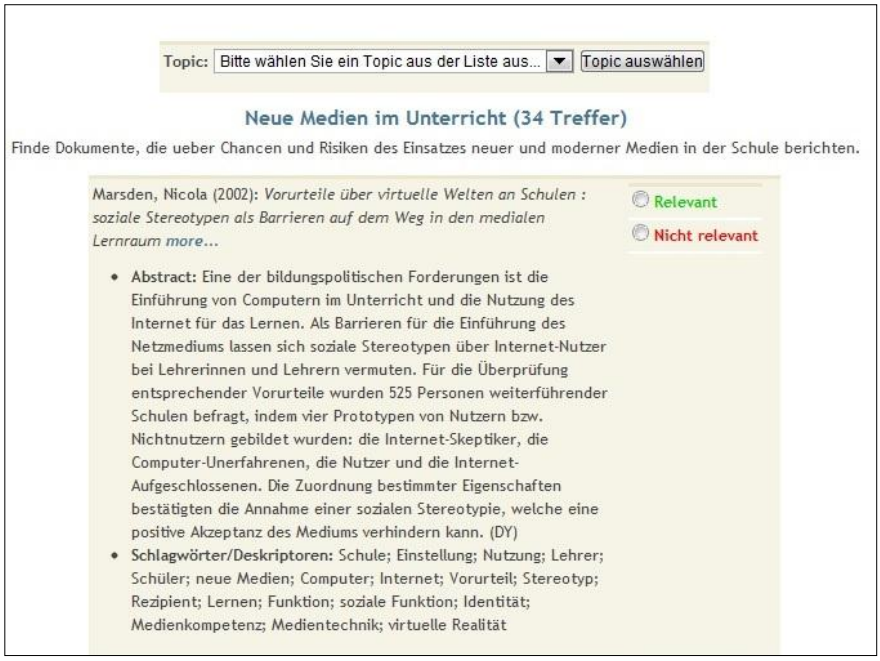

Abbildung 3: Ausschnitt der Bewertungsoberfläche mit einem Beispieldokument (inkl. Titel, Abstract und Schlagwort-Information)

[3] http://www.gesis.org/en/research/external-funding-projects/archive/irm/

Auswertung der Eyetracking-Daten

Bei vier von 12 Teilnehmern traten Ungenauigkeiten in den Eyetracking-Daten auf, so dass diese nicht verwendet werden konnten. Alle nachfolgenden Angaben, die sich speziell auf Eyetracking beziehen, gelten daher nur für acht Teilnehmer. Die Aufzeichnungen der Bewertungs-Sessions wurden über Video-Exporte zum Teil am heimischen PC ausgewertet, detaillierte statistische Analysen unter Verwendung der Software Tobii Studio konnten nur im Uselab am Campus Dieburg der Hochschule Darmstadt erfolgen. Letztere Analysen waren sehr zeitaufwendig und konnten bei jedem Teilnehmer nur für 17 der 34 bewerteten Dokumente einer Session durchgeführt werden. Um bestimmte Bereiche einer Szene genauer zu untersuchen, müssen Interessensbereiche, sogenannte Areas of Interest (AOI) bestimmt werden. Für diese definierten Sektionen, auch Lookzones genannt, können dann statistische Daten wie Anzahl und Dauer der Fixationen ermittelt werden.

Daten, die mithilfe der Software Tobii Studio erhoben werden konnten:

- Die Dokumentrepräsentationen wurden in jeweils 3 Areas of Interest aufgeteilt (vgl. Abbildung 4): T für Titel/Autor/Jahr; A für Abstract und D für Deskriptoren. Anschließend wurde die Anzahl der Fixationen für jede AOI ermittelt.
- Scanpfade für jede Bewertungssequenz wurden visualisiert (vgl. Abbildung 5)

Daten, die über Video-Exporte gewonnen werden konnten:

- Bewertungsdauern für Einzeldokumente und Gesamtbewertungsdauern; 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 eines Users bildet die Gesamtlänge.
- Erfassung der Relevanzbewertungen (relevant / nicht relevant), Vergleich mit Expertenbewertung
- Bestimmung von "Absprungmarken" als jene Stellen, an denen der Nutzer das Dokument verlässt um seine Bewertung durchzuführen
- Beschreibung des Blickverlaufs jeder Bewertungssequenz über Zeichenketten (z.B. TTDA: User liest zwei mal den Titel, dann die Deskriptoren, dann das Abstract)
- Bestimmung der Länge der Scanpfade jeder Bewertung (z.B. TTDA = 4)
- Auszählung wie viele AOI ein User während der Session insgesamt besucht

Zusätzlich wurden die Post-Search-Interviews analysiert und die darin enthaltene Frage zur Wichtigkeit der Datenelemente ausgewertet.

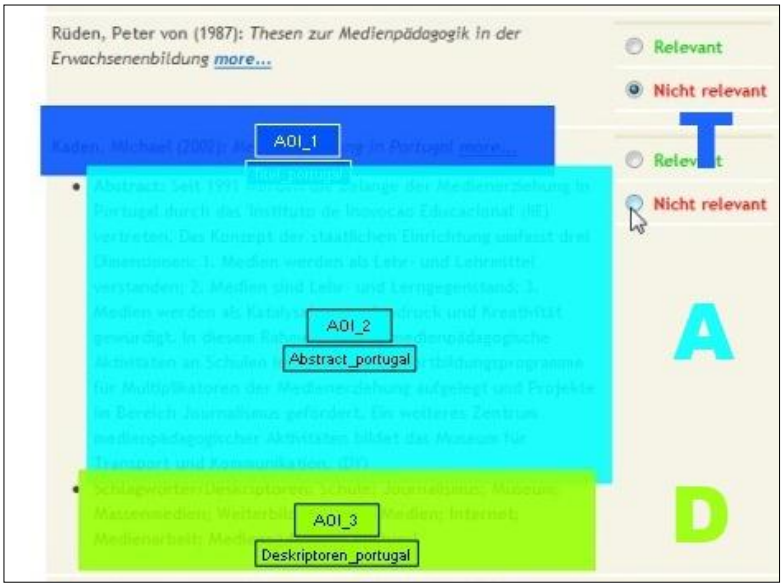

Abbildung 4: Definition von drei Lookzones T, A und D für jedes ausgewertete Dokument

4 Ergebnisse

Wie oben beschrieben, konnten Aula et al. (2005) die Teilnehmer ihrer Eyetracking-Studie in zwei Klassen von Bewertungstypen einordnen (vgl. Abb. 2). Zum einen die besonders gründlichen Bewerter, die exhaustive evaluators. Zum anderen die economic evaluators, die sich mit wenigen Informationen zufriedengeben bevor sie eine Entscheidung treffen.

Die Teilnehmer dieser Studie lassen sich nach der Auswertung der Eyetracking-Daten ebenfalls in die beiden Gruppen einsortieren. Die Zugehörigkeit zu einer der beiden Klassen wurde durch folgende Auswertungskriterien festgestellt: Gesamtlänge der Bewertung, durchschnittliche Anzahl der Fixationen pro Dokumentbewertung, Anzahl der besuchten AOIs insgesamt sowie durchschnittliche Länge des Scanpfades pro Nutzer über alle Bewertungen (Tab. 1).

	Economic Evaluators	**Exhaustive Evaluators**
Dauer Gesamtbewertung	Schneller (12,67 sek.)	Langsamer (19,78 sek.)
Anzahl Fixationen pro Dokument	Wenig (62,45)	Viele (100,7)
Anzahl besuchter AOI insgesamt	Wenig (89,25)	Viele (120,5)
Durchschnittliche Länge d. Scanpfade	Kürzer (2,63)	Länger (3,55)
Übereinstimmung mit Expertenmeinung	Eher höher (18,75)	Eher geringer (16,5)
Anzahl relevant bewerteter Dokumente	13,5	20

Tabelle 1: Vergleich zwischen Economic und Exhaustive Evaluators

Abbildung 5: Beispiel für visualisierte Scanpfade. Links: Economic Evaluator, Rechts: Exhaustive Evaluator

Bewertungsdauer

Da jeder linke Mausklick eine Bewertungsentscheidung darstellt, bildet die Zeit von einem linken Mausklick zum nächsten eine Bewertungssequenz. Die Dauer von der ersten Fixation im ersten

Dokument bis zur letzten Relevanzentscheidung bildet die Gesamtlänge der Bewertung. 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. 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. Die Lesedauer eines einzelnen Dokuments liefert allerdings keine Rückschlüsse auf Relevanz. 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.

Bewertungsgeschwindigkeit

Die Lesedauern für einzelne Dokumente nehmen im Laufe der Bewertung bei acht von zwölf Teilnehmern ab, was für nachlassende Konzentration und/oder Motivation sprechen könnte. Bei zwei Teilnehmern blieben die Bewertungsdauern konstant, zwei Tester wurden nach schnellem Beginn langsamer. Die Liste war mit 34 Dokumenten aber auch vergleichsweise lang. Auf SERPs werden in der Regel nur 3-5 Abstracts betrachtet (vgl. Lorigo et al. 2008). Da die Dokumente jedes Mal neu gemischt und in willkürlicher Reihenfolge angezeigt wurden, konnte gewährleistet werden, dass sich die längeren Abstracts nicht immer an derselben Stelle befinden.

Übereinstimmung

Nur 4 von 34 Dokumenten wurden von allen gleich bewertet. Inklusive Expertenmeinung waren es sogar nur 2 Dokumente, bei denen sich alle einig waren. Das spricht dafür, dass die Nutzer sehr unterschiedliche Relevanzkriterien angelegt haben bzw. ähnliche Kriterien mit unterschiedlichen Gewichtungen versehen haben. Hier zeigt sich deutlich, dass Subjektivität ein zentrales Merkmal der Relevanz ist.

Während in der Gesamtheit wenig Konsens herrscht, sind die Übereinstimmungen im Einzelnen umso größer. Saracevic (2007b) schätzte die Übereinstimmung zweier Bewertender auf ca. 30 %, was in dem aktuellen Experiment zehn gleich bewerteten Dokumenten entsprechen würde. Mit durchschnittlich 17,8 Übereinstimmungen zwischen dem Experten und einzelnem Tester liegen die Werte in diesem Experiment deutlich darüber.

Einflussfaktoren auf die Relevanzbewertung

Eines der beiden wichtigsten Relevanzkriterien nach Xu und Chen (2006) ist die Neuheit eines Dokuments. Im Feedback-Gespräch haben einige Nutzer angegeben, dass sie ältere Dokumente aus den 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. Das Dokumentalter hatte aber dennoch keinen messbaren Einfluss auf die Relevanzentscheidung.

Bei der Dokumentlänge hingegen konnte ein Einfluss festgestellt werden. Dokumente mit sehr langen Abstracts wurden im Test tendenziell eher nicht-relevant bewertet. 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 längerer Texte ohne weitere Formatierungen anstrengend und abschreckend sei.

Wichtigkeit der Metadatenelemente für das Erkennen von Relevanz

Um herauszufinden welche Metadatenelemente die meisten Hinweise auf Relevanz liefern, wurden drei Kriterien untersucht:

1. Punktevergabe im Feedback-Interview: Das Abstract wurde mit Abstand als wichtigstes Metadatenelement bewertet. Dicht gefolgt vom Titel, der mit 3 mal 10 Punkten am häufigsten die Höchstwertung erhalten hat (Tabelle 2). Der Titel wurde jedoch von einigen zurückhaltender bewertet, da er in manchen Fällen allein nicht ausreichend genug den tatsächlichen Inhalt beschreibt.

2. "Absprungmarken" (Tabelle 3): Um herauszufinden, zu welchem Zeitpunkt die Relevanzentscheidungen fallen, wurde der zuletzt vor der Entscheidung betrachtete Bereich im Dokument isoliert geprüft. Das Abstract ist insgesamt die häufigste Absprungmarke, für fünf User waren die Deskriptoren letzter Fixationspunkt. 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 (Tabelle 4): Die Lookzones „Titel“ und „Abstract“ wurden beide fast gleich oft besucht, wobei das Abstract in diesem Ranking knapp vorne liegt. Deskriptoren werden deutlich weniger häufig gelesen. 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. 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.

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). Allerdings konnte anhand der Scanpfade festgestellt werden, dass Abstracts sehr häufig nicht komplett gelesen werden. Lange Texte schienen eher abzuschrecken.

	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

Tabelle 2: Wichtigkeit der Datenelemente. Die Tester konnten 0-10 Punkte vergeben (10 = am wichtigsten)

	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

Tabelle 3: Auszählung der „Absprungmarken“

AOI	Anzahl Besuche
T	372
A	373
D	204

Tabelle 4: Visit Count: Anzahl der Besuche in den AOIs insgesamt über alle Teilnehmer

5 Zusammenfassung und Diskussion

Grundsätzlich ist zu bemerken, dass die Ergebnisse aufgrund der sehr geringen Fallzahl nicht verallgemeinerbar sind, sondern nur in dem aktuellen Zusammenhang gültig sind. Die Studie wurde angesichts des erheblichen Arbeitsaufwandes bei der Auswertung mit nur zwölf Testpersonen durchgeführt, Eyetracking-Daten konnten aufgrund von Ungenauigkeiten bei der Kalibrierung nur für acht Tester erhoben werden. Die Test-Teilnehmer waren darüber hinaus alle Studenten der Informationswissenschaft, was möglicherweise Einfluss auf Verhalten oder die Beantwortung der Frage nach den wichtigsten Datenelementen hatte (gefühlte Verpflichtung z.B. Deskriptoren höher zu bewerten). Außerdem ist die Wahrnehmung nicht nur auf Fixationen beschränkt, sondern findet auch im peripheren Sichtfeld statt. Absolute Zahlen sind daher nur als ungefähre Anhaltspunkte zu verstehen.

In der Studie konnten die beiden Bewertungsgruppen nach Aula et al. von 2005 bestätigt werden. Die economic evaluators, im Test 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, trotz insgesamt weniger als relevant bewerteten Dokumenten, höher als die der exhaustive evaluators. Der gründliche Bewertungstyp hingegen, im Test größtenteils weiblich, braucht insgesamt länger, liest mehr und genauer, hat längere Scanpfade und bewertet auch deutlich mehr Dokumente relevant als nicht relevant. Dennoch ist die Übereinstimmung mit der Expertenmeinung geringer.

Die Aussagen der Nutzer im Feedback-Gespräch stützen diese Resultate. Die langsameren exhaustive evaluators betonten, 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 nicht-relevant. Wichtiges Kriterium für diesen Typ der Bewerter war die hohe thematische (topical relevance) Übereinstimmung der Dokumente. Zu allgemeine Werke oder Sammelbände hätten für die economic evaluators zu viel Arbeitsaufwand für zu wenig Information bedeutet.

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. 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. 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 in Fisher et al. 2005). Entscheidungen werden außerdem beeinflusst durch Persönlichkeitstyp, Arbeitstyp, Erinnerungen, Befinden, Abneigungen, Stress, Desinteresse, Ablenkung, innere Haltung (Geisteshaltung), Tagesform und vieles mehr.

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 stand, schreckten eher ab. Sehr häufig wurde nur die obere Hälfte der Abstracts oder weniger gelesen. Sogar kurze Abstracts wurden häufig nicht komplett gelesen. Das Abstract ist aber nach Auswertung dreier verschiedener Parameter das wichtigste Metadatenelement für die Ableitung von Relevanz. Für Dokumentare und Autoren 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 könnten 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 zeigten 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 zu überblicken.

Literatur

1. 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
2. BITKOM (2010): Suchmaschinen im Boom. URL: http://www.bitkom.org/de/presse/66442_65444.aspx, Stand: 7.10.2010
3. Broder, Andrei (2002): A taxonomy of web search. In: Newsletter ACM SIGIR Forum, Vol. 36, Nr. 2, Herbst 2002, S. 3-10
4. 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 SIGCHI conference on Human factors in computing systems, April/Mai 2007, San José, USA
5. Enquiro, EyeTools, Did-It (Hrsg.) (2005): Eye Tracking Study. URL: http://www.enquiroresearch.com/images/eyetracking2-sample.pdf, Stand: Juni 2005
6. Fisher, Karen; Erdelez, Sanda; McKechnie, Lynne (Hrsg.) (2009): Theories of Information Behavior. Medford: Information Today, Inc., 2009
7. 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
8. Ingwersen, Peter (1992): Information Retrieval Interaction. London: Taylor Graham, 1992.
9. 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
10. 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
11. 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
12. 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
13. 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
14. 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
15. Mutschke, Peter; Mayr, Philipp; Schaer, Philipp; Sure, York (2011): Science models as value-added services for scholarly information systems. In: Scientometrics, Vol. 89, Nr.1, S. 349-364

16. 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

17. Reichert, Stefanie (2011): Messung von Relevanz in einem kontrollierten Information Seeking Experiment. Masterarbeit, Hochschule Darmstadt

18. 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

19. 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

20. 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

21. 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

22. Xu, Yunjie; Chen, Zhiwei: Relevance Judgement: What Do Information Users Consider Beyond Topicality? In: Journal of the American Society for Information Science and Technology, 57 (7):961 – 973, 2006