Bezugsrahmen für die Evaluation von Information Retrieval Systemen mit Visualisierungskomponenten

von Sonja Hierl

1 Zunehmende Entwicklung von Suchmaschinen mit visuellen Komponenten

Aktuell ist auf dem Retrieval- und Suchmaschinenmarkt ein steigendes Angebot von Systemen zu beobachten, die eine visuelle Darstellung von Suchergebnissen umsetzen oder anstreben. Zum einen handelt es sich dabei um Anbieter freier und kostenpflichtig zugänglicher Web- und Desktop-Suchmaschinen (14, 30, 10) und Provider von Fachinformationen (7, 18). Zum anderen berichten diverse Publikationen (16, 24, 25, 35, 19, 28, 32) von Forschungsprojekten, die eine Entwicklung unterschiedlicher Retrieval Applikationen mit visuellen Komponenten behandeln.

Ziel derartiger Information Retrieval Systeme (IRS) ist die Optimierung der Unterstützung des Nutzers während des Suchprozesses. Dies geschieht beispielsweise durch die visuelle Repräsentation von Relationen zwischen selektierten Ergebnisobjekten (14), der Darstellung von Themenclustern, die eine Aussage über die inhaltlichen Zusammenhänge der Treffer ermöglichen (10) oder der Unterstützung des Nutzers in der Anfrageformulierung oder Reformulierung von Suchanfragen durch visuell aufbereitete Begriffswolken (23).

Wenn auch die Wirksamkeit solcher Ansätze durchaus einleuchtend erscheinen, muss diese dennoch empirisch nachgewiesen werden, wie Cugini et al. feststellen: "One of the lessons of our experience is that no matter how much intuitive appeal a given interface might have, without some systematic testing, its real value remains unknown. Especially in the field of visualization, it is all too common for technical wizardry to be unaccompanied by any gain in efficiency." (6).

2 State-of-the-Art der Evaluation von Information Retrieval Systemen mit Visualisierungskomponenten

2.1 Aktuelle Vorgehensweisen

Während sich bereits vor längerer Zeit ein Konsens über die Vorgehensweise zur Evaluation gängiger IRS gebildet hat (33), weisen bisherige Studien zur Evaluation der oben beschriebenen Systemtypen ein sehr uneinheitliches und häufig wenig fundiertes bzw. methodisch schwach abgestütztes Vorgehen auf. So stellt Vaughan fest, dass die Entwicklung valider Evaluationstechniken derzeit nicht Schritt halten kann mit der rapiden Geschwindigkeit der Entwicklung neuartiger Suchapplikationen (31).

Im Rahmen von Forschungsprojekten werden zwar in der Regel Studien zur Evaluation und Qualitätssicherung der entwickelten Systeme durchgeführt, diese verfolgen jedoch zumeist unterschiedliche Zielsetzungen und weisen folglich ein sehr unterschiedliches Untersuchungsdesign auf.

Die Analyse diverser Evaluationsstudien (31, 9, 3, 4, 13, 17, 12, 19, 24, 35) ergibt, dass zur Evaluation von IRS mit Visualisierungskomponenten (VK) im wesentlichen Methoden aus den Bereichen der Retrievalperformanzmessung und der Gebrauchstauglichkeitsmessung angesetzt werden. So können neben den von Plaisant (22) aufgeführten, häufig vertretenen Ansätzen

- Kontrollierte Experimente zum Vergleich von Designelementen
- Usability-Evaluation einer Anwendung
- Kontrollierte Experimente zum Vergleich zweier oder mehrerer Anwendungen sowie
- Case Studies in realistischen Szenarien (22)

vor allem Herangehensweisen identifiziert werden, bei denen anhand klassischer Retrievaleffektivitäts-maße wie Recall und Precision die Retrievaleffektivität von IRS mit VK gemessen wird.

In den meisten Fällen werden entweder Methoden aus einem der beiden Bereiche eingesetzt, oder es erfolgt eine Kombination von Methoden, die zwar gemeinsam die Messung der Qualität des jeweiligen Systems bezwecken, hierbei jedoch in der Regel keine Interdependenzen zwischen Usability-Evaluation und Retrievaleffektivitäts-Evaluation berücksichtigen (siehe hierzu die detaillierten Ausführungen in Kapitel 3). Vielmehr findet eine losgelöste Betrachtung der Methoden statt, Auswirkungen der Interaktion des Nutzers mit dem System und deren Einflüsse auf Retrievalperformanz sowie Einflüsse der Usability und der daraus entstehenden Ansprüche an die Gestaltung der Nutzeroberfläche werden nicht in ausreichendem Maße berücksichtigt.

2.2 Studien zur allgemeinen Wirksamkeit von IRS mit Visualisierungskomponenten

Es liegen derzeit nur sehr wenige Studien vor, die die Wirksamkeit der in IRS integrierten Visualisierungen im Allgemeinen evaluieren und mit den erzielbaren Ergebnissen konventioneller Suchmaschinen und IRS mit Textausgabe vergleichen (1).

Chen und Yu führten 2000 erstmals eine umfassende Studie durch, die eine Meta-Analyse aktueller, empirischer Evaluationen von visuellen Informationssystemen zum Ziel hat und identifizierten dabei folgende Problemstellung: Ohne ein einheitliches, systematisch basiertes Evaluationskonzept lassen sich keine kontrollierten und methodisch breit abgestützten Tests durchführen, die anschließend eine Vergleichbarkeit der Ergebnisse über verschiedene Evaluationsstudien hinweg ermöglichen. Durch das Fehlen einer gemeinsamen Grundlage und nachhaltiger Untersuchungen können folglich kaum repräsentative Schlüsse und empirisch abgestützte Aussagen zur generellen Wirksamkeit von Visualisierungen in der Ergebnisrepräsentation von IRS getroffen werden. So blieben nach einem systematischen Auswahlverfahren von anfänglich 27 identifizierten Publikationen, die die Evaluation eines IRS mit VK diskutieren, lediglich sechs Studien für den eigentlichen Meta-Vergleich übrig, da die anderen Beiträge zu große Unterschiede aufwiesen oder so uneinheitliche Ergebnisdaten beinhalteten, dass keine direkte Gegenüberstellung sinnvoll durchführbar war (5).

Shneiderman und Plaisant untersuchten 2006 auf Grundlage dieser Metastudie die gängige Evaluationspraxis für IRS mit VK und kamen zum Schluss, dass sich aktuell noch kein einheitliches Evaluationsdesign durchgesetzt hat (27). Die Autoren fordern einen Trend weg von experimentellen Labor-Usabilitystudien hin zu ethnographischen Studien in der üblichen Arbeitsumgebung der Probanden, bei denen Unterbrechungen, Arbeitsplatz, Hilfeleistungen und der soziale Austausch wie gewohnt vorliegen. Weiterhin betonen sie die Relevanz der Durchführung von Studien, bei denen Probanden reale Aufgaben ihrer täglichen Arbeit durchführen und nicht vorgegebene Testaufgaben.

3 Herausforderungen bei Schaffung eines Bezugsrahmens für die Evaluation von IRS mit Visualisierungskomponenten

3.1 Interdependenzen zwischen Usability-Evaluation und Retrievaleffektivitäts-Evaluation

Die in 2.1. angesprochene fehlende Integration gewählter Ansätze zur Evaluation der Gebrauchstauglichkeit und der Retrievaleffektivität erscheint unbedingt erforderlich, da sich durch die Kombination von IRS und einer visuellen Oberfläche Interdependenzen ergeben, die sich auf die Qualität des Systems auswirken, wie in folgender *Abbildung 1* verdeutlicht:

So haben beispielsweise Interaktionen von Nutzern auf der visuellen Oberfläche einen Einfluss auf die Retrievalfunktionalität eines Systems, was wiederum Auswirkungen auf die Usability des IRS hat.

Weiterhin lassen sich in Abhängigkeit von Retrievalfunktionalität und Gebrauchstauglichkeit Ansprüche an die Gestaltung der Nutzeroberfläche identifizieren. Durch die Kombination von Retrieval- und VK ergeben sich folglich Interdependenzen, die sich auf die Qualität des Systems auswirken, wie im Folgenden ausgeführt.

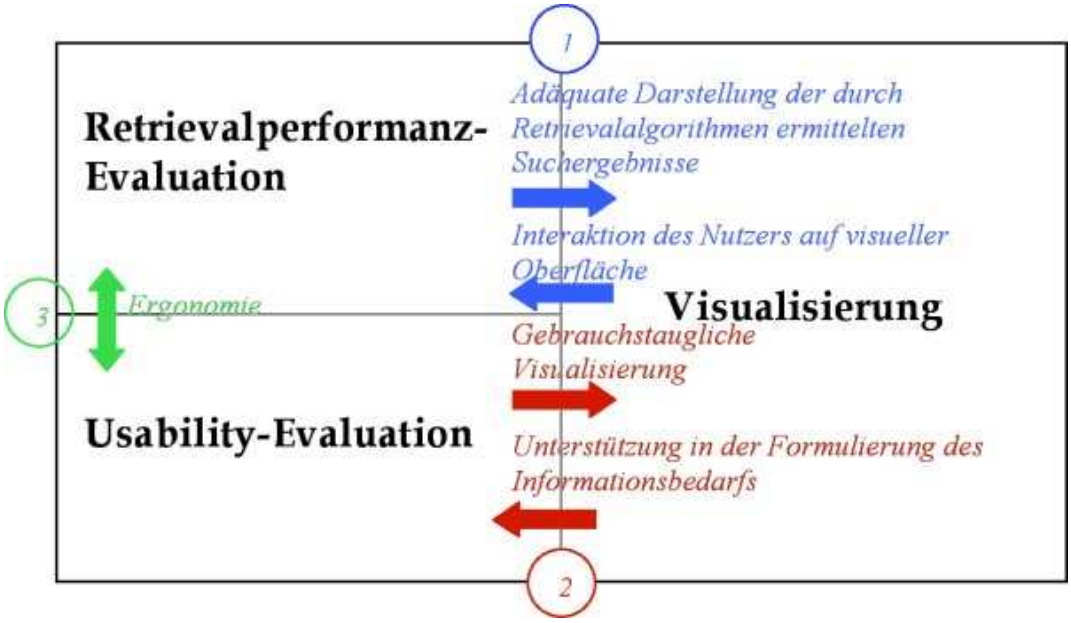

Abbildung 1: Interdependenzen Retrivalperformanz-Visualisierung-Usability-Evaluation

1 Retrievaleffektivitäts-Evaluation und Visualisierung

Bei der Durchführung einer Evaluation muss berücksichtigt werden, dass die Visualisierung der Ergebnisse durch die zum Einsatz gebrachten Retrievalalgorithmen beeinflusst werden. Repräsentiert ein System beispielsweise Ergebnismengen durch eine nicht-hierarchische Darstellung, so kann ein durch die Algorithmen berechnetes Ranking mit einer in der Relevanz abnehmenden Reihenfolge nicht adäquat dargestellt werden.

Andererseits haben Interaktionen seitens der Nutzer mit den visuell aufbereiteten Ergebnismengen einen Einfluss auf die im System

implementierten Retrieval-Algorithmen: In Abhängigkeit der Interaktionsmöglichkeiten des Nutzers muss das IRS die entsprechenden Veränderungen der Suchanfrage und der Ergebnismenge verarbeiten um anschließend die erneut ermittelten und verfeinerten Ergebnisse visuell zu präsentieren.

2 Visualisierung und Usability-Evaluation

Zielführend ist der Einsatz intuitiver Visualisierungen, die den Nutzer während des Suchprozesses unterstützen - beispielsweise in der adäquaten Formulierung seines Informationsbedarfs. Hierfür müssen einerseits die Bedeutungen der verwendeten Visualisierungen dem Nutzer bekannt und einfach verständlich nachvollziehbar sein. Andererseits muss der Erfolg einer Visualisierung in einer Evaluation des IRS bewertet werden. Durch die verwendeten Methoden in der Usability-Evaluation sollte also nicht nur grundsätzlich festgestellt werden, ob ein IRS mit VK für den Nutzer allgemein gebrauchstauglich gestaltet ist, sondern auch, ob die gewählten Visualisierungen dem Zweck der Definition des Informationsbedürfnisses dienlich sind. Es ergibt sich folglich eine Wechselwirkung zwischen Ursache und Wirkung, die nur schwer ermittelt werden kann.

Synnestvedt und Chen formulieren die Herausforderung folgendermaßen: "The problem with usability testing based on information retrieval tasks (...) is that the testing reports success or failure of a task but not why the user failed. Evaluation [is] needed of tasks requiring users to compare, associate, distinguish, rank, cluster or categorize." (29)

3 Usability-Evaluation und Retrievaleffektivitäts-Evaluation

Eine weitere Wechselwirkung ergibt sich durch den Aspekt der Ergonomie: Konzepte zur Visualisierung im Rahmen des IR mögen zwar überzeugend sein und in der Theorie die Retrievaleffektivität erhöhen; wenn diese Ansätze jedoch nicht in ergonomischer Weise umgesetzt werden, unterstützen die entsprechenden Systeme den Nutzer nicht bei der Befriedigung seiner Informationsbedürfnisse.

Soll zum Beispiel bei einem IRS die Art des Informationsbedürfnisses des Nutzers berücksichtigt werden, um einen darauf optimierten Retrievalalgorithmus anzuwenden, hat dies zur Folge, dass das System aus Nutzersicht komplexer und dadurch schwerer anzuwenden wird. Während für versierte Anwender die Auswahl verschiedener Einschränkungen und Systemeinstellungen von Vorteil ist, könnte sie sich für ungeübte Nutzer aus Sicht der Gebrauchstauglichkeit überflüssig oder gar nachteilig gestalten. Es müssen folglich Wege gefunden werden, diese aus der Sicht der Retrievalperformanz sinnvolle Ergänzung auf ergonomische Weise in die Systemoberfläche zu integrieren um Einbussen in der Gebrauchstauglichkeit zu vermeiden. Auch bezüglich dieses Aspekts ist die bereits oben erwähnte Messung von Ursache und Wirkung eine große Herausforderung.

3.2 Weitere Problemfelder

Beim Entwurf eines einheitlichen Evaluationsdesigns für IRS mit integrierter VK treten diverse weitere Herausforderungen auf, die sowohl durch den Umstand der eingesetzten Visualisierungen bedingt sind, als auch allgemein bei vergleichenden Evaluationen auftreten. Im Folgenden werden zwei dieser Problemfelder andiskutiert:

1 Verschiedenartigkeit von Visualisierungen

Die in Informationssystemen eingesetzten Visualisierungen unterscheiden sich massiv, sowohl hinsichtlich ihrer Art und Interaktionsmöglichkeiten, als auch bezüglich des verfolgten Ziels. Koshman leitet daraus ab, dass bei visuellen Systemen ein unterschiedliches Evaluationsdesign zugrunde gelegt werden muss, das jeweils die Besonderheiten der eingesetzten Visualisierung berücksichtigt (16). Im Gegensatz zu textbasierten Ausgaben von Suchmaschinenergebnissen können wesentlich mehr Informationen über die Ergebnismenge dargestellt werden (so z. B. Relationen zwischen Suchergebnissen, Mengenverhältnisse, Ergebniskategorien u. v. m). Folglich ergeben sich je nach eingesetzter Visualisierung unterschiedliche Funktionen und Interaktionsschritte, die durch den Nutzer ausgelöst werden können.

2 Unterschiede in Bekanntheitsgrad und Intuitivität

Der Erfahrungsgrad eines Nutzers hat einen unmittelbaren Einfluss auf seine Interaktion mit einem Informationssystem. Durch die weite Verbreitung und den großen Bekanntheitsgrad von klassischen Suchmaschinen mit textbasierter Ergebnisausgabe wie *Google*, *Yahoo* oder *MSN* besteht folglich die Gefahr, dass Ergebnisse von vergleichenden Evaluationen verfälscht werden, wenn der Nutzer hinsichtlich der Nutzung des textbasierten IRS über einen hohen Grad an Systemerfahrung verfügt und somit auf sein implizites Wissen und Erfahrungen zurückgreifen kann, jedoch das neue visuelle System nicht kennt und es somit nicht mit dem gleichen Erfahrungsschatz und Wissen nutzen kann.

3.3 Auswirkungen auf das Evaluationsdesign

Zusammenfassend lässt sich aus den Herausforderungen die Notwendigkeit zur Verwendung eines integrierten Methodenmix aus Ansätzen der Usabilitymessung und der Retrievaleffektivitäts-Evaluation unter Berücksichtigung oben dargestellter Interdependenzen ableiten. Die Messung des gleichen Aspekts unter Verwendung unterschiedlicher Methoden zur Erlangung sowohl quantitativer als auch qualitativer Aussagen sollte angestrebt werden. Durch den Vergleich der voraussichtlich zum Teil widersprüchlichen Ergebnisse lassen sich die Aussagen ganzheitlich interpretieren. So sollte beispielsweise ein von einem Nutzer ausgefüllter Fragebogen zur subjektiv empfundenen Qualität der getesteten Systeme jeweils mit den Ergebnissen von auf Recall und Precision basierenden ermittelten Effektivitätswerten verglichen und interpretiert werden.

Hinsichtlich der Verschiedenartigkeit von Visualisierungen gilt es für die Evaluation einen Bezugsrahmen anzusetzen, das zwar eine gemeinsame Grundlage hinsichtlich der Aspekte bildet, die bei allen IRS mit VK gleich sind, gleichzeitig jedoch die Möglichkeit zur freien Wahl passender Methoden bietet. Hierfür dient die im anschließenden Kapitel vorgeschlagene Morphologie, die eine Grundlage zur Auswahl geeigneter Methoden darstellt.

Der Aspekt des unterschiedlichen Bekanntheitsgrades neuartiger Systeme mit Visualisierungen sollte weiterhin ausgeglichen werden durch den kombinierten Einsatz von Feldstudien, bei denen das Alltagsverhalten von Nutzern berücksichtigt wird und eine Vertrautheit im Umgang mit visuellen Komponenten entstehen kann.

Die Herausforderung bei der Planung einer Evaluation besteht folglich darin, anhand eines integrierten Designs einen Bezugsrahmen zu entwickeln, das nicht nur Methoden kombiniert, sondern auch den oben genannten neu auftretenden Fragestellungen begegnet.

Weiterhin sollte eine Grundlage geschaffen werden, auf der künftig ein Vergleich durchgeführter Evaluationen möglich ist, um Unterschiede sowie Ähnlichkeiten in den Resultaten

4 Morphologischer Kasten für die Auswahl geeigneter Methoden zur Evaluation von IRS mit Visualisierungskomponenten

4.1 Übersicht geeigneter Evaluationsmethoden

Als Grundlage für die Auswahl eines Methoden-Mix wurden im Rahmen der Literaturanalyse rund 40 Methoden aus den Bereichen der Retrievaleffektivitäts- und der Gebrauchstauglichkeits-Messung zusammengetragen und beschrieben[1]:

BEZEICHNUNG	ERLÄUTERUNG	ABK
Recall	Vollständigkeit des Retrievalergebnisses, die auf Basis der gefundenen EOs berechnet wird mit einem Ergebniswert zwischen 0 und 1. (1 wird vergeben, wenn alle relevanten EOs einer Testkollektion gefunden wurden, 0 wenn keines der EOs durch das IRS ermittelt wurde.)	R
Ability to retrieve top ranked pages	Berechnung des Recall in Bezug auf intellektuelles Relevanzurteil von Probanden, dem jeweils die ersten x Ergebnisobjekte einer Recherche unterzogen wurden.	
Relative Recall@n	Relative Vollständigkeit des Retrievalergebnisses bei den ersten n Treffern	R@n
Precision	Genauigkeit des Retrievalergebnisses, die auf Basis der gefundenen Ergebnisobjekte berechnet wird mit einem Ergebniswert zwischen 0 und 1 (1 wird vergeben, wenn alle gefundenen EOs relevant sind, 0 wenn keines der ausgegebenen EO's relevant ist.)	P
Precision@n	Precisionwert der ersten n Treffer auf die Suchanfrage	P@n
Mean Average Precision	Gewogener Mittelwert der Precision über die ersten n Treffer hinweg	MAP@n
e-Mass nach Van Rijksbergen	Gewichtung von Recall und Precision unter Einbezug einer Konstante â	E-Mass
First Retrieved Document Rank	Ranking des ersten Dokuments, das Information enthielt, die der Nutzer als so relevant einstuft, dass keine weitere Suche mehr erforderlich ist.	FRDR
Jewel Measure	Rank des Dokuments, das am besten auf die Suchanfrage des Nutzers passt	JM
System- oder Ergebnisstabilität	Stabilitätsmass der Retrievalalgorithmen und somit der Ergebnistreffer (welche Abweichungen ergeben sich bei den Treffermengen bei wiederholter gleicher Suchanfrage unter gleichen Bedingungen)	
Suchqualität	Ableitung der Nutzerzufriedenheit aus der Interaktion des Nutzers mit Ergebnisobjekten	
Algorithmen-Analyse	Algorithmen, die einem IRS zugrunde liegen und Aussagen über dieses zulassen. Auf erster Ebene wird dichrom das Vorhandensein ermittelt, liegen Algorithmen vor, werden diese detaillierter analysiert	
Heuristische Expertenevaluation	Überprüfung der Systemkonformität anhand festgelegter Heuristiken durch Experten, die die Rolle typischer Nutzer einnehmen und Verstösse gegen die Heuristiken identifizieren	
Expertenevaluation mit Guidelines	Überprüfung des Erfüllungsgrades von Leitfäden und Guidelines aus Sicht der Softwareergonomie	
Walkthrough-methoden	Identifikation von logischen Brüchen oder Fehlfunktionen durch das Durchspielen klassischer Nutzungsszenarien durch Experten, die die Rolle typischer Nutzer einnehmen	
Lautes Denken	Erhebung eines Verbalprotokolls der Problemlösungsstrategien von Probanden während der Taskdurchführung	
Tagebuchstudien	(systematische) schriftliche Aufzeichnung der Probanden von Erfahrungen, die sich bei der Systemnutzung ergeben.	
Beobachtung	Beobachtung von Probanden über Kontrollraum oder per Videoaufnahme während der Taskdurchführung.	
Aufmerksamkeits-Analyse mit Eye-Tracking	Aufnahme der Augenbewegungen von Probanden während der Taskdurchführung über befestigte Apparatur	
Aufmerksamkeits-Analyse mit Video-Capturing	Automatische, softwaregestützte Aufnahme der Aktivitäten auf dem Bildschirm der Probanden während der Taskdurchführung	
Monitoring bzw. Data-Logging	Automatische, softwaregestützte Aufnahme von Interaktionsschritten der Probanden während der Taskdurchführung	
Nutzerbefragung durch Fragebogen	Befragung von Probanden unmittelbar nach Taskdurchführung oder Befragung von regelmässigen Systemnutzern	
Nutzerbefragung durch Fokusgruppen und Interviews	Systematische und strukturierte Befragung von Probanden und Systemnutzern bezüglich ihrer Einschätzung und des Umgangs mit dem System	
Taskanalyse	Analyse der erzielten Ergebnisse sowie der Aktivitäten zur Durchführung vereinbarter Aufgabenstellung durch Probanden	
(ethnographische) Feldstudien	Durchführung von Studien unterschiedlicher Art (beispielsweise Taskanalysen) in der gewohnten Arbeitsumgebung der Probanden unter Einbezug ihrer täglich anfallenden Fragestellungen ohne Vorgabe von Zeitlimiten, Tasks, Testkollektionen o. ä.	
Coaching Methode	Bereitstellung von Hilfeleistungen für die Taskdurchführung der Probanden durch einen Testleiter	
Bereitstellung von Supportfunktionen	Bereitstellung von Hilfestellungen, wie sie in der gewohnten Arbeitsumgebung üblicherweise zur Verfügung stehen	
Langzeitstudien	Durchführung von Studien unterschiedlichster Art über eine längere Zeit hinweg, in der die jeweiligen Ergebnisse, Vorgehensweisen, Abweichungen und Veränderungen systematisch festgehalten werden und somit der Lernfortschritt bzw. Verhaltensweisen in der Systemnutzung identifizierbar werden	

Tabelle 1: Kurzcharakteristik von zur Evaluation von IRS mit VK geeigneten Methoden

4.2 Spannungsfeld bei der Kombination geeigneter Methoden

Bei der Kombination von Methoden zur Evaluation von IRS mit VK befinden sich die Instrumente im Spannungsfeld zwischen Laborstudien einerseits, die durch eine künstlich geschaffene Umgebung sehr präzise, aber unter Umständen auch leicht verfälschte Ergebnisse aufweisen

können und Feldstudien andererseits (22). Bei letzteren werden den Probanden für eine Evaluation möglichst authentische und reale Arbeitsbedingungen geboten, wobei die Störfaktoren nicht behoben werden und Ursachen für die Artung der Ergebnisse aus diesem Grund nicht immer identifizierbar sind.

Als weitere Dimension im Spannungsfeld sind zum einen die Objektivität der Ergebnisse und zum anderen ein hoher Grad an Nutzerbeteiligung zu berücksichtigen. Die höhere Objektivität auf der einen Seite wird erlangt durch in Laborkontexten durchgeführte Studien, bei denen nach Möglichkeit alle zu untersuchenden Variablen von den Störfaktoren isoliert werden. Stehen hingegen Aspekte im Vordergrund, bei denen die reale und möglichst natürliche Interaktion zwischen Nutzer und System untersucht werden, sollte eine verstärkte Nutzerbeteiligung (beispielsweise in Feldstudien) angestrebt werden, bei denen zwar weniger objektive, dafür jedoch der realen Nutzung eher entsprechende Ergebnisse erzielt werden.

Die Zusammenstellung eines integrierten Methoden-Mix für die Evaluation eines IRS mit VK sollte dieses Spannungsfeld berücksichtigen. Die Einordnung der charakterisierten Methoden in ein Portfolio mit den vier erwähnten Spannungspolen erfolgt in Anlehnung an die Bewertungen dieser Methoden in der einschlägigen Literatur (11, 26) sowie entsprechend der Analysen in den vorangegangenen Kapiteln.

Hierbei werden die Methoden der Usability-Evaluation und der Retrievaleffektivitäts-Evaluation jeweils getrennt in das Portfolio eingetragen, um eine bessere Übersichtlichkeit zu erzielen.

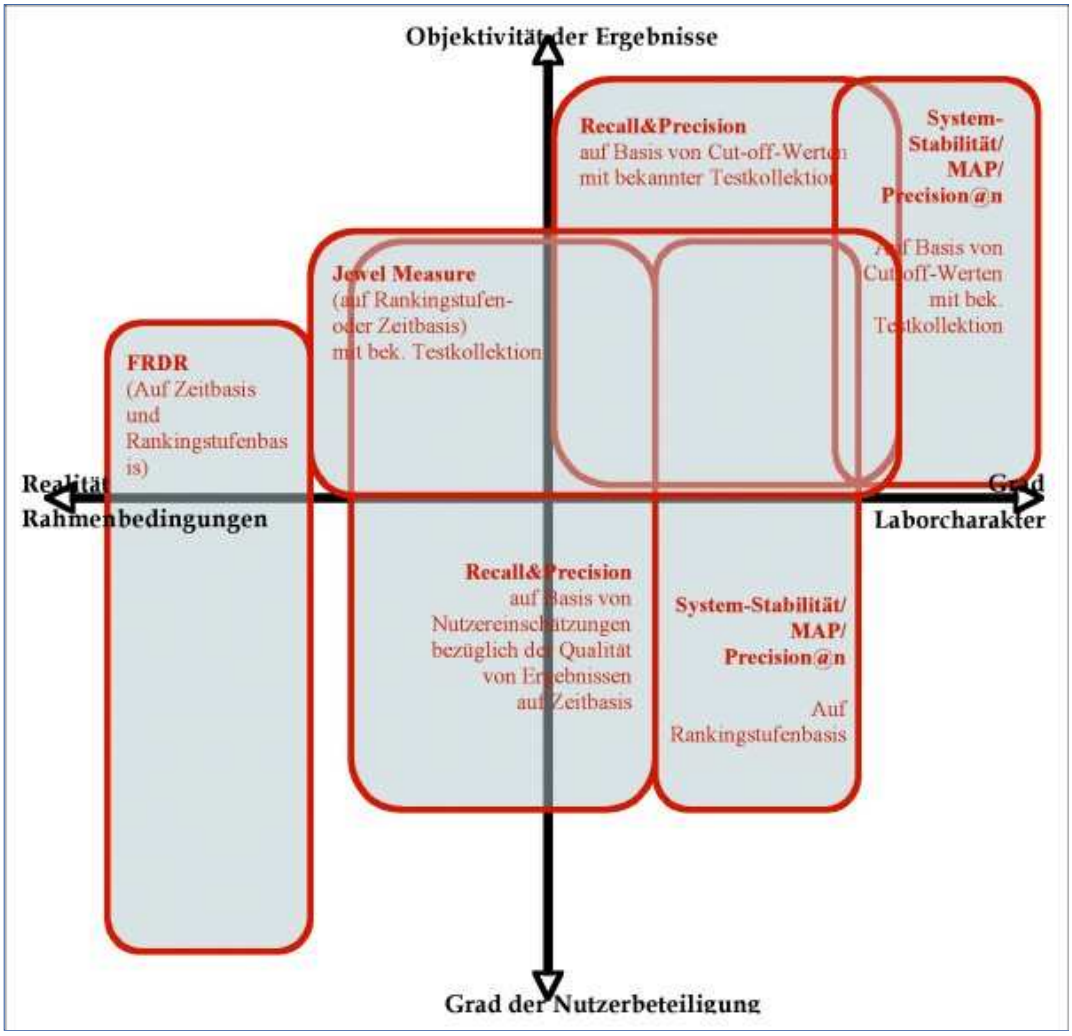

Abbildung 2: Portfolioklassifikation von Retrievaleffektivitätsmessmethoden

Aus der Abbildung wird ersichtlich, dass bei Methoden zur Messung der Retrievaleffektivität unter realen Rahmenbedingungen mit zunehmender Angleichung an die tatsächlichen Arbeitsbedingungen eines Anwenders die Nutzerbeteiligung steigt, gleichzeitig jedoch auch die Objektivität der Ergebnisse sinkt. Im Gegenzug bringen Methoden mit starkem Laborcharakter weitaus objektivere Ergebnisse hervor, die Nutzerbeteiligung sinkt jedoch bis hin zum vollständigen Ausschluss des Anwenders aus den Evaluationen, wodurch Nutzerfeedback, Relevanzurteile und Einschätzungen von Probanden nicht berücksichtigt werden. Gleichermaßen gestaltet sich die Situation bei den Methoden der Gebrauchstauglichkeits-Evaluation, weshalb in beiden Bereichen die Ausgewogenheit der eingesetzten Methoden essenziell ist für eine möglichst hohe Aussagekraft der zu erzielenden Resultate.

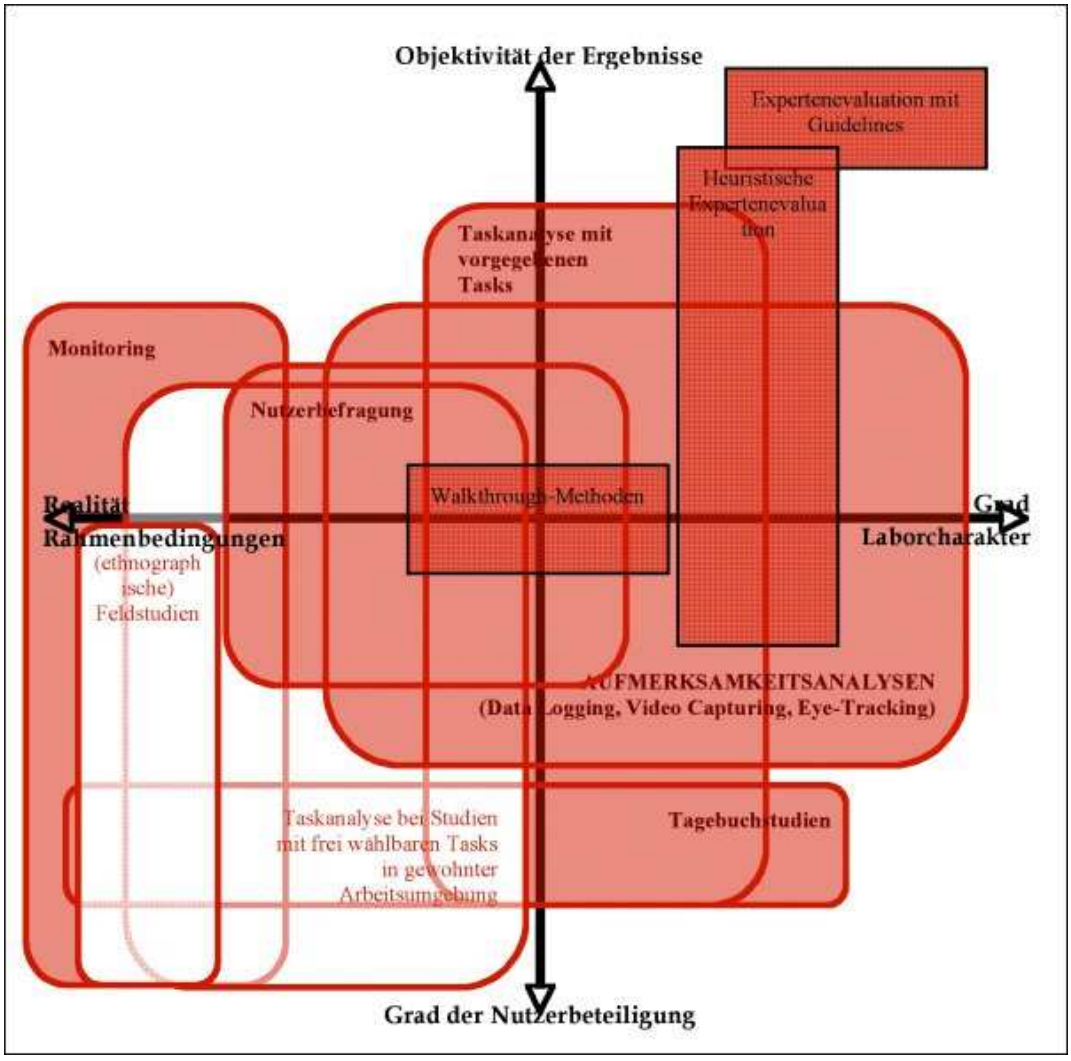

Abbildung 3: Portfolioklassifikation von Gebrauchstauglichkeitsmessmethoden

Bei den in Abbildung 3 dargestellten Methoden wird unterschieden zwischen analytischen Verfahren, empirischen Verfahren, sowie feldstudienähnlichen Verfahren. Die Differenzierung nach den drei Bereichen erfolgt anhand folgender leichten Farbcodierung:

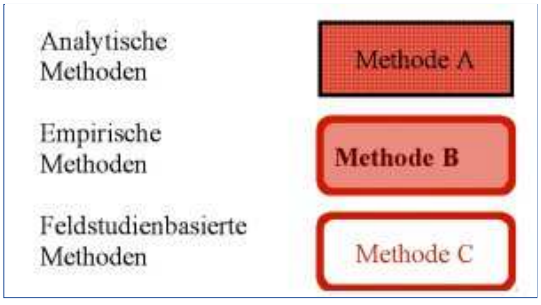

Abbildung 4: Legende der Portfolioklassifikation in Abbildung 3

Anhand der oben dargestellten Klassifikation von Methoden entsprechend der Faktoren

- Laborcharakter
- Realität der Rahmenbedingungen
- Objektivität der Resultate und
- Nutzerbeteiligung bei der Bewertung

soll und kann keine absolute und trennscharfe Einteilung erfolgen, da je nach Anwendungskontext und Art der Umsetzung einer Methode die Zuteilung und Ausprägung der oben genannten Größen durchaus schwanken kann. Vielmehr soll anhand der Portfolios die Tendenz aufgezeigt werden, in welchem Bereich sich eine Methode in etwa bewegt und welche Aspekte bei ihrem Einsatz berücksichtigt werden und welche nicht.

4.3 Morphologischer Kasten für die Klassifikation von Evaluationsmethoden

Für die Klassifikation von geeigneten Methoden und Instrumenten, die den Zielsetzungen einer Evaluation entsprechen, wurde in Anlehnung an Mussgnug und Stowasser (21) ein morphologischer Kasten entwickelt, anhand dessen sich unterschiedliche Ansätze zwischen Usability- und Retrievaleffektivitäts-Evaluation einordnen lassen und der gleichzeitig dem oben aufgezeigtem Spannungsfeld Rechnung trägt (Vergleiche *Tabelle 1*).

DIMENSION	**SKALA bzw. Ausprägung**			
Zielsetzung				
Zielsetzung und Art der Evaluation	Vergleichend	Funktionalität	Leistung (Algorithmen)	Design
Rahmenbedingungen				
Kosten	Gering		Mittel	Hoch
Trainingsaufwand	Gering		Mittel	Hoch
Menge erforderlicher Probanden	Keine		Wenige	Viele
Erforderliche Apparaturen bzw. Software	Auflistung (m) = Muss, (k) = Kann			
Durchführung				
Zeitlicher Aufwand	Gering		Mittel	Hoch
Untersuchungsort	Feld		Kontrolliertes Feld	Labor
Auswertungsdimensionen				
Ergebnisdimension	Qualitativ		Quantitativ	
Bezugsdimension	Subjektiv		Objektiv	
Untersuchungsart	Analytisch	Experimentell	Beobachtend	Fragend

Tabelle 2: Morphologischer Kasten für die Klassifikation von Evaluationsmethoden für IRS mit Visualisierungskomponenten

Auf Grundlage dieses morphologischen Kastens wurde eine Morphologie erstellt mit den identifizierten Methoden und Evaluationsinstrumenten (http://www.informationswissenschaft.ch/index.php?id=299).

4.4 Ergebnisse einer exemplarischen vergleichenden Evaluation

Auf den bisherigen Erkenntnissen aufbauend wurde im Rahmen einer Fallstudie der vorgeschlagene Ansatz in der Empirie exemplarisch überprüft. Die Evaluation erhebt selbstverständlich keinen Anspruch darauf, repräsentative Ergebnisse zu liefern, sondern ist als eine erste empirische Validierung des vorgeschlagenen Bezugsrahmens für die Evaluation zu verstehen.

Es wurde eine vergleichende Evaluation der Suchmaschinen *Yahoo* mit konventioneller Listenausgabe und *Grokker* mit einer visuellen Ergebnisrepräsentation mit fünf Probanden vorgenommen, bei der unter anderem folgende Methoden zum Tragen kamen:

- Kontrolliertes Experiment (Usability-Test) mit vorgegebenen Aufgabenstellungen (5 Probanden)
- "Lautes Denken" (5 Probanden)
- Screencapturing mit Auswertung (5 Probanden)
- Fragebogen (5 Probanden)
- Tagebuchstudie über 2 Wochen (ein Proband)
- Retrievaleffektivitäts-Evaluation zur Erhebung u. a. folgender Maße: Relative Recall@n, Precision@n, Jewel Measure, First Retrieved Document Rank (17).

Die Auswertung ergibt einerseits, dass sich das Suchverhalten sowie die subjektive Einschätzung des Systems *Grokker* durch den Probanden der Tagebuchstudie bereits über die kurze Zeit von zwei Wochen sehr stark veränderten. Während zu Beginn der Studie, bei der der Proband täglich

mindestens 3-5 Recherchen in beiden Systemen durchführte und dokumentierte, große Vorbehalte dem visuellen System gegenüber deutlich wurden, bevorzugte er gegen Ende der Studie eindeutig *Grokker* und erzielte bei dort getätigten Recherchen bessere Resultate. Dieses Ergebnis ließ sich sowohl durch die persönlichen Einschätzungen in den Tagebucheinträgen und im abschließenden Fragebogen feststellen, als auch durch die in einem Usabilitytest erzielten Suchresultate.

Bei der Auswertung des kontrollierten Experiments kann festgestellt werden, dass sich die in einem Fragebogen von Probanden geäußerten Bewertungen nicht immer decken mit den anderweitig erhobenen Kennzahlen. Beispielsweise gaben alle Probanden an, sie hätten das Gefühl, die Anzahl der erforderlichen Interaktionen seien bei *Grokker* höher gewesen als bei *Yahoo*. Die Auswertung des Screencapturings ergibt jedoch, dass während des Usabilitytests zur Erfüllung der Aufgaben im Schnitt rund 20% mehr Interaktionen auf der Oberfläche von *Yahoo* vorgenommen wurden als auf der visuellen Oberfläche von *Grokker*. Ähnliche scheinbare Widersprüche ergaben sich hinsichtlich der Qualität der erzielten Treffer, die mit Methoden der Retrievaleffektivitätsmessung und der Auswertung des Screencapturings erhoben und gleichzeitig durch Befragung der Probanden eingeschätzt wurde.

In diesen Beispielen kann anhand der breiten methodischen Abstützung festgestellt werden, dass sich durch das IRS mit VK zwar bessere Ergebnisse erzielen ließen, diese Feststellung jedoch nicht immer von den Probanden wahrgenommen wurde. Bereits mit Hilfe einer sehr kurzen Tagebuchstudie wurde das Problem des höheren Bekanntheitsgrades und des routinierteren Umgangs mit listenbasierten Ergebnis-repräsentationen deutlich gemacht. Der Proband der Tagebuchstudie erzielte im Vergleich zu den Probanden, die lediglich am kontrollierten Experiment teilnahmen, deutlich bessere Ergebnisse in *Grokker* und seine subjektive Einschätzung deckte sich wesentlich besser mit den anderweitig gemessenen Ergebnissen. Anekdotisch anzumerken bleibt, dass sein verändertes Suchverhalten im beruflichen und privaten Bereich zu einer nachhaltigen Nutzung von Suchmaschinen mit visuellen Ergebnisrepräsentationen geführt hat.

Die exemplarische Evaluation verdeutlicht die durch einen gezielten Einsatz und der integrierten Kombination geeigneter Methodenansätze erzielten Vorteile. Bereits mit wenigen Mitteln lässt sich eine breit abgestützte Evaluation durchführen, die durch den Einbezug qualitativer und quantitativer Maße eine gute Ausgangsbasis bietet für die Interpretation von Ergebnissen sowie die Identifikation von Auslösern scheinbarer Differenzen im Erhebungsmaterial.

Die Schwächen einiger Evaluationsmethoden werden somit durch die Stärken anderer Methoden ausgeglichen und die erhobenen Ergebnisse weisen insgesamt eine höhere Qualität auf.

5 Ausblick

Die Auswahl und Kombination geeigneter Methoden für eine nachhaltige und umfassende Evaluation von IRS mit VK bedeutet nach wie vor eine große Herausforderung. Anhand einer Morphologie lässt sich ein integrierter Ansatz verfolgen, der eine möglichst breite Abstützung der Ergebnisse hinsichtlich aller relevanten Aspekte eines IRS mit visueller Ausgabe gewährleistet. Für den zukünftigen evaluationsübergreifenden Vergleich von Evaluationsresultaten bezüglich der Wirksamkeit von Visualisierungen im Information Retrieval im Allgemeinen empfiehlt sich ein einheitliches methodisches Vorgehen.

Künftig gilt es auf der Grundlage des Bezugsrahmens eine konkrete Evaluationsumgebung für die Durchführung von Evaluationen von IRS mit VK zu gestalten, die als Ausgangsbasis für den evaluationenübergreifenden Vergleich dient. Langfristig lassen sich damit repräsentativere Aussagen zur Eignung von Visualisierungen im Information Retrieval ableiten.

Den vorgelegten Bezugsrahmen gilt es in nächsten Schritten durch weitere Ergebnisse empirischer Erprobung sukzessive zu verfeinern und zu optimieren. Der vorliegende Vorschlag ist somit als erster Schritt in Richtung einer einheitlichen Evaluationsgrundlage zu verstehen - ein Weg, der für die nachhaltige Evaluation neuer IRS unerlässlich ist.

Zur Autorin

Sonja Hierl, MSc BIS, Dipl.-Informationsspezialistin (FH) ist Projektleiterin bei

Swiss Institute for Information Research der
Hochschule für Technik und Wirtschaft Chur
Ringstraße/Pulvermühlestraße 57
CH-7004 Chur

Referenzen

(1) C. Arnold. Visualisierung im Information Retrieval. Magisterarbeit in der Philosophischen Fakultät IV (Informationswissenschaft) der Universität Regensburg: Regensburg, 2004.

(2) J. Bar-Ilan, M. Levene, M. Mat-Hassan. Dynamics of search engine rankings - a case study. In *Proceedings of the 3rd international workshop on web dynamics*, New York, 2004.

(3) M. M. S. Beg. A subjective measure of web search quality. In *Information Sciences* Volume 169, Issues 3-4, 1 February 2005, S. 365-381.

(4) C. Buckley, E. M. Voorhees. Evaluating Evaluation Measure Stability. In *SIGIR 2000*. Belkin, N. J., Ingwersen, P., und Leong, M.-K. (eds.); ACM, S. 33-40, Athen, 2000.

(5) C. Chen, Y. Yu. Empirical studies of information visualization: a metaanalysis. In *International Journal of Human-Computer Studies*, 53 (2000) 5, Seiten 851-866, 2000.

(6) J. Cugini. Presenting Search Results: Design, Visualization and Evaluation. In: *Workshop: Information Doors - Where Information Search and Hypertext Link*, San Antonio.

(7) Fachdatenbank Factiva, URL: http://www.factiva.com, Stand: 02.04.2007.

(8) Suchmaschine Google, URL: http://www.google.com, Stand: 02.04.2007.

(9) F. Gremy, J. M. Fessler, M. Bonnin. Information systems evaluation and subjectivity. In *International Journal of Medical Informatics* Volume 56, Issues 1-3 , December 1999, S. 13-23.

(10) Suchmaschine Grokker, URL: http://www.grokker.com, Stand: 02.04.2007.

(11) Görner, C./Ilg, R.: Evaluation der Mensch-Rechner-Schnittstelle, in: Ziegler, J./Ilg, R. (Hrg.): Benutzergerechte Software-Gestaltung. Standards, Methoden und Werkzeuge, 1993, München.

(12) D. Hawking, N. Craswell, P. Bailey *et al*. Measuring search engine quality. In *Information Retrieval*, 4 Nr.1, S. 33-59, 2001.

(13) E. C. Jensen, S. M. Beitzel, O. Frieder, O. *et al.* A framework for determining necessary query set sizes to evaluate web search effectiveness. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, May 10-14, 2005, Chiba, Japan.

(14) Suchmaschine Kartoo, URL: http://www.kartoo.com, Stand: 25.06.2006.

(15) D. A. Keim. Information Visualization and Visual Data Mining. In *IEEE Transactions on visualization and computer graphics,* Vol. 7, No. 1, January-March 2002.

(16) Koshman. Testing user interaction with a prototype visualization-based information retrieval system. In *Journal of the American Society for Information Science and Technology*, 56(8) 2005, Seiten 824-833, 2005.

(17) S. K. Kwan, S. Venkatsubramanyan. An Economic Model for Comparing Search Services. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences* (HICSS'06). S. 107-116, 2006.

(18) Nachrichtendatenbank Lexis Nexis, URL: http.//www.lexisnexis.com, Stand: 02.04.2007.

(19) T. M. Mann. Visualization of Search Results from the World Wide Web, Dissertation, Universität Konstanz, 2002.

(20) Suchmaschine MSN, http://search.msn.de, Stand: 02.04.2007.

(21) J. Mussgnug, S. Stowasser. Ein Schema zur Auswahl geeigneter Usability-Methoden - Dargestellt am Beispiel der Blickbewegungsanalyse. In *Proceedings of the 2nd annual GC-UPA Track Paderborn*, September 2004, Paderborn.

(22) C. Plaisant. The Challenge of Information Visualization Evaluation. In *Proceedings of Conference on Advanced Visual Interfaces AVI'04*.

(23) Suchmaschine Quintura, URL: www.quintura.com/, Stand: 02.04.2007.

(24) H. Reiterer, G. Tullius, T. M. Mann. INSYDER: a content-based visual-information-seeking system for the Web. In International Journal on Digital Libraries, Volume 5, Issue 1, Mar 2005, Seiten 25-41.

(25) H. Reiterer. Visuelle Recherchesysteme zur Unterstützung der Wissensverarbeitung. In Hammwöhner, R.; Rittberger, M.; Semar, W. (Hrg.): *Wissen in Aktion. Der Primat der Pragmatik als Motto der Konstanzer Informationswissenschaft. Festschrift für Rainer Kuhlen*. Konstanz, 2004. Seiten 1-21.

(26) Sarodnick, F./Brau, H.: Methoden der Usability Evaluation. Wissenschaftliche Grundlagen und praktische Anwendungen, 2006, Bern.

(27) B. Shneiderman, C. Plaisant. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. In *Proceedings of the BELIV'06 Workshop*, Venice Seiten 61-77, 2006.

(28) M. Sebrechts, J. Vasilakis, M. Miller, et al. (1999): Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,* S. 3-10, 1999.

(29) Synnestvedt, M./Chen, C.: Design and Evaluation of the Tightly Coupled Perceptual-Cognitive Task in Knowledge Domain Visualiza-tion, in The 11th International Conference on Human-Computer In-teraction (HCII 2005), 2005, Las Vegas.

(30) Suchmaschine Ujiko, URL: www.ujiko.com, Stand: 02.04.2007.

(31) Vaughan, L. (2004): New measurements for search engine evaluation proposed and tested. In *Information Processing and management* 40 (2004) S. 677-691, 2004.

(32) A. Veerasamy, N. J. Belkin. Evaluation of a tool for visualization of information retrieval results. In *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Zurich, Switzerland, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, Seiten 85-92, 1996.

(33) Womser-Hacker, C.: THeorie des Information Retrieval III: Evaluierung, in: Kuhlen, R./Seeger, T./Strauch, D. (Hrg.) (2004): Grundlagen der praktischen Information und Dokumentation. Bd. 1, S. 227-235, 2004, München.

(34) Suchmaschine Yahoo, URL: http://search.yahoo.de, Stand: 02.04.2007.

(35) R. Van Zwol, H. Van Oostendorp. Google's "I'm feeling lucky", Truly a Gamble?. In *Zhou, X. et al. (Hrg.) (2004): Web Information Systems - WISE 2004, Proceedings of the 5th International Conference on Web Information Systems Engineering*. Brisbane, Australia, Seiten 378-390, 2004.

Anmerkung

1. Eine detailliertere Ausführung zu unterschiedlichen Variationen im Einsatz dieser Methoden findet sich unter folgender URL: http://www.informationswissenschaft.ch/fileadmin/uploads/sonstiges/TAB_Methodenklassifikation.html

---

Danksagung

Ganz herzlich möchte ich mich bedanken beim Verein zur Förderung der Informationswissenschaft (VFI) für die Auszeichnung meiner Master-Arbeit und die Möglichkeit, vorliegenden Beitrag daraus zu publizieren. Ebenso bedanke ich mich recht herzlich bei meinen Betreuern, Herrn Prof. Dr. Bernard Bekavac und Herrn Prof. Dr. Siegfried Weinmann für die Begleitung der Master-Thesis.
