Handbuch Internet-Suchmaschinen 2
D. Lewandowski (Hrsg.)

Evaluierung von Suchmaschinen

Dirk Lewandowski
Hochschule für Angewandte Wissenschaften Hamburg,
Department Information,
Finkenau 35, 22081 Hamburg
dirk.lewandowski@haw-hamburg.de

Abstract. Die Evaluierung von Suchmaschinen ist von hoher Bedeutung, sowohl wenn es um die Überprüfung der Leistungsfähigkeit selbst entwickelter Systeme geht als auch, um die Qualität der bekannten Suchdienste untereinander zu vergleichen. In diesem Kapitel wird der Standardaufbau von Tests zur Messung der Retrievaleffektivität von Suchmaschinen beschrieben, um darauf aufbauend systematisch die Grenzen dieser Tests aufzuzeigen und erste Lösungsmöglichkeiten zu diskutieren. Es werden Hinweise für die Praxis gegeben, wie sich Retrievaltests mit vertretbarem Aufwand gestalten lassen, die trotzdem zu verwertbaren Ergebnissen führen.

Keywords. Suchmaschinen, Retrievaleffektivität, Retrievaltest, Nutzerverhalten, Klickdaten, Suchanfragetypen.

Einleitung

Die Bedeutung von Suchmaschinen für die Informationsbeschaffung kann kaum überschätzt werden. Die Suche im Netz ist nach der Verwendung von E-Mails die zweithäufigste Anwendung des Internet [1]. Die Anzahl der monatlich allein in Deutschland abgeschickten Suchanfragen liegt bei mehreren Milliarden [2].

Vor dem Hintergrund dieser enormen Bedeutung der Websuche verwundert es auch nicht, dass die Frage nach der Qualität der Suchmaschinentreffer von großem Interesse ist. Auf der einen Seite geht es dabei um die Frage, inwieweit die Ergebnisse der Suchmaschinen mit denen von spezialisierten Datenbanken mithalten können (s. z. B. [3]), auf der anderen Seite geht es um einen Vergleich der verschiedenen Suchmaschinen untereinander (u.a. [4][5][6]). Vor allem vor dem Hintergrund der starken Dominanz einer einzigen Suchmaschine – Google – (vgl. [7]) stellt sich die Frage, ob diese Dominanz tatsächlich durch die überragende Qualität dieser Suchmaschine gerechtfertigt ist oder ob andere Suchmaschinen zumindest eine vergleichbare Leistung bieten.

So wichtig die Qualität der Suchmaschinentreffer ist, so muss doch klargestellt werden, dass sich die Qualität von Suchmaschinen nicht auf einen Faktor reduzieren lässt. Vielmehr bildet sich die Qualität von Suchmaschinen durch ein Zusammenspiel unterschiedlicher Faktoren, mit deren Hilfe sich ein Bild von der Qualität dieser Suchwerkzeuge gewinnen lässt. Lewandowski und Höchstötter [8] benennen die vier Bereiche Qualität des Index, Qualität der Suchresultate, Qualität der Suchfunktionen und Nutzerfreundlichkeit. Allerdings kann auch in diesen Bereichen nicht jeweils nur eine Kennzahl zur Qualitätsmessung herangezogen werden, sondern in jedem Bereich gibt es zahlreiche Fragestellungen, die hier nur angerissen werden können:

Die Qualität des Indexes umfasst beispielsweise die Größe und die Vollständigkeit (im Sinne der Abdeckung des World Wide Web) des Indexes, seine Aktualität sowie länderspezifische Unterschiede zwischen den Indizes mehrerer Suchmaschinen.

Bei der Qualität der Suchresultate geht es auf der einen Seite um die Messung der Retrievaleffektivität, also der „objektiven" Qualität der Treffer, wobei tatsächliche Nutzungsszenarien möglichst ausgeblendet bleiben. Zusätzlich zur alleine stehenden Qualität der Treffer ist bei der Betrachtung verschiedener Suchmaschinen allerdings auch die Einzigartigkeit von Suchergebnissen von Bedeutung.

Hierbei ist anzumerken, dass ein Großteil der Nutzer von sich sagt, dass sie in der Regel ohne größere Probleme in den Suchmaschinen das finden, was sie suchen (vgl. [9][10]). Auf der einen Seite ist diese hohe Zufriedenheit natürlich als Erfolg für die Suchmaschinen zu sehen, auf der anderen Seite liegt aber die Vermutung nahe, dass zumindest Laiennutzern nicht bekannt ist, dass sie relevante Treffer bei ihrer Recherche verpassen könnten und falls doch, in welchem Umfang dies geschehen kann. Studien zur Überlappung der Suchergebnisse der bekannten Suchmaschinen zeigen, dass, wenn man die ersten 10 Ergebnisse der Suchmaschinen betrachtet, nur geringe Überschneidungen festzustellen sind und sich daher (neben einem Weiterblättern in der Trefferliste einer Suchmaschine) das Aufrufen der Ergebnisse für dieselbe Suchanfrage in einer anderen Suchmaschine in vielen Fällen lohnen würde [11].[1] Vor diesem Hintergrund ist es nicht verwunderlich, dass eine über den wissenschaftlichen Bereich hinausgehende Diskussion um die Qualität von Suchmaschinenergebnissen zumindest bislang nicht stattfindet.

Als dritten Bereich benennen Lewandowski und Höchstötter die Qualität der Suchfunktionen. Hier geht es einerseits um den Umfang der angebotenen Funktionen (im Rahmen einer erweiterten Suche bzw. durch die Eingabe von Kommandos), andererseits aber auch um die Funktionsfähigkeit dieser Funktionen. In Studien konnte gezeigt werden, dass Suchmaschinen zumindest einige erweiterte Funktionen anbieten, die nur unzuverlässig funktionieren [12][13].

Unter der Nutzerfreundlichkeit von Suchmaschinen wird einerseits die Usability verstanden[2], andererseits geht es auch um alle weiteren Berührungspunkte von Nutzern mit Suchmaschinen, die widerspiegeln, wie gut Suchende mit Suchmaschinen umgehen können. Es werden dazu das Design des Interface betrachtet, die Akzeptanz spezieller Suchfunktionen und Operatoren, die Verarbeitung der Anfragen und die Benutzerführung.

Ausgehend von dem skizzierten Qualitätsmodell soll in diesem Kapitel der Stand der Forschung zum Bereich „Qualität der Suchresultate" dargestellt werden. Aus dem Vorangegangenen ist allerdings schon deutlich geworden, dass dieser nicht isoliert betrachtet werden kann, sondern dass insbesondere die Qualität des Indexes und Fragen der Nutzerfreundlichkeit mit in die Studien eingehen. Im letztgenannten Bereich geht es im beschriebenen Kontext vor allem um die Zufriedenheit der Nutzer mit den Suchergebnissen, die sich, wie noch zu zeigen sein wird, nur schwer von der „objektiven" Qualität der Treffer trennen lässt. Eine besondere Problematik bei der Evaluierung der Treffer von Web-Suchmaschinen ergibt sich aus der Tatsache, dass die Anbieter ihre

[1] Dies gilt umso mehr, wenn man mit den Ergebnissen zu einer Suchanfrage nicht zufrieden ist. Untersuchungen zeigen, dass in vielen Fällen der Wechsel der Suchmaschinen (bei Beibehaltung der Suchanfrage) zu einem Erfolg führt [4]. Bei den Studien zur Überschneidung der Suchergebnisse darf allerdings nicht übersehen werden, dass diese Vergleiche nur auf Basis der URLs erfolgen. Jedoch kann dasselbe Dokument unter verschiedenen URLs abgelegt sein. Daher ist davon auszugehen, dass die Überschneidungen zwischen den Suchmaschinen zumindest etwas höher liegen als in den zitierten Studien konstatiert.

[2] S. dazu den Beitrag von Sonja Quirmbach in diesem Band.

Systeme nicht offenlegen: In den Tests wird damit neben den Treffern selbst auch gleichzeitig der Index, also die zugrunde liegende Datenbank, mitbewertet. Dokumente, die die Suchmaschine nicht kennt, kann sie auch auf eine Anfrage hin nicht ausgeben. Zwar mag dieses Problem in der Testpraxis in vielen Fällen keine Rolle spielen, da die Menge der potenziell relevanten Dokumente meist über der Menge der vom Nutzer zu bewältigenden Menge liegen dürfte, eine Beeinflussung der Tests kann aber trotzdem nicht ausgeschlossen werden.

Die in diesem Kapitel dargestellten Verfahren zur Messung der Trefferqualität basieren auf einer langen Tradition der Evaluierung von Information-Retrieval-Systemen. Zwar mag aus der kurzen Diskussion des Qualitätsmodells schon klar geworden sein, dass die Bezeichnung „Evaluierung" nicht passgenau ist, da man unter diesem Begriff doch mehr erwarten würde als nur einen Ausschnitt aus den möglichen Qualitätsmessungen, jedoch soll aufgrund der Tradition auch in diesem Kapitel an der Bezeichnung festgehalten werden.

Die in den folgenden Abschnitten geführte Diskussion um die Qualität der Suchergebnisse fragt stets danach, was eigentlich mit bestimmten Tests bzw. mit bestimmten Kennzahlen gemessen werden soll. Aus der Tradition der Evaluierung im Information Retrieval lässt sich zwar erklären, warum bestimmte Verfahren mit der Qualitätsmessung von Suchsystemen gleichgesetzt werden, eine solche Tradition birgt aber auch immer die Gefahr, aufgrund von scheinbaren Selbstverständlichkeiten die Grundannahmen der eigenen Methodik nicht weiter zu überprüfen.

Wie gezeigt werden wird, eignen sich die klassischen Verfahren der IR-Evaluierung nur (noch) bedingt zur Evaluierung von Suchmaschinen. Dies ist vor allem auf zwei Punkte zurückzuführen:

- Die Informationssuche im Web ist nur zum Teil durch das Anfrage-Ergebnis-Paradigma abgedeckt. In vielen Fällen entstehen weitergehende Interaktionen zwischen Nutzern und Suchmaschinen, die in klassischen Tests nicht abgedeckt werden können. Diese Einschränkung der klassischen Methodik ist allerdings nicht allein auf die Web-Suchmaschinen bezogen, sondern betrifft die meisten Information-Retrieval-Systeme. Schon seit einigen Jahren wird eine Erweiterung bzw. Veränderung unter der Bezeichnung „Interactive Information Retrieval" diskutiert (vgl. u. a. [14]).
- Die Präsentation der Treffer in Suchmaschinen unterscheidet sich mittlerweile grundlegend von der sonst üblichen Listenpräsentation. Zwar bildet die Trefferliste weiterhin den Kern der Trefferpräsentation in Suchmaschinen, allerdings werden Ergebnisse innerhalb der Liste auf unterschiedliche Weise präsentiert sowie weitere Treffer, wiederum in unterschiedlicher Präsentation, um die Trefferliste herum platziert [15]. Dies führt dazu, dass in den Augen der Nutzer nicht mehr alle Treffer gleich sind, wie das noch in der traditionellen IR-Evaluierung angenommen wird. Gemessen werden kann diese unterschiedliche Beachtung der Treffer einerseits durch Eyetracking-Studien (welche die Wahrnehmung bestimmter Elemente messen, s. u. a. [16][17][18,19]), andererseits durch Messung der Klicks auf bestimmte Elemente der Trefferseiten (wobei die Klickhäufigkeiten erfasst werden; s. [20]).[3]

[3] Einen Überblick über die Ergebnispräsentation in Suchmaschinen bietet der Artikel von Lewandowski und Höchstötter im ersten Band dieses Handbuchs.

Das Ziel dieses Artikels ist, aufbauend auf den Standardverfahren für die Evaluierung der Treffer von Suchmaschinen aufzuzeigen, wo die Schwachstellen dieser Verfahren liegen und welche weiteren Faktoren für eine valide Evaluierung hinzugenommen werden sollten. Ausgangspunkt dieser Überlegungen bilden Erkenntnisse zum typischen Verhalten der Suchmaschinennutzer.

Der Rest dieses Artikels ist wie folgt aufgebaut: Nach einem Überblick über die Standards beim Aufbau von Suchmaschinen-Retrievaltests werden die Schwachstellen solcher Tests diskutiert. Alternative Verfahren zu den konventionellen Tests werden dargestellt, wobei hier vor allem auf die Messung der Trefferqualität durch die Auswertung von Klickdaten sowie auf die sessionbasierte Evaluierung eingegangen wird. Auf dieser Basis werden Empfehlungen für die Evaluierung von Suchmaschinen in der Praxis gegeben, wobei auch auf die Übertragbarkeit der Verfahren auf andere Kontexte als die Websuche eingegangen wird. Im Fazit wird der Stand der Suchmaschinen-Evaluierung zusammengefasst und ein Ausblick auf die weitere Entwicklung gegeben.

1. Standardaufbau von Tests zur Retrievaleffektivität

In diesem Abschnitt wird auf den Aufbau von Tests zur Trefferqualität von Suchmaschinen eingegangen und die wesentlichen Ergebnisse aus den bislang durchgeführten Tests präsentiert.

Retrievaleffektivität bezeichnet die Fähigkeit einer Suchmaschine, auf eine Anfrage relevante Dokumente auszugeben. In zahlreichen Tests wurden populäre Suchmaschinen evaluiert. Meist wurden für diese Tests englischsprachige Anfragen verwendet (u. a. [21][22][23]; ein Überblick findet sich in [4], es finden sich aber auch Tests mit deutschsprachigen Anfragen (z. B. [24][4] und (in geringerem Maße international veröffentlicht) mit Anfragen in anderen Sprachen [25][26][27][28]. Einen Überblick über nicht-englischsprachige Suchmaschinentests der letzten Jahre, der über die reinen Retrievaleffektivitätstests hinausgeht, findet sich in [29].

Für die meisten Tests wird ein (teils leicht modifizierter) Standardaufbau verwendet, wie er aus der Information-Retrieval-Literatur und den Evaluierungsinitiativen (v. a. TREC, vgl. [30]) bekannt ist. Eine bestimmte Menge an Suchanfragen wird an unterschiedliche Suchmaschinen gesendet, die zurückgegebenen Ergebnisse werden zuerst anonymisiert und gemischt, dann werden sie Juroren zur Beurteilung vorgelegt. Für die Auswertung werden die Ergebnisse wieder den untersuchten Systemen und den Trefferplätzen, auf denen sie ausgegeben wurden, zugeordnet. Gemessen wird schließlich meist die *Präzision* der Suchergebnisse, d. h. die Fähigkeit des Systems, nur relevante Treffer auszugeben. Die Tests beschränken sich meist auf eine bestimmte Anzahl von Trefferpositionen, da insbesondere bei Web-Suchmaschinen die Trefferanzahl meist die für einen Nutzer prüfbare Menge weit übersteigt.

Die Prämissen, die dem Testaufbau in TREC zugrunde liegen, werden von den leitenden Personen dieser Evaluierungsinitiative wie folgt benannt: Der typische Nutzer des Systems ist ein „dedicated searcher“ (abgegrenzt von Novizen/Laien), der Nutzer wünscht sowohl einen hohen Recall als auch eine hohe Precision und ist bereit, eine Vielzahl von Dokumenten durchzusehen [30]. Durch die Adaption der TREC-Methodik (grundlegend in [31]), die ursprünglich mit Fokus auf technische Aspekte entwickelt wurde, haben diese Grundannahmen auch Eingang in die Suchmaschinen-Evaluierung gefunden.

Dem gegenüber stehen jedoch aktuelle Erkenntnisse über das Nutzerverhalten im Rahmen der Web-Suche, die bspw. aus der Auswertung von Logfiles kommerzieller Suchmaschinen bzw. aus Befragungen stammen und mit obigen Annahmen nicht übereinstimmen. Vielmehr ergibt sich folgendes Nutzerporträt (ausführlich zum Verhalten der Suchmaschinennutzer: [32]): Suchmaschinen werden von allen Nutzergruppen, d. h. von Laien bis Profi-Rechercheuren verwendet; die Laiennutzung allerdings ist der Standardfall. Dies zeigt sich vor allem in allgemein formulierten und sehr kurzen Suchanfragen [33][34][35]. Expertenwissen des Nutzers im Rahmen seiner Web-Suche kann damit nicht vorausgesetzt werden.

Während ihrer Suchsessions sehen Suchmaschinennutzer insgesamt nur wenige Dokumente an [35]. In den weit überwiegenden Fällen werden nur die ersten Ergebnisseiten (zusammenfassend dargestellt in [36]) und dort wiederum lediglich die ersten Trefferpositionen betrachtet [16,37][19][37]. Der Nutzer ist also nicht bereit, eine Vielzahl an Dokumenten durchzusehen.

Ein Schwachpunkt bisheriger Untersuchungen ist damit, dass sie weitgehend unabhängig vom tatsächlichen Nutzer- bzw. Klickverhalten gestaltet sind bzw. aus anderen Zusammenhängen gewonnene Annahmen zugrunde legen.

Zwar gab es bspw. mit dem HARD Track (http://ciir.cs.umass.edu/research/hard/) und dem Interactive Track (http://trec.nist.gov/data/t11_interactive/t11i.html) Ansätze, Informationen über den Suchenden und/oder den Kontext der Suche in die Ergebnisliste einfließen zu lassen, doch eine tiefgreifende Berücksichtigung des tatsächlichen Nutzerverhaltens findet bislang bei der Messung der Retrievaleffektivität nicht statt.

Die Untersuchungen im Rahmen von Evaluierungsinitiativen basieren auf Testkollektionen. Dies kommt dem Ziel einer optimalen Vergleichbarkeit der zu untersuchenden Information-Retrieval-Systeme näher, da bei einem solchen Testaufbau in der Tat einzig und allein die von den verschiedenen Systemen ausgegebenen Treffer verglichen werden, da der zugrunde liegende Index (in diesem Fall also die Testkollektion) bei allen Systemen der gleiche ist. Die Nachteile einer solchen Vorgehensweise liegen allerdings zum einen darin, dass sich die Ergebnisse nicht direkt auf größere Datenbestände übertragen lassen (Skalierung) und zum anderen darin, dass nur Systeme verglichen werden können, die freiwillig an der entsprechenden Evaluierungsinitiative teilnehmen.

Aufgrund des in der Einleitung angesprochenen übergreifenden Interesses an der Qualität der Suchresultate der bekannten Web-Suchmaschinen (die ihre Systeme nicht im Rahmen der bekannten Evaluierungsinitiativen begutachten lassen) muss jedoch der Nachteil des Index-Einflusses in Kauf genommen werden, wenn es um die Evaluierung der Retrievaleffektivität der Web-Suchmaschinen geht. Suchmaschinentests „von außen“ (also ohne einen direkten Zugang zum System) bewerten also nicht allein die Ergebnisse, sondern immer auch den zugrunde liegenden Datenbestand (der sich zwischen den Suchmaschinen trotz ihres Anspruchs, „das ganze Web“ abzudecken, durchaus deutlich unterscheidet (s. u. a. [38]).

Zwar wurden im Rahmen von Evaluierungsinitiativen auch Testkollektionen mittels Web-Crawl gebildet (so im TREC Web Track), jedoch erfüllen die mit diesen durchgeführten Tests nicht den Anspruch, die wichtigsten Suchmaschinen abzudecken (da ja die großen kommerziellen Anbieter ihre Systeme explizit an dem Test teilnehmen lassen müssten). Zusätzlich ergibt sich das Problem der Übertragbarkeit von einer Testkollektion auf das gesamte Web; die Skalierbarkeit spielt im Web-Kontext eine weit größere Rolle als bei anderen Typen von Informationssystemen. Trotz der beschriebenen Nachteile bieten die Evaluierungsinitiativen erprobte Methoden, die im Folgenden in ihrer Adaption auf Suchmaschinentests dargestellt werden sollen.

1.1. Testaufbau

Der Aufbau von Retrievaltests orientiert sich meist an den von Tague-Sutcliffe [39] aufgestellten Schritten. Die Besonderheiten von Web-Suchmaschinen haben Gordon und Pathak [22] sowie (darauf aufbauend) Hawking et al. [31] berücksichtigt. Die von Hawking et al. genannten fünf Kriterien für einen validen Suchmaschinentest beziehen sich auf die Abbildung realer Informationsbedürfnisse, auf die Mitteilung des Informationsbedürfnisses (falls ein Informationsvermittler eingesetzt wird), auf die genügend große Zahl von Testanfragen, auf die Auswahl der wichtigsten Suchmaschinen sowie auf den sorgfältigen Aufbau der Untersuchung und eine ebensolche Durchführung. Diese Kriterien dürften zumindest bei den neueren Studien erfüllt sein.

Im Folgenden soll der Aufbau und die Durchführung eines Retrievaltests in seinen einzelnen Schritten genauer beschrieben werden. Dabei soll auch herausgearbeitet werden, an welchen Punkten Entscheidungen zu treffen sind, die sich auf die Ergebnisse des Tests auswirken können. Die zu verwendenden Kennzahlen für die Auswertung des Tests werden in einem eigenen Abschnitt diskutiert, auch wenn das Testdesign zumindest, wenn bestimmte Kennzahlen verwendet werden sollen, entsprechend modifiziert werden muss.

1.1.1. Auswahl der Suchmaschinen

Die Auswahl der zu testenden Suchmaschinen hängt natürlich vom Zweck der Untersuchung ab: Soll ein selbst entwickeltes System mit bestehenden Systemen verglichen werden? Sollen die wichtigsten Web-Suchmaschinen miteinander verglichen werden? Oder sollen neue Suchmaschinen untereinander bzw. mit einer Referenzsuchmaschine (wie bspw. Google; man spricht hier auch von einem „Gold-Standard“) verglichen werden?

Da die Anzahl der zu testenden Suchmaschinen einen wesentlichen Einfluss auf den Umfang des Tests hat, ist abzuwägen zwischen einer wünschenswerten, möglichst großen Abdeckung von Suchdiensten und dem Aufwand, den man selbst betreiben sowie den Juroren zumuten kann.

Für Vergleichstests gängiger Suchmaschinen lässt sich damit empfehlen, zumindest die hinsichtlich ihrer Marktanteile bedeutendsten Suchmaschinen auszuwählen. Allerdings ist dabei einerseits zu beachten, dass vor allem Suchportale, aber auch scheinbar eigenständige Suchmaschinen oft auf die Ergebnisse einer der bekannten Suchmaschinen zurückgreifen [40][41]. Andererseits unterscheiden sich die Marktanteile der Suchmaschinen in den unterschiedlichen Ländern erheblich [42], sodass die populärsten Suchmaschinen für das jeweilige Land berücksichtigt werden sollten. Insbesondere bei länderübergreifenden Tests ist dies zu berücksichtigen.

1.1.2. Anzahl der Suchanfragen

Wünschenswert ist natürlich eine möglichst hohe Anzahl von Suchanfragen. In der Praxis zeigt sich allerdings, dass es oft Probleme gibt, genügend Juroren zu finden, um die Ergebnisse zu diesen Suchanfragen bewerten zu lassen. In den meisten Tests wird mittlerweile ein Minimum von 50 Suchanfragen verwendet, allerdings gibt es für diese Zahl keinen „Beweis“, sondern sie hat sich aufgrund der Erfahrungen in solchen Tests herausgebildet. Stark abhängig ist die Anzahl der zu verwendenden Suchanfragen natürlich vom Zweck des Tests. Sollen möglichst viele Themenbereiche und ein verschiedenartiges Suchverhalten abgedeckt werden, muss die Anzahl der Suchanfragen entsprechend erhöht werden.

1.1.3. Art der Suchanfragen

Bei der Auswahl der Suchanfragen muss unterschieden werden zwischen Tests, die versuchen, Aussagen über die Trefferqualität der Suchmaschinen allgemein zu treffen und solchen, die sich bewusst auf ein bestimmtes Thema bzw. die Anfragen einer bestimmten Nutzergruppe (z.B. von Kindern) beschränken.

Soll der Test eine allgemeine Aussage über die Trefferqualität der Suchmaschinen treffen, so sind die Suchanfragen möglichst breit zu wählen. Hilfreich bei der Auswahl passender Suchanfragen sind Suchhäufigkeiten, die beispielsweise aus den Logfiles einer Suchmaschine gewonnen werden können. Dies setzt allerdings voraus, dass man zumindest den Zugriff auf die Suchanfragedaten einer Suchmaschine hat. Ist dies nicht der Fall, so kann man zur Ermittlung von Suchhäufigkeiten auf die Tools der Suchanzeigenanbieter (wie etwa Google AdWords) zurückgreifen, welche (wenn auch mit einer anderen Zielsetzung) solche Daten zur allgemeinen Nutzung anbieten.

Allgemeine Tests sollten sowohl populäre als auch selten gestellte Anfragen abdecken, dazu sollte auch die Länge der Suchanfragen beachtet werden. Durchschnittliche Längen (in Wörtern) und die Verteilung der Suchanfragen nach Wortanzahl sind bekannt [36] und können bei der Auswahl der Suchanfragen berücksichtigt werden. Gerade bei Retrievaltests, die nur auf einer relativ kleinen Auswahl von Suchanfragen beruhen, müssen diese mit besonderer Sorgfalt ausgesucht werden.

Ist es das Ziel des Tests, die Eignung verschiedener Suchmaschinen für die Recherche in bestimmten Themenfeldern zu untersuchen, so gelten die gleichen Voraussetzungen wie beschrieben, nur dass sich die Suchanfragen natürlich auf das zu untersuchende Themenfeld beziehen müssen. Auch bei der Beschränkung auf bestimmte Nutzergruppen ist auf deren Anfrageverhalten einzugehen, was sich allerdings als schwierig erweisen kann, da man zwar in den allgemeinen Suchanfragendatenbanken die Häufigkeit auch von Anfragen aus bestimmten Themenbereichen überprüfen kann, diese Datenbanken aber keine Aufschlüsselung nach Nutzergruppen erlauben.

1.1.4. Herkunft der Suchanfragen

Die Möglichkeit, Suchanfragen aus Logfiles oder den Tools der Anbieter von Suchanzeigen zu gewinnen, wurde bereits genannt. Der Nachteil dieser Methode ist, dass man damit nur die „nackten" Suchanfragen bekommt. Gerade allerdings, wenn man die Suchergebnisse hinterher Juroren vorlegt, die die jeweilige Suchanfrage nicht selbst gestellt haben, sind oft Kontextinformationen zur Suchanfrage nötig.

Eine Lösung für dieses Problem besteht darin, Suchanfragen von echten Nutzern abzufragen. In einer solchen Befragung lassen sich auch zusätzliche Informationen abfragen („Bitte beschreiben Sie kurz, was Sie mit dieser Suchanfrage erreichen wollten" o.ä.), die die Juroren bei der Bewertung unterstützen können. Schwierig ist es bei einer solchen Sammlung von Suchanfragen allerdings, Suchanfragen zu gewinnen, die in ihren Häufigkeiten (die man zusätzlich aus den genannten Quellen ermitteln kann) dem tatsächlichen Anfrageverhalten der Nutzer entsprechen.

Soll ein eher allgemeiner Test durchgeführt werden, können auch besonders populäre Suchanfragen verwendet werden. Listen solcher Suchanfragen werden von den bekannten Suchmaschinen jeweils zum Jahresende veröffentlicht.[4] Zu bedenken ist

[4] Bing: http://www.bing.com/community/site_blogs/b/search/archive/2010/11/29/the-top-bing-searches-for-2010-the-year-of-the-celebrities.aspx; Google: http://www.google.com/intl/en/press/zeitgeist2010/; Yahoo: http://yearinreview.yahoo.com/.

allerdings, dass diese Suchanfragen auf der einen Seite gefiltert sind (so finden sich bspw. keine Suchanfragen aus dem pornographischen Bereich) und dass die Suchmaschinen auch manuell auf die möglichst gute Beantwortung gerade dieser Anfragen ausgerichtet sein könnten, was einen Test mit diesen Anfragen verfälschen würde.

1.1.5. Darstellung der Suchanfragen

Wie auch immer die Suchanfragen gewonnen wurden, empfiehlt es sich auf jeden Fall, die reinen Suchanfragen um eine Beschreibung des dahinterstehenden Informationsbedürfnisses anzureichern. Diese Beschreibung wird dann den Juroren zur Unterstützung ihrer Bewertung mit angezeigt.

Um den Juroren eine möglichst genaue Bewertung der Dokumente zu ermöglichen, kann auch zwischen der Beschreibung der Suchanfrage und expliziten Bewertungsinformationen, in denen benannt wird, welche Art von Dokumenten relevant ist, unterschieden werden. Dieses Verfahren ist insbesondere geeignet, wenn die Suchanfragen von echten Nutzern abgefragt werden, da diese in der Regel am besten beschreiben können, welche Intention hinter ihrer Anfrage stand und wie die idealen Dokumente zu dieser Anfrage für sie aussehen würden.

Folgendes Beispiel verdeutlicht die Unterschiede zwischen Suchanfrage, Beschreibung und Bewertungsinformationen:

> **Suchanfrage**: Lebenshaltungskosten USA
> **Beschreibung**: Wie hoch sind die Lebenshaltungskosten in den USA? Welcher Anteil des Gehalts ist für die Miete einzurechnen, welcher für Nebenkosten und welcher für Lebensmittel?
> **Bewertungsinformationen**: Relevant sind Dokumente, die einen Überblick über die Lebenshaltungskosten in den USA geben und nicht nur einen der genannten Aspekte behandeln.

Sofern dem Testleiter klar ist, welche Eigenschaften die als relevant zu bewertenden Dokumente haben sollen, ist es sinnvoll, den Juroren eine entsprechende Hilfestellung zu geben. Wenn allerdings eher Suchanfragen von allgemeinem Interesse verwendet werden, ist eine so genaue Anweisung an die Juroren nicht nötig, da diese selbst in der Lage sind, relevante von nicht relevanten Dokumenten zu unterscheiden.

Sollen Juroren die Ergebnisse für mehrere (oder gar ein kleines Jurorenteam die Ergebnisse für alle) Suchanfragen bewerten, so sind konkrete Bewertungsanweisungen vonnöten, da sonst die Gefahr besteht, dass der persönliche Geschmack der Juroren einen zu starken Einfluss auf die Ergebnisse hat.

1.1.6. Anzahl der Ergebnisse pro Suchanfrage

Da in Retrievaltests das typische Verhalten der betreffenden Nutzergruppe untersucht werden soll, ist auch bei der Festlegung der Anzahl der Ergebnisse pro Suchanfrage dieses Verhalten zu berücksichtigen. Für allgemeine Suchmaschinennutzer gilt, dass diese in der Regel nur wenige Suchergebnisse betrachten [33][32], sodass hier der Cut-Off-Wert relativ niedrig gewählt werden kann. Die gängigen Untersuchungen beschränken sich in der Regel auf die ersten 10 Ergebnisse zu jeder Suchanfrage.

Nicht unerwähnt bleiben soll hier allerdings, dass für bestimmte Fragestellungen die Untersuchung einer (wesentlich) höheren Anzahl von Ergebnissen sinnvoll sein kein. Als Erstes ist hier an besondere Nutzergruppen zu denken. Wenn etwa Aussagen über die Eignung der Suchmaschinen für Profi-Rechercheure, die in der Regel sehr tief in ein Thema einsteigen, untersucht werden soll, so sollten auch entsprechend viele Treffer berücksichtigt werden. Auch wenn der Frage nachgegangen werden soll, ob

Suchmaschinen tatsächlich die besten Treffer auf den vorderen Plätzen zeigen oder ob sich weiter hinten in den Trefferlisten noch bessere (oder zumindest ebenbürtige) Treffer finden lassen, ist die Betrachtung einer größeren Treffermenge nötig.

Auch bei der Festlegung der Anzahl der Treffer ist wiederum zu beachten, dass der Umfang des Tests mit einer höheren Trefferzahl wiederum ansteigt. So finden sich in der Literatur auch kaum Tests, die mehr als die ersten 20 Treffer berücksichtigen.

1.1.7. Anzahl der Juroren

Die Anzahl der für einen Test nötigen Juroren richtet sich natürlich nach der Anzahl der zu untersuchenden Suchanfragen. Üblicherweise bewertet jeder Juror sämtliche Treffer, die zu einer Suchanfrage von allen Suchmaschinen ausgegeben werden. Insbesondere bei einer hohen Zahl von Suchanfragen ist dies oft nicht möglich, dann kann jeder Juror auch die Ergebnisse zu mehreren Suchanfragen bewerten. Allerdings ist darauf zu achten, dass die Ergebnisse des Tests nicht dadurch verfälscht werden, dass die Ergebnisse nur von einer kleinen Jurorengruppe, die bestimmte Eigenschaften oder Vorlieben hat, verfälscht werden.

1.1.8. Auswahl der Juroren(gruppen)

Führt man einen Test für einen bestimmten Themenbereich durch, sollten die Juroren selbstverständlich mit dem Thema vertraut sein. In bestimmten Kontexten kann es sinnvoll sein, Experten als Juroren einzusetzen, wobei sich die Gewinnung von Experten allerdings oft als schwierig erweist.

Werden in dem Retrievaltest Anfragen eher allgemeiner Natur verwendet, so können als Juroren Laien eingesetzt werden. Sehr häufig werden Studierende als Juroren verwendet, was bei Retrievaltests auch oft kritisiert wird. Bislang nicht untersucht wurde allerdings, inwieweit sich verschiedene Jurorengruppen bei „General-Interest-Tests“ in ihren Bewertungen voneinander unterscheiden.

1.1.9. Anzahl der Juroren pro Suchanfrage

In der Regel wird jeder Treffer nur von einer Person beurteilt. Dies führt zu Kritik insbesondere in Verbindung mit der Verwendung einer bestimmten Jurorengruppe (s.o.). Da aber davon auszugehen ist, dass auch innerhalb einer einzelnen Gruppe die Interrater-Reliabilität nur ein gewisses Maß erreicht [43], ist es sinnvoll, die Treffer zumindest von zwei Personen bewerten zu lassen. Es gibt allerdings bislang keine gesicherten Erkenntnisse darüber, ob sich die Zuverlässigkeit der Tests durch die Hinzunahme weiterer Juroren signifikant erhöhen lässt.

Sollen in einem Test die Bedürfnisse unterschiedlicher Nutzergruppen abgebildet werden, müssen die Dokumente natürlich zumindest jeweils einmal von einem Vertreter jeder Nutzergruppe bewertet werden. Wenn zum Beispiel der Frage nachgegangen werden soll, inwieweit sich allgemeine Suchmaschinen zur Befriedigung der professionellen Informationsbedürfnisse von Mitarbeitern von Krankenhäusern eignen, so sind die Treffer zumindest jeweils einmal von Juroren aus der Berufsgruppe Ärzte und einmal von Juroren aus der Berufsgruppe Pflegepersonal zu bewerten.

1.1.10. Bewertung der Dokumente

Bei der Bewertung der Dokumente ist zuerst einmal zwischen einer binären (ja/nein) und einer Skalenbewertung zu unterscheiden. Im Kontext der Web-Suche hat sich

dabei gezeigt, dass eine binäre Bewertung nur wenig diskriminierend ist: Nutzer tendieren dazu, Dokumente auch bei einem relativ geringen Informationsgehalt als relevant zu bewerten. Werden die gleichen Dokumente auf einer Skala bewertet, zeigen sich dann allerdings größere Unterschiede. Es scheint für Suchmaschinen kein allzu großes Problem zu sein, *irgendwie* relevante Dokumente auf den vorderen Rängen der Trefferlisten anzuzeigen; die Suchmaschinen unterscheiden sich oft nur dann signifikant, wenn gemessen wird, ob sie in der Lage sind, *hoch* relevante Dokumente anzuzeigen.

In der Praxis hat es sich bewährt, die Juroren die Dokumente sowohl binär als auch auf einer Skala bewerten zu lassen. So kann in der Auswertung auch festgestellt werden, wo bestimmte Juroren(-gruppen) ihre Grenze zwischen relevant und nicht relevant setzen.

Zur Gestaltung der Skalen ist zu sagen, dass sich Skalen mit fünf Punkten weitgehend bewährt haben. Experimente mit Prozentbewertungen zeigen, dass die Nutzer dort sehr häufig grobe Bewertungen („20%“, „80%“) vornehmen und die eigentlich möglichen Differenzierungen nicht ausnutzen.

1.1.11. Berücksichtigung der Trefferbeschreibungen

In den meisten Retrievaltests werden nur die Trefferdokumente selbst berücksichtigt. Ausgehend von der Grundannahme, dass ein Nutzer die Trefferliste von oben nach unten abarbeitet (also jeden Treffer auswählt), erscheint ein solches Vorgehen auch sinnvoll. Allerdings werden Suchmaschinennutzer stark durch die Treffer*beschreibungen* innerhalb der Trefferliste beeinflusst. Diese sind wesentlich für die Entscheidung, ob ein Treffer angeklickt wird oder nicht. Ist eine Suchmaschine nun nicht in der Lage, einen relevanten Treffer auch so zu beschreiben, dass der Nutzer davon ausgeht, dass es sich um einen relevanten Treffer handelt, so ist dieser Treffer genaugenommen als verloren anzusehen, da er in einer realen Nutzungssituation nicht berücksichtigt werden würde. Zusammen mit der Tatsache, dass die gängigen Web-Suchmaschinen in einer nennenswerten Zahl von Fällen den Nutzer durch fehlende Konformität von Beschreibung und Treffer in die Irre führen [4], sollten die Trefferbeschreibungen auf jeden Fall in Tests berücksichtigt werden.

1.1.12. Weitere Hinweise für den Testaufbau

Es gibt weitere Faktoren, die die Testergebnisse verfälschen können. Insbesondere ist die Herkunft der Treffer (also von welcher Suchmaschine sie ursprünglich ausgegeben wurden) zu verschleiern, da sonst in der Bewertung starke Markeneffekte zu beobachten sind [44][45]. Weiterhin sollte die ursprüngliche Reihung der Treffer für die Juroren nicht sichtbar sein, um Lerneffekte auszuschließen [46]. Auch sollten die Treffer der verschiedenen Suchmaschinen in der Bewertung vermischt werden. Dubletten, also Treffer, die von mehreren Suchmaschinen ausgegeben werden, sollten den Juroren nur einmal (pro Suchanfrage) zur Bewertung vorgelegt werden, um einheitliche Urteile zu erhalten.

Wie bereits angeklungen ist, sind beim Aufbau eines Retrievaltests einige Entscheidungen zu treffen, die direkte Auswirkungen auf den für den Test nötigen Aufwand haben. Die Anzahl der zu bewertenden Items lässt sich folgendermaßen berechnen:

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

Für einen einfachen Test mit 3 Suchmaschinen, 50 Suchanfragen, einem Cut-off bei 10 Suchergebnissen und der Bewertung durch nur jeweils einen Juror ergeben sich 1.500 zu bewertende Items. Nimmt man nur eine Suchmaschine mehr hinzu, sind wei-

tere 500 Dokumente zu bewerten; möchte man die Dokumente jeweils von zwei Juroren bewerten lassen, verdoppelt sich die Anzahl der zu bewertenden Items. Mit diesen Zahlen soll beispielhaft verdeutlicht werden, dass selbst bei einem Test mit vermeintlich geringem Umfang ein relativ hoher Aufwand für die Bewertungen nötig ist und dass viele wünschenswerte Erweiterungen schnell zu einer extremen Erhöhung des Aufwands führen.

1.2. Kennzahlen

In diesem Abschnitt wird auf gängige Kennzahlen, wie sie in der Suchmaschinen-Evaluierung verwendet werden, eingegangen. Ausgehend von den klassischen Retrievalmaßen Precision und Recall soll gezeigt werden, dass in Retrievaltests immer die Frage im Vordergrund stehen sollte, was eigentlich gemessen werden soll, da sich die Kennzahlen erheblich in ihren Intentionen unterscheiden und damit die auf ihnen basierenden Ergebnisse eines Tests entsprechend unterschiedlich ausfallen können.

1.2.1. Precision und Recall

Um die Qualität der Treffer zu bewerten, wurden unterschiedliche Kennzahlen entwickelt. Entsprechend der Idealvorstellung einer vollständigen und vollständig relevanten Treffermenge wurden die mittlerweile „klassischen" Retrievalmaße entwickelt. An ihnen kann schon gezeigt werden, in welcher Weise die Auswahl der in einem Test verwendeten Kennzahlen eine bestimmte Annahme über den suchenden Nutzer enthält. Während eben im klassischen Information Retrieval, welches von einem Profi-Rechercheur in der Wissenschaft oder der Wirtschaft ausgeht, der aus beruflichen Gründen recherchiert und durch seine Recherche ein vollständiges Bild von einem Thema erhalten will, die Vollständigkeit des Ergebnisses eine Rolle spielt, so ist dies bei der Web-Recherche in den allermeisten Fällen überhaupt nicht gewünscht. Vielmehr werden meist nur ein paar relevante Dokumente, oder gar nur eines, gewünscht.

Das bedeutendste (und in Retrievaltests am häufigsten verwendete) Maß ist die *Precision*. Diese gibt den Anteil der relevanten ausgegebenen Treffer an der Gesamtheit der ausgegebenen Treffer an. Dieses Maß ist relativ leicht zu bestimmen: Alle ausgegebenen Treffer werden einem oder mehreren Juroren zur Beurteilung vorgelegt und anschließend wird ausgezählt, wie hoch der Anteil der relevanten Treffer ist. Bei großen Treffermengen (wie sie bei Suchmaschinen die Regel sind) wird die Precision nur bis zu einem Cut-off-Wert gemessen; meist werden die ersten 10 ausgegebenen Ergebnisse bewertet. Zwar werden immer häufiger Ableitungen aus der „reinen" Precision verwendet, jedoch dominiert diese Kennzahl mit ihren Ableitungen weiterhin die Information-Retrieval-Evaluation.

Das zweite klassische Retrievalmaß ist der *Recall*. Dieser wird bestimmt als der Anteil der relevanten ausgegebenen Treffer an der Gesamtzahl der insgesamt vorhandenen relevanten Treffer. Die Gesamtzahl der relevanten Treffer bezieht sich dabei auf die zugrunde liegende Datenbasis, also im klassischen Fall auf die gesamte Datenbank und im Web-Kontext auf alle im Web vorhandenen relevanten Seiten. Damit wird deutlich, dass sich der Recall nur sehr schwer messen lässt, im Web-Kontext ist seine Messung nicht möglich bzw. kann nur über Hilfsmethoden (wie *Pooling*, bei dem die Gesamtzahl *der von allen getesteten Systemen* zu einer Suchanfrage ausgegebenen relevanten Dokumente als Basis genommen werden) erfolgen.

Bei der Messung von Precision und Recall sollte darauf geachtet werden, dass verschiedene Systeme auf unterschiedliche Ansprüche hinsichtlich der beiden Werte

ausgerichtet sind. Die klassische Gegenüberstellung ist hier die eines Patent-Retrievalsystems und einer Web-Suchmaschine: Bei der Recherche nach Patenten ist es essenziell, kein einziges relevantes Dokument zu verpassen, während in der Websuche in den meisten Fällen der Schwerpunkt auf die Precision auf den vorderen Rängen der Trefferliste gelegt wird. Aus diesem Grund lassen sich auch Precision- und Recall-Werte zwischen Systemen, die unterschiedliche Zwecke verfolgen, nicht vergleichen. Es gibt also nicht den Mindestwert, den man von einem System erwarten können muss, sondern ein Vergleich kann höchstens mit den Systemen, die denselben Zweck erfüllen, angestellt werden. Auch bei der Verwendung ein- und desselben Systemtyps kann sich der Wunsch nach Vollständigkeit vs. Präzision der Suchergebnisse aber je nach momentaner Aufgabe unterscheiden. Su schlägt vor, diesen Wunsch in die Evaluierung mit einzubeziehen und definiert die Kennzahlen *Importance of completeness of search results* und *Importance of precision of the search to user* [47]. Damit wird versucht, das typische Nutzerverhalten in den Evaluierungsprozess einzubinden.

Sowohl für die beiden klassischen Kennzahlen als auch für ihre modernen Ableitungen gilt, dass sie Suchergebnisse stets auf Dokumentebene betrachten, wobei die einzelnen Treffer unabhängig voneinander sind. So könnte eine Suchmaschine, die zehnmal hintereinander dasselbe relevante Dokument ausgibt, eine perfekte Precision erreichen, auch wenn die Ergebnismenge als Ganzes für einen Nutzer natürlich nicht relevant wäre.

Weiterhin ist bei der Evaluierung nach Suchanfragetypen zu unterscheiden: Angelehnt an die Unterteilung von Broder [48] kann gesagt werden, dass mit Precision und Recall nur die Performance von Suchmaschinen für informationsorientierte Anfragen gemessen werden kann, also Anfragen, bei denen eine bestimmte Menge von Ergebnissen gewünscht wird und es auch mehr als ein relevantes Dokument gibt. Im Gegensatz dazu stehen navigationsorientierte Anfragen, bei denen eine bestimmte Webseite gefunden werden soll und damit nur ein relevantes Dokument vorhanden ist. Hier sind andere Kennzahlen zu wählen, beispielsweise success@n, welche angibt, auf der wievielten Position das relevante Dokument bei einer bestimmten Suchmaschine im Durchschnitt angezeigt wird (vgl. [49]).

Broders dritter Anfragetyp, transaktionsorientierte Anfragen, bei denen der Nutzer zuerst eine Website auffinden will, auf der dann eine Transaktion (bspw. Kauf eines Produkts, Download einer Datei, Spielen eines Spiels) durchgeführt wird, stellt die Evaluation vor ein weiteres Problem: Hier interagiert der Nutzer zuerst mit dem System, dann mit der gefundenen Website. Dieses Verhalten kann im klassischen Anfrage-Ergebnis-Paradigma nicht abgebildet werden und fällt in den Bereich des Interaktiven Information Retrieval (s. Abschnitt 4).

1.2.2. Weitere Kennzahlen

Im Laufe der Jahre wurden zahlreiche weitere Retrievalmaße entwickelt, sowohl für die allgemeine Evaluierung als auch speziell für den Web-Kontext. Im Web-Bereich hat sich allerdings bislang noch kein allgemeines Maß durchsetzen können; klar ist jedoch, dass besondere Maße nötig sind, um den besonderen Anforderungen an die Web-Suchmaschinen gerecht werden zu können. Insbesondere ist zu unterscheiden nach Kennzahlen, die auf die Bewertung der einzelnen Treffer eingehen und solchen, die sich auf die gesamte Trefferliste (wiederum bis zu einem bestimmten Cut-off-Wert) beziehen.

Aus dem weiten Fundus der Kennzahlen sollen hier exemplarisch einige herausgegriffen werden, die zum einen das Suchergebnis als Ganzes (also auf der Ebene der

Trefferliste) betrachten, zum anderen die Performance aller getesteten Systeme auf der Ebene der einzelnen Suchanfrage einbeziehen.

Auf der Ebene der Trefferlisten wurden unterschiedliche Kennzahlen vorgeschlagen; gemessen wird hier die Übereinstimmung zwischen dem Ranking, welches von einer Suchmaschine ausgegeben wurde, und einem Ranking, welches (auf Basis der gleichen Dokumente) von Juroren erstellt wurde [50] [51]. Zu beachten ist hier allerdings, dass auch zwischen menschlichen Juroren erhebliche Unterschiede in den Rankings bestehen können und dass nicht davon auszugehen ist, dass es das eine richtige Ranking überhaupt gibt. Insofern sind diese Maße allenfalls als Ähnlichkeitsmaße zu betrachten und eine hundertprozentige Übereinstimmung zwischen maschinellem und menschlichem Ranking ist weder erreichbar noch wünschenswert.

Kennzahlen, die das Abschneiden der anderen im Test einbezogenen Suchmaschinen in Relation zum erreichten Ergebnis einer Suchmaschine stellen, basieren auf der Feststellung, dass die Chancen, relevante Dokumente auszugeben, stark von der Suchanfrage abhängen und die Schwierigkeit je nach der Suchanfrage stark variiert. Höher bewertet werden soll nun eine Suchmaschine, die im Vergleich zu ihren Konkurrenten wesentlich besser abschneidet. Einer solchen Annahme folgt beispielsweise die Kennzahl *Salience* [52]. Noch einen Schritt weiter geht *Ability to retrieve top ranked pages* [50]: Hierbei werden von unterschiedlichen Suchmaschinen jeweils die top-gerankten Dokumente bis zu einem bestimmten Cut-off-Wert (z. B. 10) zusammengeführt und menschlichen Gutachtern zur Bewertung vorgelegt. Dann werden die von den Menschen am besten bewerteten Dokumente ausgefiltert, wobei wieder ein Cut-off festgelegt wird (bspw. 75 Prozent der Dokumente sollen in die Wertung eingehen). Letztlich wird für jede Suchmaschine berechnet, wie hoch der Anteil dieser Dokumente im Ergebnis ist.

Der in diesem Abschnitt gegebene Überblick über die Retrievalmaße verfolgte vor allem das Ziel, für die Auswahl geeigneter Maße zu sensibilisieren und strebte nicht an, eine vollständige Übersicht zu geben.[5] Wie sich noch in den weiteren Abschnitten dieses Kapitels zeigen wird, ist die Auswahl geeigneter Kennzahlen von großer Bedeutung, da häufig mit vermeintlich objektiven Kennzahlen gemessen wird, die allerdings etwas anderes messen als das, was in dem entsprechenden Test eigentlich angestrebt war.

2. Grenzen der Standardverfahren

Die in den vorangegangenen Abschnitten beschriebenen Standardverfahren gehen, wenn auch durch die Modifikation der Tests gewisse Anpassungen erreicht wurden, von einem Nutzer aus, der die Treffer nacheinander durchgeht, sich aufgrund der Trefferbeschreibung entscheidet, welchen Treffer er auswählt, sich aber nicht von einer hervorgehobenen Präsentation der Treffer beeinflussen lässt oder seine Anfrage aufgrund der ausgegebenen Treffermenge (bzw. der gesichteten Treffer) verändert. Zwar bildet insbesondere die Einbeziehung der Trefferbeschreibungen eine wesentliche Verbesserung gegenüber den rein auf die Treffer selbst abzielenden Evaluationen, jedoch kann weiterhin kaum von einer realistischen Abbildung des Nutzerverhaltens gesprochen werden. Ob sich eine solche überhaupt im Rahmen praktikabler Tests

[5] Ein guter Überblick über die klassischen Maße findet sich in dem Lehrbuch von Korfhage [69]; eine Vollständigkeit anstrebende Auflistung von Retrievalmaßen findet sich in [70].

erreichen lässt, sei hier dahingestellt. Grundsätzlich wird damit auch nicht die Validität der Tests infrage gestellt, sondern es soll hier verdeutlicht werden, dass die Testergebnisse nur für einen Teil der Interaktionen mit Suchmaschinen Gültigkeit beanspruchen können.

Die beschriebenen Standardverfahren finden ihre Grenzen einerseits in ihrer Unfähigkeit, den oftmals interaktiven und in mehreren Schritten verlaufenden Prozess der Recherche abzubilden. Dieses Problem ist aus der Information-Retrieval-Evaluierung bekannt und betrifft nicht allein die Evaluierung von Suchmaschinen. Die Herausforderung besteht hier darin, die eher systemorientierten Tests (welche hier behandelt werden) mit den nutzerorientierten Tests (in der Regel qualitative Untersuchungen) zusammenzuführen.

Eine weitere Beschränkung der Standardverfahren ist die Annahme, dass die Ergebnisqualität wesentlich für die Bevorzugung eines bestimmten Informationssystems vor anderen Systemen verantwortlich ist. Zwar lassen sich unterschiedliche Systeme mit den Verfahren vergleichen, eine mögliche Wechselbereitschaft der Nutzer kann aus den Ergebnissen jedoch nicht abgeleitet werden. Auch hier unterscheidet sich wieder die Evaluation von Systemen zum Zweck der Auswahl eines bestimmten Systems für den eigenen Zweck oder zum Vergleich des eigenen Systems mit anderen Systemen von der Evaluation von Web-Suchmaschinen, die sich ja gerade dadurch auszeichnet, dass sie mehrere fremde, in ihren Prozessen nicht zugängliche Systeme unterscheidet. Zwar mögen hier Empfehlungen für oder gegen die Nutzung einer bestimmten Suchmaschine ausgesprochen werden, in der Praxis dürften jedoch auch oder gar vor allem andere, nicht direkt der Ergebnisqualität zurechenbare Faktoren für die Wahl einer Suchmaschine ausschlaggebend sein. Zu denken ist hier beispielsweise an Zusatzangebote (wie E-Mail-Dienste und Office-Anwendungen). Im Suchmaschinenbereich zeigt sich auch eine starke Markenbindung; Untersuchungen haben gezeigt, dass Nutzer ihre Lieblingssuchmaschine bevorzugen, wenn die zu evaluierenden Ergebnisse als dieser Suchmaschine zugehörig markiert sind [53].

Im Folgenden soll dargestellt werden, welche Faktoren die Qualität der Testergebnisse beeinflussen können. Das Ziel dieser Darstellung ist es, durch eine bessere Berücksichtigung der genannten Faktoren die Aussagekraft der Tests zu verbessern.

2.1. Anfragetypen

Bereits bei der Darstellung der Kennzahlen zur Messung der Retrievaleffektivität wurde kurz auf die unterschiedlichen Anfragetypen eingegangen. Wenn nun beim Testdesign berücksichtigt wird, mit welchem Typ von Anfragen getestet werden soll, so lassen sich belastbare Ergebnisse erreichen. Um generelle Aussagen über die Qualität der Ergebnisse einer Suchmaschine zu erreichen, ist die Verwendung nur eines Anfragetyps jedoch nicht ausreichend, da sich Suchmaschinen ja gerade dadurch auszeichnen, dass sie alle Anfragetypen bedienen können und auch die Nutzerschaft davon ausgeht, dass sie dazu in der Lage sind [54]. Im Gegensatz dazu sind fachliche Informationssysteme in aller Regel auf einen Anfragetyp ausgerichtet (zur Befriedigung eines konkreten Informationsbedarfs auf Faktenfragen; zur Befriedigung eines problemorientierten Informationsbedarfs auf thematische Fragen), was die Evaluierung erheblich erleichtert.

Geht man davon aus, dass die drei Anfragetypen informationsorientiert, navigationsorientiert und transaktionsorientiert jeweils einen nennenswerten Anteil der an die allgemeinen Suchmaschinen gestellten Anfragen ausmachen [54], so lässt sich aus der Evaluierung, welche nur Anfragen eines Typs verwendet, keine allgemeine Aussage

über die Qualität der Treffer der Suchmaschinen ableiten. Die meisten Suchmaschinentests beziehen sich auf informationsorientierte Anfragen und lassen damit eine wesentliche Aufgabe der Suchmaschinen, nämlich ihre Nutzer direkt auf vorher bereits bekannte Webseiten zu führen, außer Betracht.

2.2. Elemente der Trefferseiten und ihre Darstellung

Die gängigen Tests zu Retrievaleffektivität von Suchmaschinen berücksichtigen nur die sog. organischen Ergebnisse, d.h. die reguläre Trefferliste. Unberücksichtigt bleiben die weiteren Elemente der Ergebnisseite, vor allem die Textanzeigen, sog. Shortcuts und Treffer aus weiteren Kollektionen (ein Überblick der Elemente der Trefferseiten findet sich in [55]).

Die Darstellung von Suchergebnisseiten als Zusammenstellung unterschiedlicher Elemente (gegenüber der konventionellen Listendarstellung) war zuerst noch ein Phänomen, das nur bei populären Suchanfragen oder solchen mit besonders aktuellem Bezug auftauchte. Welches Ausmaß diese alternative Art der Ergebnispräsentation inzwischen allerdings angenommen hat bzw. welches Potenzial in ihr steckt, zeigt eine Aussage von Shashi Seth, dem Vice President Search Products bei Yahoo. Demnach können etwa 50 Prozent aller Suchanfragen entweder mit direkten Antworten oder hervorgehobenen Hinweisen auf Treffer aus anderen Kollektionen (sog. Shortcuts) beantwortet werden [56]. Dies wirft die Frage auf, inwieweit die heutige Web-Suche überhaupt noch als Web-Suche bezeichnet werden kann, wenn man unter diesem Begriff die Abfrage eines einzigen Web-Index, in dem alle für die Suche relevanten Dokumente gespeichert sind, versteht.

In der Testkonzeption ist stets zu berücksichtigen, welche Elemente für den jeweiligen Test von Bedeutung sind. So kann man beispielsweise bei einem Test, in dem exklusiv Anfragen wissenschaftlicher Natur verwendet werden, davon ausgegangen werden, dass nur in wenigen Fällen Textanzeigen eingeblendet werden und dass die Nutzergruppe Wissenschaftler in der Lage ist, diese Anzeigen ggf. von den organischen Treffern unterscheiden zu können. Hier würde sich wohl der Aufwand, die Textanzeigen zusätzlich als Ergebnisse zu erfassen und auswerten zu lassen, nicht lohnen.

Im Folgenden werden die wichtigsten auf den Ergebnisseiten der gängigen Web-Suchmaschinen auftauchenden Elemente beschrieben und ihre Bedeutung für die Trefferevaluation herausgestellt:

- *Textanzeigen* sind auf den Trefferseiten der Suchmaschinen prominent platziert und erreichen gute Klickraten. Neben den von den Werbetreibenden gebotenen Beträgen für Klicks auf die entsprechende Anzeige beeinflussen auch Relevanzkriterien die Platzierung der Anzeigen. Da die Anzeigen außerdem auf die eingegebenen Suchbegriffe angepasst sind, ist zumindest von einer gewissen Relevanz auch der bezahlten Werbeplätze für die Anfrage auszugehen. Und obwohl die Anzeigen von den Suchmaschinenbetreibern mehr oder weniger deutlich als solche kenntlich gemacht werden, ist davon auszugehen, dass ein hoher Anteil der Suchmaschinennutzer nicht zwischen den Werbetreffern und den organischen Treffern unterscheiden kann.
- Ein weiteres zu berücksichtigendes Element der Trefferseiten sind sog. *Shortcuts*, also hervorgehobene, besonders aufbereitete Treffer innerhalb der regulären Trefferliste. Solche Shortcuts beantworten oft schon die gestellte Anfrage entweder durch eine direkte Antwort (Bsp. Wettervorhersage) oder durch die Weiterleitung auf eine sog. Autorität (bspw. bei der Eingabe einer Wertpapierkennnummer).

Abbildung 1: Typische Elemente der Ergebnispräsentation

- Mit *Treffern aus weiteren Kollektionen* sind Treffer gemeint, die nicht direkt aus dem Web-Index einer Suchmaschine kommen, jedoch im Rahmen einer sog. „Universal Search" (vgl. [57]) in die reguläre Trefferliste eingebunden werden. Beispiele hierfür sind Bilder aus dem Bilderindex oder Nachrichten aus einem speziellen Nachrichtenindex.

Die Frage, welche Ergebnisse bei der Evaluierung von Suchmaschinen zu berücksichtigen sind, stellt sich noch auf einer anderen Ebene. In den bekannten Untersuchungen werden meist die ersten zehn oder zwanzig Treffer der organischen Trefferliste berücksichtigt, wobei der Annahme gefolgt wird, dass Nutzer nur diese Ergebnisse der ersten (oder eben der ersten und zweiten) Ergebnisseite ansehen würden. Verschiedene Untersuchungen haben jedoch gezeigt, dass Nutzer sich in einem sehr hohen Maß auf die ersten angezeigten Ergebnisse verlassen und die Reihung der Ergebnisse einen hohen Einfluss auf die Klickraten hat [46][58]. So erreichen Ergebnisse, die auf dem Ergebnisbildschirm sichtbar sind, ohne dass auf der Seite gescrollt werden muss, weit höhere Klickraten als die Ergebnisse, die „unter dem Knick" stehen. Man unterscheidet daher den sichtbaren und den unsichtbaren Bereich der Ergebnisseite [55]. Abbildung 1 zeigt die typischen Elemente einer Ergebnisseite am Beispiel von Google.

Weiteren Einfluss auf die Klickraten haben die Trefferbeschreibungen. Treffer, deren Beschreibung nicht auf einen relevanten Treffer hinweist, werden kaum angesehen. Es konnte gezeigt werden, dass Suchmaschinen nur eingeschränkt in der Lage sind, relevante Treffer auch mit relevanten Beschreibungen zu versehen. Daraus ergibt sich die Situation, dass Nutzer durch ihre Beschreibung als relevant erscheinende Treffer anklicken und enttäuscht werden oder aber, dass relevante Treffer aufgrund ihrer als nicht relevant erscheinenden Beschreibungen außer Acht gelassen werden [4]. Aus diesem Grund erscheint die Berücksichtigung der Trefferbeschreibungen in Suchmaschinen-Retrievaltests als unbedingt notwendig.

2.3. Relevanzbewertungen

Die Bewertung der Relevanz der Treffer bzw. die Art dieser Bewertung spielt selbstverständlich bei allen Retrievaltests eine Rolle und ist kein exklusives Problem der Suchmaschinenevaluierung. Jedoch ergibt sich in diesem Bereich eine Besonderheit dadurch, dass im Web zu vielen nachgefragten Themen eine Menge grundsätzlich relevanter Treffer vorhanden sind und Juroren generell recht leicht bereit sind, ein Dokument als relevant einzuschätzen. Erst durch eine differenzierte Relevanzbewertung lassen sich oft Differenzen zwischen den Suchmaschinen feststellen. Zusammengefasst lässt sich sagen, dass es für die gängigen Web-Suchmaschinen weniger ein Problem ist, irgendwie relevante Treffer anzuzeigen als aus dieser Menge der relevanten Treffer die tatsächlich *hoch relevanten* zu identifizieren.

In den bekannten Retrievaltests wurden allerdings meist binäre Bewertungen verwendet (also die Unterscheidung zwischen relevanten und nicht relevanten Treffern), seltener wurden Skalen verwendet, die eine differenziertere Bewertung möglich machen. Da in Suchmaschinen in den meisten Fällen nach einigen hoch relevanten Dokumenten gesucht wird und weniger (wie teils in anderen Informationssystemen) nach einer vollständigen Treffermenge, ist diese Abstufung hier von besonderer Bedeutung. Mehr noch: Es zeigt sich, dass die bei einer binären Bewertung oft nur graduellen Unterschiede zwischen den populären Suchmaschinen bei einer Skalenbewertung deutlicher werden.

Die Bewertung der Relevanz erfolgt in der Regel anhand der einzelnen ausgegebenen Treffer. Zumindest zusätzlich erscheint jedoch eine Bewertung der Trefferliste (bzw. des sichtbaren Bildschirmbereichs wie oben beschrieben) sinnvoll, da diese für den Gesamteindruck von einer Suchmaschine entscheidender sein dürfte als ein oder mehrere angeklickte Treffer, deren Relevanz bei verschiedenen Suchmaschinen ja ähnlich sein kann.

Ein weiteres Problem ergibt sich aus der Entscheidung, wie viele Dokumente überhaupt in die Relevanzbewertung eingehen. Hier ist es wichtig zu erfassen, wann ein Nutzer die Suche abbrechen würde, entweder auf Basis der gesamt ausgegebenen Ergebnisse („Wenn bis Trefferplatz drei nichts Relevantes dabei ist, wird die Suche abgebrochen oder modifiziert“) oder auf Basis der angesehenen Ergebnisse (der Nutzer ist bereit, eine bestimmte Anzahl von Ergebnissen anzusehen).

Weiterhin wird in den bekannten Tests nicht erfasst, inwieweit die Ergebnisse dem Anspruch der Nutzer nach einer Vielfalt der Ergebnisse entsprechen. Gerade weil Nutzer nur wenige Positionen der Trefferliste überhaupt in Betracht ziehen, ist auf diesen Positionen eine gewisse Vielfalt gefragt. Zwar mag eine Suchmaschine auf den ersten drei Positionen nur relevante Treffer ausgeben, wenn es sich dabei aber jeweils um ähnliche Treffer (im inhaltlichen Sinne oder hinsichtlich des Dokumenttyps) handelt, wird die Bewertung durch den Nutzer in der Regel schlecht ausfallen.

2.4. Durchmischung der Trefferlisten

Um den unterschiedlichen Bedürfnissen verschiedener Nutzer, welche eine identische Suchanfrage abschicken, gerecht zu werden, haben Suchmaschinenbetreiber früh erkannt, dass es hilfreich ist, die Trefferlisten zu durchmischen: Einerseits sollen Ergebnisse unterschiedlichen Typs (Nachrichten, Blogs, offizielle Homepages) gezeigt werden, andererseits Ergebnisse aus unterschiedlichen Medienkollektionen (Bilder, Videos). Unumstritten ist, dass durch solche Durchmischungen die Qualität der ausgegebenen Ergebnisse verbessert werden kann [57]. Allerdings liegen bisher noch keine Modelle für die Evaluierung vor, die die Kombination der verschiedenen Ergebnistypen berücksichtigen.

Eine Durchmischung der Trefferlisten soll auch bedeuten, dass auf den vorderen Plätzen Dokumente präsentiert werden, die unterschiedliche bzw. im Idealfall alle Aspekte des gesuchten Themas abdecken. Zur Evaluierung der Suchmaschinen hinsichtlich dieser Fähigkeiten kann als Ausgangsbasis auf die Erkenntnisse aus dem TREC Diversity Track (s. [59]) und die dort verwendeten Maßzahlen [60][61] zurückgegriffen werden. Untersuchungen im Rahmen von Tests der gängigen Web-Suchmaschinen stehen aber noch aus.

2.5. Such-Sessions

Als allgemeine Schwäche von Studien zur Retrievaleffektivität von Suchmaschinen kann festgestellt werden, dass sich diese in der Regel auf ein einfaches Such-Modell beziehen: Ein Nutzer gibt eine Anfrage ein, bekommt eine Trefferliste angezeigt und selektiert aus dieser die ihm geeignet erscheinenden Ergebnisse. Nicht berücksichtigt wird hier, dass im Suchprozess oft Zwischenschritte, etwa eine Reformulierung der Suchanfrage oder die Nutzung von Filterelementen auf der Ergebnisseite, stattfinden.

Eine besondere Schwierigkeit ergibt sich dadurch, dass die Suchmaschinennutzung nicht klar in eine Richtung weist – auch hier zeigt sich wieder die Vielfältigkeit der Suchmaschinennutzung: Auf der einen Seite werden zuhauf Anfragen gestellt, bei denen nach Ansicht der Trefferseite keine Reformulierung erfolgt (die also dem alten Anfrage-

Ergebnis-Paradigma entsprechen), auf der anderen Seite ergeben sich (teils sehr komplexe) Such-Sessions, die in den konventionellen Tests nicht abgebildet werden können.

Spink und Jansen [33][35] konnten in Logfile-Untersuchungen zeigen, dass die Such-Sessions bei Web-Suchmaschinen in der Regel kurz sind und im Verlauf einer Session nur wenige Dokumente gesichtet werden. Allerdings handelt es sich hier um Durchschnittswerte und die Varianz der Sessionlängen ist sehr hoch. Noch weiter als die Betrachtung von Sessions gehen sog. explorative Suchen (*exploratory searches*; [62]), in welchen teils über Tage oder Wochen hinweg immer wieder an einer Suchaufgabe gearbeitet wird (wie zum Beispiel an der Planung einer Urlaubsreise). Die Schwierigkeiten, die sich bei der Evaluierung komplexerer Interaktionen mit Suchmaschinen ergeben, sind evident.

Erste Untersuchungen weisen allerdings darauf hin, dass sich die Zufriedenheit der Nutzer mit den Ergebnissen einer Session, welche Umformulierungen der Suchanfrage beinhalten kann, durch die Bewertung der Trefferliste der initialen Suchanfrage voraussagen lässt [63]. Inwieweit sich solche Voraussagen auch unter den Bedingungen der veränderten Trefferpräsentation treffen lassen, ist jedoch noch unklar. Für explorative Suchen jedenfalls dürften solche Voraussagen nicht zu treffen sein.

3. Messung der Retrievaleffektivität mittels Klickdaten

Aus der generellen Vorgehensweise bei der Durchführung von Retrievaltests wurde bereits deutlich, dass es sich dabei um sehr arbeitsintensive Testverfahren handelt, bei denen eine große Zahl von Juroren gewonnen werden muss, auch wenn die Tests selbst dann nur relativ wenige Suchanfragen, vor allem im Vergleich zu der schieren Masse verschiedener Suchanfragen, die tatsächlich an Suchmaschinen gestellt werden, abdecken können. Insofern ist es nicht verwunderlich, dass nach Verfahren gesucht wurde, die auf der einen Seite weniger arbeitsintensiv sind und auf der anderen Seite eine Masse von Suchanfragen und auch Bewertern berücksichtigen können.

Eine Lösung liegt in der Verwendung von Interaktionsdaten echter Nutzer mit Suchmaschinen. Diese Verfahren basieren auf den Daten aus den Logfiles der Suchmaschinen. Zur Messung herangezogen wird zwar auch hier nur ein Ausschnitt aus den vorhandenen Daten, es ist aber beispielsweise möglich, sämtliche Suchanfragen eines Monats zur Analyse heranzuziehen, wobei es sich mitunter um Millionen von Suchanfragen handeln kann. Die Vorteile eines solchen Verfahrens liegen auf der Hand: Es werden die echten Suchanfragen echter Nutzer und ihr Verhalten auf den Trefferseiten ausgewertet, dazu kommt, dass man die Suchanfragen eines bestimmten Zeitraums vollständig abdecken kann.

Solche auf den Klickdaten der Nutzer basierenden Verfahren berücksichtigen die Suchanfragen selbst, die auf der Trefferseite ausgewählten Ergebnisse sowie ggf. die Zeit, in der sich ein Nutzer ein Suchergebnis ansieht, bevor er auf die Trefferseite zurückkehrt. Gemessen werden auf der einen Seite die Klicks auf einen bestimmten Treffer, auf der anderen Seite die sog. *dwell time*, also die Verweildauer auf dem Ergebnis. Handelt es sich nur um eine sehr kurze Verweildauer, nach der der Nutzer auf die Trefferseite zurückkehrt, um dort einen anderen Treffer auszuwählen, so spricht man auch von der *bounce rate*. Gemessen wird damit der Anteil derjenigen Nutzer, die ohne genaue Durchsicht des Treffers auf die Ergebnisseite zurückkehrt.

In Klick-Tests kann gemessen werden, ob ein bestimmtes Ergebnis mehr Klicks erhält als ein vor diesem Ergebnis angezeigter Treffer. Dies wird als klarer Indikator

dafür gesehen, dass das Ranking verbessert werden kann, indem der häufiger geklickte Treffer weiter oben angezeigt wird. Zu berücksichtigen ist hier allerdings auch die Verweildauer auf dem Treffer bzw. die bounce rate, da genau genommen ein Klick in der Trefferliste ja nur bedeutet, dass die Trefferbeschreibung auf einen relevanten Treffer hindeutet. In Kombination mit den Kennzahlen zur Verweildauer lassen sich jedoch erhebliche Verbesserungen des Rankings erreichen.

Da solche Tests nicht mit Juroren arbeiten, sondern die Daten der echten Nutzer einer Suchmaschine aufzeichnen, können auch viele Bewertungen zu einem einzigen Treffer gesammelt werden [20]. Allerdings ist zu bedenken, dass bei selten gestellten Anfragen nicht unbedingt viele Klicks auf die Trefferdokumente vorliegen. Weiterhin werden in solchen Tests keine *expliziten* Relevanzurteile erfasst, sondern *implizite*. Es bleibt unklar, ob Nutzer tatsächlich auf dem besten Treffer verweilt haben oder ob sie sich schlicht mit dem aufgrund des Ranking der Suchmaschine als am besten angesehenen Treffer zufriedengegeben haben. In solchen Tests findet also keine systematische Evaluierung einer vorher bestimmten Treffermenge statt, da von den Nutzern, wie oben dargestellt, nur die ersten bzw. besonders hervorgehobenen Treffer überhaupt angeklickt werden.

Während klickbasierte Tests unbestreitbare Vorteile haben, ist die Durchführung eines solchen Tests nur möglich, wenn man Zugriff auf die tatsächlich bei der Suchmaschine anfallenden Daten hat. Dadurch ist der Personenkreis der Testdurchführenden auf die Suchmaschinenbetreiber selbst und mit diesen kooperierende Institutionen eingeschränkt. Außerdem ist mit diesen Verfahren ein Vergleich zwischen verschiedenen Systemen nur theoretisch möglich, da die Suchmaschinenbetreiber kaum ihre Daten für solche Zwecke zur Verfügung stellen dürften. Insofern kann man nur empfehlen, die aus den entsprechenden Tests gewonnenen Erkenntnisse zu beachten und mit eigenen, Juroren-basierten Tests zu kombinieren. Auch, wenn man einzig die Ergebnisqualität seines eigenen Systems testen möchte, sollte man ein kombiniertes Verfahren anwenden, da die alleinige Analyse des Klickverhaltens zu wenig aussagekräftig ist.

4. Evaluierung im Interaktiven Information Retrieval

Während die bislang beschriebenen Verfahren die Qualität der Suchergebnisse entweder auf der Ebene der Treffer oder aber auf der Ebene der Trefferlisten messen, gerät zunehmend die sessionbasierte Evaluierung in den Fokus. Auf der einen Seite können hier im Rahmen von Klicktests die Ergebnisse, die im Verlauf einer Session gesichtet wurden, aufgrund der Verweildauern und des Zeitpunkts des Abrufs bewertet werden. Andererseits können sessionbasierte Evaluierungen auch konkret auf eine weitergehende Nutzerbeobachtung bezogen werden. Dabei wird das Verhalten ausgewählter Nutzer protokolliert, wobei es auch möglich ist, deren Suche mit unterschiedlichen Suchwerkzeugen (Suchmaschinen, Verzeichnisse, Wikipedia, Soziale Netzwerke, usw.) abzubilden. Der große Vorteil eines solchen Verfahrens liegt darin, dass Nutzer, die einer solchen Untersuchung zugestimmt haben, auch über längere Zeiträume beobachtet werden können, dass zusätzliche Daten wie Alter, Geschlecht, usw. der Nutzer abgefragt werden können und dass die Untersuchung um weitergehende Befragungen ergänzt werden kann. Solche Untersuchungen können damit sowohl quantitative als auch qualitative Daten einbeziehen, indem etwa ergänzend zur Protokollierung des tatsächlichen Suchverhaltens einerseits nach expliziten Relevanzbewertungen zu den angesehenen Treffern und andererseits nach der Motivation für das Suchverhalten gefragt wird.

Der Nachteil der Evaluierung solcher interaktiver Szenarien liegt darin, dass zumindest, wenn den Nutzern die Wahl der Suchmaschine(n) freigestellt wird, nur unzureichend Vergleichsdaten anfallen und Ergebnismengen nicht systematisch evaluiert werden können, da ja den Nutzern nicht ein vorgegebenes Set von Ergebnissen zur Bewertung vorgelegt wird. Auch lassen sich Markeneffekte und andere Präferenzen nur schwer ausschließen. Und nicht zuletzt sind Probanden für solche Tests schwieriger zu gewinnen, da der Test entweder im Labor durchgeführt werden muss oder der Nutzer eine besondere Software zur Protokollierung installieren muss. Nichtsdestotrotz stellen solche Untersuchungen zumindest eine sinnvolle Ergänzung zu den konventionellen Retrievaltests dar; für bestimmte Fragestellungen sind sie auch alleine geeignet.

Für die sessionbasierte Evaluierung stehen verschiedene Werkzeuge zur Datenerhebung kostenlos zur Verfügung. Zu nennen sind hier Search-Logger [64], der HCI Browser [65] und Wrapper [66]. Auch wenn diese Tools nicht zuvorderst zur Messung der Retrievaleffektivität erstellt wurden und daher etwa keine Auswertungsmodule für die Trefferqualität enthalten, sind sie zumindest ein unkompliziertes Mittel zur Erfassung der Daten für solche Untersuchungen. Leider gibt es jedoch bislang keine Tools, die beide Testarten zusammenführen.

5. Empfehlungen für die Praxis

Die vorangegangene Darstellung der verschiedenen Testverfahren und der zahlreichen Entscheidungen, die beim Aufbau eines Tests getroffen werden müssen, mögen zunächst einmal den Eindruck hinterlassen, dass valide Tests in der Praxis nur schwer durchgeführt werden können. In der Tat ist zu beobachten, dass solche Tests oft minderkomplex sind und daher die damit erzielten Ergebnisse angezweifelt werden können. Dies ist umso bedauerlicher, wenn man den Aufwand für die Testdurchführung mit Juroren in Betracht zieht.

Belastbare Testergebnisse lassen sich jedoch auch mit vertretbarem Aufwand erreichen. Bedingung ist allerdings, dass die Untersuchungsfrage klar benannt wurde und das Testdesign entsprechend ausgerichtet wurde. Wie auch in anderen Kontexten kann es auch hier sinnvoll sein, lieber einige Kernfragen genau zu behandeln als zu versuchen, einen alle Fragen umfassenden Test zu erstellen, der hinterher nur zu fragwürdigen Ergebnissen führt.

Die folgenden Empfehlungen beinhalten bei Weitem nicht alle im Text genannten Punkte und stellen daher zwar keinen idealen Test dar, wenn man jedoch den nötigen Aufwand betrachtet, wird ersichtlich, dass dieser schon weit über die üblichen Tests hinausgeht. Es wurde damit versucht, eine Abwägung zu schaffen zwischen möglichst perfektem Testdesign und dem in der Praxis Machbaren.

Zusammengefasst sollten folgende Punkte berücksichtigt werden:

1. Der Test wird nach Anfragetypen unterteilt. Eine Vermischung der Anfragetypen ist unzulässig und verfälscht die zu messenden Kennwerte. Aufgrund der Häufigkeit ihres Vorkommens sollen informationsorientierte sowie navigationsorientierte Anfragen verwendet werden.
2. Berücksichtigung finden alle Elemente der ersten Ergebnisseite. Anhand der bekannten Bildschirmauflösungen und ihrer Verbreitung wird der jeweils sichtbare Bereich der Ergebnisseite ermittelt. Die Elemente finden unabhängig von ihrer Zuordnung zu den organischen Treffern Eingang in die Untersuchung.

3. Die verwendeten Kennzahlen richten sich nach den Anfragetypen. Für navigationsorientierte Anfragen werden Erfolgsraten berechnet, für informationsorientierte Anfragen Relevanzwerte.
4. Die Relevanzbewertungen erfolgen auf einer differenzierenden Skala. Es werden sowohl die einzelnen Treffer als auch das Ergebnisset im Gesamten bewertet.
5. Sowohl die Trefferbeschreibungen als auch die Treffer selbst werden bewertet.
6. Die Ergebnisse werden nach Dokumenttypen klassifiziert, um eine zusätzliche Messung der Vielfalt der Ergebnisse zu ermöglichen.

Diese Empfehlungen beziehen sich auf die Gestaltung von Retrievaltests mit Juroren, da diese Form des Tests als die praktikabelste und technisch am unaufwändigsten anzusehen ist. Allerdings ist auch anzumerken, dass für diese Testform keine frei verfügbaren Werkzeuge verfügbar sind, die dabei helfen könnten, den Aufwand zu minimieren.

In Tabelle 1 sind wird der Retrievaleffektivitätstest mit Juroren noch einem den anderen beiden Testverfahren gegenübergestellt. Alle Verfahren werden hinsichtlich ihrer Eignung für praktische Tests bewertet.

Tabelle 1.: Vergleich der Testverfahren.

Testverfahren	**Anwendungsfall**	**Bewertete Dokumente**	**Bewertung**
Retrievaleffektivitätstest	Überprüfung des eigenen Systems Vergleich eines eigenen Systems mit fremden Systemen Vergleich fremder Systeme untereinander	alle bis zu einem bestimmten Cut-off-Wert	Geeignet, wenn explizite Bewertungen zu einer vorher bestimmten Trefferanzahl ausgewertet werden sollen. Einzige Möglichkeit, wenn vollständige Treffermengen (bis zu einem bestimmten Cut-off-Wert) beurteilt werden sollen.
Klicktest	Überprüfung des eigenen Systems	von tatsächlichen Nutzern angeklickte Dokumente	Gut geeignet, um Massendaten zu analysieren und automatische Verbesserungen am Ranking des eigenen Systems durchzuführen.
Protokollbasierter Test	Überprüfung des eigenen Systems Vergleich eines eigenen Systems mit fremden Systemen Vergleich fremder Systeme untereinander	von den Nutzern im Test angeklickte Dokumente	Geeignet, um Sessions oder auch explorative Suchen abzubilden. Geeignet, um ausgewählte Nutzer bei der Interaktion mit (und dem Wechsel zwischen) mehreren Systemen zu beobachten.

Während die Verfahren in diesem Text im Kontext der Websuche diskutiert wurden, ergibt sich in der Praxis natürlich die Frage, inwieweit die Erkenntnisse auch auf die Evaluation anderer Suchsysteme übertragen werden können, beispielsweise auf die Suche in einem Intranet oder innerhalb einer Webpräsenz.

Eine grundsätzliche Übertragbarkeit ist gegeben, auch wenn je nach Anwendungsfall einige Modifikationen vorzunehmen sind. Bei der Evaluation einer Intranet-Suche ergibt sich das Problem, dass kein direkter Vergleich mit anderen Suchsystemen möglich

ist, es sei denn, man kann Testinstallationen verschiedener Anbieter miteinander vergleichen. Ähnlich sieht es bei der Suche innerhalb einer Webpräsenz (bspw. eines Nachrichtenportals) aus; auch hier lässt sich wieder ein direkter Vergleich ziehen. Zwar können natürlich vergleichbare Portale vorhanden sein, diese arbeiten jedoch jeweils mit exklusiven Datenbeständen. Damit wird der Einfluss des Indexes auf die Ergebnisbewertung zu groß.

6. Fazit

In diesem Kapitel wurde ein Überblick über Verfahren zur Bewertung der Trefferqualität von Suchmaschinen gegeben. Dabei hat sich gezeigt, dass es mit Standardverfahren möglich ist, eine valide Evaluation durchführen, sofern es der Anwendungskontext erlaubt. In Bezug auf die Websuche haben sich allerdings noch keine Standardverfahren etabliert, sondern man modifiziert die aus anderen Kontexten bekannten Verfahren. Da die jeweils durchgeführten Modifikationen aber je nach Test unterschiedlich sind, lassen sich die Tests (und ihre Ergebnisse) nur schwer vergleichen. Hier ist für die Zukunft zu hoffen, dass sich entsprechende Standards herausbilden werden.

Es wurde gezeigt, dass die gängigen Tests zur Messung der Retrievaleffektivität durch eine stärkere Orientierung am Verhalten der Suchmaschinennutzer verbessert werden können. Einige der genannten Elemente wurden bereits in für sich stehenden Studien untersucht, jedoch fehlt bislang ein Rahmen, der alle genannten Elemente zusammenführt und in einem Test prüfbar macht.

In der Praxis der Testdurchführung ergeben sich Schwierigkeiten, da zwischen einem optimalen Testdesign und dem für einen Test vertretbaren Aufwand abgewogen werden muss. Deutlich wurde jedoch einerseits, dass eine Evaluierung sowohl im Prozess der Erstellung eines Suchsystems (formativ) als auch die Evaluierung bestehender Systeme (summativ) sinnvoll ist. Außerdem führt ein Test, welcher nicht den skizzierten Idealbedingungen entspricht, nicht notwendigerweise zu ungültigen Ergebnissen. Vielmehr lassen sich mit vertretbarem Aufwand Tests erstellen, die sorgfältig gestellte Fragen beantworten können.

Zum Abschluss soll noch ein Ausblick auf weitere Forschungen zur Trefferqualität von Suchmaschinen gegeben werden. Bereits deutlich wurde, dass zukünftige Tests wesentlich stärker auf das Verhalten der Suchmaschinennutzer hin angepasst werden müssen. Zum anderen ist aber auch die Tendenz festzustellen, nicht mehr nur nach der Relevanz der Ergebnisse zu fragen, sondern beispielsweise auch nach ihrer Intention. Damit ist in diesem Fall gemeint, dass bspw. durch Verfahren der Suchmaschinenoptimierung platzierte Ergebnisse bevorzugt kommerzielle Treffer auf die vorderen Ränge der Trefferlisten bringen können [67].

Diese Diskussion führt bereits in Richtung der weitergehenden Diskussion um die Verantwortlichkeit der Suchmaschinenbetreiber für die Neutralität ihrer Ergebnisse und eine gerechte Repräsentation der unterschiedlichen Akteure im Web.

Selbst wenn eine einzige Suchmaschine ihren Konkurrenten in Bezug auf die Trefferqualität so weit überlegen wäre, dass sie für alle Suchanfragen die besten Ergebnisse liefern würde, so wäre doch eine größere Vielfalt auf dem Suchmaschinenmarkt zu wünschen. Die Diskussion um die Qualität der Suchergebnisse unterliegt leider in weiten Teilen dem Irrtum, dass es „die einzig richtigen“ Ergebnisse zu einer Suchanfrage gibt und die unterschiedlichen Suchmaschinen dem Ziel, genau diese Ergebnisse in der richtigen Reihung anzuzeigen, nur mehr oder weniger nahe kommen. Dabei lassen sich

zumindest informationsorientierte Suchanfragen auf vielfältige Weise beantworten. Und nicht zuletzt kann eine Suchmaschine, die aus vielen relevanten Ergebnissen zu einer Suchanfrage auswählen kann, bestimmte Ergebnisse willentlich bevorzugen. Dieser unter dem Label „Search Engine Bias" geführte Bereich wird in der letzten Zeit verstärkt diskutiert, vor allem in Bezug auf die Bevorzugung eigener Angebote durch die Suchmaschinen Google [68].[6]

Literatur

[1] B. van Eimeren & B. Frees, Der Internetnutzer 2009 – multimedial und total vernetzt? Ergebnisse der ARD/ZDF-Onlinestudie 2009, *media perspektiven* **7**, 2009, 334-348.

[2] ComScore, comScore Reports Global Search Market Growth of 46 Percent in 2009, http://comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009, 2010.

[3] J. Brophy & D. Bawden, Is Google enough? Comparison of an internet search engine with academic library resources, *Aslib Proceedings* **57**, 2005, 498-512.

[4] D. Lewandowski, The retrieval effectiveness of web search engines: considering results descriptions, *Journal of Documentation* **64**, 2008, 915-937.

[5] W. Tawileh, J. Griesbaum & T. Mandl, *Evaluation of five web search engines in Arabic language*, *Proceedings of LWA2010*, Kassel, Germany, http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir1.pdf\, 2010.

[6] J. Véronis, A comparative study of six search engines, http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf, 2006

[7] Webhits, Webhits Web-Barometer, http://www.webhits.de/deutsch/index.shtml?webstats.html, 2011.

[8] D. Lewandowski & N. Höchstötter, Qualitätsmessung bei Suchmaschinen – System- und nutzerbezogene Evaluationsmaße, *Informatik-Spektrum* **30**, Juni 2007, 159-169.

[9] M. Machill, Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen, *Wegweiser im Netz*, M. Machill & C. Welp (Hrsg.), Bertelsmann Stiftung, Gütersloh, 2003, 13-490.

[10] N. Schmidt-Maenz & C. Bomhardt, Wie suchen Onliner im Internet?, *Science Factory/Absatzwirtschaft* **2**, 2005, 5-8.

[11] A. Spink, B. J. Jansen, C. Blakely & S. Koshman, A study of results overlap and uniqueness among major Web search engines, *Information Processing & Management* **42**, 1379-1391.

[12] D. Lewandowski, Date-restricted queries in web search engines, *Online Information Review* **28** (6), 2004, 420-427.

[13] D. Lewandowski, Problems with the use of web search engines to find results in foreign languages, *Online Information Review* **32**, 2008, 668-672.

[14] P. Ingwersen & K. Järvelin, *The turn: Integration of information seeking and retrieval in context*, (The Kluwer International Series on Information Retrieval,) Springer, Dordrecht, 2005.

[15] N. Höchstötter & D. Lewandowski, What users see – Structures in search engine results pages, *Information Sciences* **179**, 2009, 1796-1812.

[16] L. A. Granka, T. Joachims & G. Gay, Eye-tracking analysis of user behavior in WWW search, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,* July, Citeseer, 2004, 25–29.

[17] G. Buscher, S. T. Dumais & E. Cutrell, The good, the bad, and the random: an eye-tracking study of ad quality in web search, *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010, 42–49.

[18] G. Hotchkiss, Eye Tracking Report: Google, MSN and Yahoo! Compared, http://new.enquiroresearch.com/812-DT/ET2.pdf, 2006.

[19] E. Cutrell & Z. Guan, Eye tracking in MSN Search: Investigating snippet length, target position and task types, http://www.bahlco.com/images/uploads/ms-eye-tracking-01-07.pdf, 2007.

[20] T. Joachims, L. Granka, B. Pan, H. Hembrooke & G. Gay, Accurately interpreting clickthrough data as implicit feedback, *Conference on Research and Development in Information Retrieval*, Salvador, Brazil, ACM, 2005, 154–161.

[21] H. V. Leighton & J. Srivastava, First 20 Precision among World Wide Web Search Services (Search Engines), *Journal of the American Society for Information Science* **50**, 1999, 870-881.

[6] Einen Überblick über die Diskussion zum Search Engine Bias findet sich in dem Kapitel von Karsten Weber in diesem Band. Darüber hinaus ist das Buch von Röhle [71] zu empfehlen.

[22] M. Gordon & P. Pathak, Finding information on the World Wide Web: the retrieval effectiveness of search engines, *Information Processing and Management* **35**, 1999, 141-180.
[23] A. MacFarlane, Evaluation of web search for the information practitioner, *Aslib Proceedings: New Information Perspectives* **59**, 352-366.
[24] J. Griesbaum, Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de, *Information Research* **9**, 2004.
[25] R. G. Demirci, V. Kismir & Y. Bitirim, An evaluation of popular search engines on finding turkish documents, *Second International Conference on Internet and Web Applications and Services, ICIW'07*, Computer Engineering Department, Eastern Mediterranean University, Famagusta, Cyprus: IEEE Computer Society, 2007.
[26] H. Halim & K. Kaur, Malaysian web search engines: a critical analysis, *Malaysian Journal of Library & Information Science* **11**, 2006, 103-122.
[27] S. Tongchim, V. Sornlertlamvanich & H. Isahara, Measuring the effectiveness of public search engines on thai queries, *Communications, Internet, and Information Technology*, M. H. Hamza (Hrsg.), St. Thomas, US Virgin Islands: ACTA Press, 2006.
[28] E. Toth, Exploring the Capabilities of English and Hungarian Search Engines for Various Queries, *Libri* **56**, 2006, 38-47.
[29] F. Lazarinis, J. Vilares, J. Tait & E. Efthimiadis, Current research issues and trends in non-English Web searching, *Information Retrieval* **12**, 2009, 230-250.
[30] D. K. Harman & E. M. Voorhees, TREC: An overview, *Annual review of information science and technology* **40**, 2006, 113-155.
[31] D. Hawking, N. Craswell, P. Bailey & K. Griffiths, Measuring Search Engine Quality, *Information Retrieval* **4**, 2001, 33-59.
[32] N. Höchstötter, Methoden der Erhebung von Nutzerdaten und ihre Anwendung in der Suchmaschinenforschung, *Handbuch Internet-Suchmaschinen*, D. Lewandowski (Hrsg.), Akademische Verlagsgesellschaft Aka GmbH, Heidelberg, 2009, 175-203.
[33] B. J. Jansen & A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, *Information Processing & Management* **42**, 2006, 248-263.
[34] N. Schmidt-Mänz, *Untersuchung des Suchverhaltens im Web: Interaktion von Internetnutzern mit Suchmaschinen*, Kovac, Hamburg, 2007.
[35] A. Spink, *Web search: public searching on the Web*, Kluwer Academic Publishers, Dordrecht, 2004.
[36] N. Höchstötter & M. Koch, Standard parameters for searching behaviour in search engines and their empirical evaluation, *Journal of Information Science* **35**, 2009, 45.
[37] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini & B. Pan, Eye tracking and online search: Lessons learned and challenges ahead, *Journal of the American Society for Information Science and Technology* **59**, 2008, 1041-1052.
[38] A. Gulli & A. Signorini, The indexable Web is more than 11.5 billion pages, *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, Chiba, Japan, 2005, 902-903.
[39] J. Tague-Sutcliffe, The pragmatics of information retrieval experimentation, revisited, *Information Processing & Management* **28**, 1992, 467-490.
[40] B. Clay, Search Engine Relationship Chart, http://www.bruceclay.com/searchenginechart.pdf, 2006.
[41] C. Maaß, A. Skusa, A. Heß & G. Pietsch, Der Markt für Internet-Suchmaschinen, *Handbuch Internet-Suchmaschinen*, D. Lewandowski (Hrsg.), Akademische Verlagsgesellschaft Aka GmbH, Heidelberg, 2009, 3-17.
[42] Suchmaschinen-Marktanteile, www.luna-park.de/home/internet-fakten/suchmaschinen-marktanteile.html, 2008.
[43] P. Schaer, P. Mayr & P. Mutschke, Implications of Inter-Rater Agreement on a Student Information Retrieval Evaluation, *LWA 2010*, Kassel, 2010.
[44] B. J. Jansen, M. Zhang & Y. Zhang, The effect of brand awareness on the evaluation of search engine results, *Conference on Human Factors in Computing Systems – Proceedings*, College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802 Department of Industrial and Mechanical Engineering, College of Engineering, Pennsylvania State University, University Park, PA 16802: 2007, 2471-2476.
[45] P. Bailey & P. Thomas, Does brandname influence perceived search result quality? Yahoo!, Google, and WebKumara, *Proceedings of ADCS*, 2007.
[46] J. Bar-Ilan, K. Keenoy, M. Levene & E. Yaari, Presentation bias is significant in determining user preference for search results – A user study, *Journal of the American Society for Information Science and Technology* **60**, 2009, 135-149.
[47] L. T. Su, Value of Search Results as a Whole as the Best Single Measure of Information Retrieval Performance, *Information Processing & Management* **34**, 1998, 557-579.
[48] A. Broder, A taxonomy of web search, *ACM Sigir forum* **36**, 2002, 3-10.

[49] D. Lewandowski, The retrieval effectiveness of search engines on navigational queries, *ASLIB Proceedings* **61**, 2011.
[50] L. Vaughan, New measurements for search engine evaluation proposed and tested, *Information Processing and Management* **40**, 2004, 677-691.
[51] J. Bar-Ilan, K. Keenoy, E. Yaari & M. Levene, User rankings of search engine results, *Journal of the American Society for Information Science and Technology* **58**, 2007, 1254-1266.
[52] W. Ding & G. Marchionini, A Comparative Study of Web Search Service Performance, *Proceedings of the ASIS Annual Meeting*, Learned Information, 1996, 136-142.
[53] B. J. Jansen, M. Zhang & Y. Zhang, Brand awareness and the evaluation of search results, *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, ACM, 2007, 1139-1140.
[54] D. Lewandowski, Query types and search topics of German Web search engine users, *Information Services & Use* **26**, 2006, 261-269.
[55] D. Lewandowski & N. Höchstötter, Standards der Ergebnispräsentation, *Handbuch Internet-Suchmaschinen*, D. Lewandowski (Hrsg.), Akademische Verlagsgesellschaft AKA GmbH, Heidelberg, 2009, 204-219.
[56] G. Sterling, Yahoo cuts positions, shifting search emphasis, http://searchengineland.com/more-job-cuts-for-yahoo-search-43810, 2010.
[57] S. Quirmbach, Universal Search – Kontextuelle Einbindung von unterschiedlichen Quellen und Auswirkungen auf das User Interface, *Handbuch Internet-Suchmaschinen*, D. Lewandowski (Hrsg.), Akademische Verlagsgesellschaft Aka GmbH, Heidelberg, 2009, 220-248.
[58] M. T. Keane, M. O'Brien & B. Smyth, Are people biased in their use of search engines?, *Communications of the ACM* **51**, 2008, 49–52.
[59] TREC, Web Track Guidelines, http://plg.uwaterloo.ca/~trecweb/, 2009.
[60] R. Agrawal, S. Gollapudui, A. Halverson & S. Ieong, Diversifying Search Results, *Second ACM International Conference on Web Search and Data Mining* **10**, 2009, 5-14.
[61] C. L. A. Clarke, M. Kolla, & G. V. Cormack, Novelty and Diversity in Information Retrieval Evaluation, *SIGIR*, Singapore, 20-24.07.2008, 2008, 8.
[62] G. Marchionini, Exploratory search: from finding to understanding, *Communications of the ACM* **49**, 2006, 41-46.
[63] S. B. Huffman & M. Hochster, How well does result relevance predict session satisfaction?, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, 2007, 567-574.
[64] G. Singer, U. Norbisrath, D. Lewandowski, E. Vainikko & H. Kikkas, Search-Logger – Tool Support for Exploratory Search Task Studies, *SAC2011*, TaiChung, Taiwan, ACM, 2011.
[65] R. Capra, HCI Browser: A Tool for Studying Web Search Behavior, *ASIST 2010*, 2010.
[66] B. J. Jansen, The Wrapper: An Open Source Application for Logging User–System Interactions during Searching Studies, *Logging Traces of Web Activity: The Mechanics of Data Collection Workshop at WWW 2006*, 2006.
[67] D. Lewandowski, The Influence of Commercial Intent of Search Results on Their Perceived Relevance, *iConference 2011*, New York, ACM, 2011, 452-458.
[68] B. Edelman and B. Lockwood, Measuring bias in "organic" web search, http://www.benedelman.org/searchbias/, 2011.
[69] R. Korfhage, *Information storage and retrieval*, Wiley, New York, 1997.
[70] G. Demartini & S. Mizzaro, A Classification of IR Effectiveness Metrics, *European Conference on IR Research*, M. Lalmas (Hrsg.), Springer, London, UK, 2006, 488-491.
[71] T. Röhle, *Der Google-Komplex: Über Macht im Zeitalter des Internets*, Transcript, Bielefeld, 2010.