Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de

Joachim Griesbaum[1] / Marc Rittberger[2] / Bernard Bekavac[1]

[1]Universität Konstanz
Informationswissenschaft
Fach D 87
D-78457 Konstanz
{Joachim.Griesbaum,
Bernard.Bekavac}@uni-konstanz.de

[2]Heinrich-Heine-Universität
Düsseldorf
Institut für Sprache und Information
D-40225 Düsseldorf

Marc@Rittberger.de

Zusammenfassung
Die vier Suchmaschinen AltaVista.de, Fireball.de, Google.de und Lycos.de werden einem Retrievaltest unterzogen, um ihre Eignung für den deutschsprachigen Suchraum zu betrachten. Die Evaluierung erfolgt mit 28 Studierenden und Mitarbeitern der Informationswissenschaft und insgesamt 56 Suchfragen im Januar 2002. Es zeigen sich deutliche Vorteile für Google.de gegenüber den anderen Suchmaschinen. Die sichtbaren Vorteile von Lycos.de können den statistischen Überprüfungen nicht standhalten, so dass bei den anderen drei Suchmaschinen von einer gleich hohen Retrievalleistung ausgegangen werden muss.

1 Einleitung

Transparenz der Ergebnisse und Qualität des Rankings bei Suchmaschinen beschäftigen die Öffentlichkeit und die Wissenschaft immer mehr. Die Aktualität der Suchergebnisse, besser des Index, die möglichst vollständige Erfassung der Inhalte des Internets, aber auch Stichworte wie Spamming[1],

[1] Spamming bezeichnet neben dem Überfluten mit unerwünschten Informationseinheiten durch Werbemails oder Marketingmaßnahmen in News-Groups im Umfeld von Suchmaschinen eher unlautere Verfahren, die zu einem hohen Ranking insbesondere bei nur näherungsweise relevanten Seiten führen.

Cloaking[2] oder Doorway-Pages[3] spielen eine große Rolle in den Diskussionen.[4]

Möglichst hohe Aktualität, große Vollständigkeit und genaue Ergebnisse sind zwar auch im Interesse der Suchmaschinen, um eine hohe Kundenzufriedenheit zu erreichen, aber durch die Größe und die Vielfalt der Daten, die auf dem Internet angeboten werden, nur schwer zu erreichen.

Geht es um die Bewertung von Suchmaschinen, so spielt die Diskussion um die Positionierung von Ergebnissen im Ranking und damit natürlich die Relevanz eine herausragende Rolle. Das hat in den letzten Jahren dazu geführt, dass in fast jeder Internet- oder Computer-Zeitschrift Suchmaschinentests veröffentlicht werden (z.B. [Bager & Schulzki-Haddouti 2001; o.V. 2001]). In solchen Tests werden die jeweilige Testanordnung und Testdurchführung jedoch nur selten umfassend und transparent dargestellt,[5] womit unklar bleibt, ob und inwieweit die erzielten Testresultate neutral und „objektiv" und damit für Suchdienstenutzer tatsächlich hilfreich sind.[6]

Zur Nutzung von Suchmaschinen und den daraus resultierenden Ergebnissen gibt es mit Bezug auf die internationale Ausrichtung von Suchmaschinen zahlreiche Untersuchungen (z.B. [Bar-Ilan 2002; Su & Chen 1999; Wang et al. 1999; Gordon & Pathak 1999; Leighton & Srivastava 1999; Hawking et al. 2001; Ford et al. 2001; Mettrop 2001; Dennis et al. 2002]).

Neben den gerne genutzten internationalen Suchdiensten wird die nationale Ausrichtung von Suchmaschinen allerdings immer wichtiger. Die Zuwachsraten in Sprachräumen oder auf Nationen bezogen, die nicht dem Englischen angehören, sind deutlich höher als die Veränderungen, die sich in den durch das Englische dominierten Informationsmärkten ergeben. Die

---

[2] Cloaking ist das IP-abhängige Zurverfügungstellen unterschiedlicher Versionen ein und derselben Webseite.

[3] Doorway-Pages sind speziell für Suchroboter optimierte Seiten, die insbesondere relevante Texte, Indexierung etc. enthalten.

[4] Mit Spamming, Cloaking oder Doorway-Pages werden Verfahren beschrieben, mit denen Anbieter von Inhalten auf dem World Wide Web versuchen, ihre Seiten bei den Suchmaschinen möglichst gut zu platzieren, ohne dass die Inhalte, die über die genannten Verfahren suggeriert werden, wirklich vorhanden sind.

[5] Siehe z.B. Newsmeldung vom 13.05.02 bei Suchfibel.de „Seltsamer PC Pro Suchmaschinentest". URL http://www.suchfibel.de/news/SeltsamerPCProSuchmaschi.htm 17.05.02

[6] Eine vergleichende Gegenüberstellung solcher Tests findet sich auf der Webseite von Klaus Patzwald. URL http://www.at-web.de/Informationen/suchmaschinen-test.htm 24.04.02

Nationalisierungsstrategien der Anbieter [Dresel et al. 2001], die damit den Anforderungen der Informationsmärkte folgen, bestätigen diesen Eindruck. Bei einer Befragung von 43 Studierenden der Informationswissenschaft in Düsseldorf, Konstanz und Chur wurde die Wichtigkeit dieser Einschätzung bestätigt, da die Mehrheit der Studierenden der Einschränkung der Suchergebnisse auf nationale, mehr noch auf Sprachraumgrenzen eine hohe Bedeutung gab.[7]

Um so verwunderlicher ist es daher, dass zumindest für den deutschsprachigen Informationsmarkt kaum wissenschaftliche Untersuchungen und Ergebnisse über die Leistungsfähigkeit der auf diesen Sprachraum fokussierten Suchmaschinen zur Verfügung stehen.[8] Der Frage nachzugehen, wie gut die deutschsprachigen Suchmaschinen sind, scheint daher durchaus eine Lücke zu schließen.

Dieser allgemeine Mangel an aktuellen Studien zum Vergleich von deutschen Suchmaschinen wird im Januar 2002 in Zusammenarbeit der informationswissenschaftlichen Forschungsgruppen in Konstanz und Düsseldorf zum Anlass genommen, einen Retrievaltest der Suchmaschinen Altavista.de, Fireball.de, Google.de und Lycos.de durchzuführen. Der Retrievaltest verfolgt zwei Ziele, erstens die Retrievaleffektivität der untersuchten Suchmaschinen zu ermitteln und zweitens das in [Griesbaum 2000] konzipierte Evaluationsverfahren, auf dem dieser Retrievaltest aufbaut, weiter zu entwickeln und damit für die systematische Durchführung von Retrievaltests mit Suchmaschinen weitere Erfahrungen zu gewinnen.

2 Vorgehensweise und Zielsetzung

Die vorliegende Evaluation beruht methodisch primär auf der von [Tague-Sutcliffe 1992], S.467 vorgeschlagenen grundlegenden Vorgehensweise zur Entwicklung eines Retrievaltests.[9] Die Testart folgt dabei dem Cranfield Paradigma [Ellis 1992] und ist grundsätzlich bemüht, z.B. in Anlehnung an TREC, etablierte Evaluationsstandards einzuhalten.[10]

---

[7] To be published

[8] Ausnahmen sind [Dresel et al. 2001; Stock & Stock 2000; Wolff 2000]

[9] Dieser „Leitfaden“ löst zwar nicht die elementare Problematik jedes Retrievaltests an sich, die der adäquaten Ausgestaltung der quantitativen und qualitativen Ausgestaltung der Testparameter – kann sie nicht lösen – bietet aber in den vorgeschlagenen zehn sequentiell zu durchlaufenden Schritten grundlegende Hinweise und Entscheidungshilfen, welche die Validität, Reliabilität und Effizienz des Testsettings absichern sollen.

[10] Beispielsweise bei der Anzahl der Suchanfragen.

Die Vorgehensweise nach [Tague-Sutcliffe 1992] unterteilt einen Retrievaltest in zehn Teilbereiche, welche sequentiell abzuarbeiten sind:[11]

1. Testen oder nicht testen (Need for testing) – Motivation des Retrievaltests
2. Testart (Type of test) – Bestimmung des grundsätzlichen Testverfahrens
3. Variablendefinition und -zuordnung (Definition of variables)
4. Verwendetes Informationssystem (Database development) – Ausgewählte Suchmaschinen
5. Erschließung der Informationsbedürfnisse und Suchanfragen (Finding queries)
6. Durchführung der Suchanfragen (Retrieval software)
7. Testanordnung (Experimental design)
8. Datenerfassung (Data collection)
9. Datenauswertung (Data analysis)
10. Ergebnispräsentation (Presenting Results)

Steht das Testsetting, wird durch einen Pretest überprüft, ob das Evaluationssetting tatsächlich adäquat ausgestaltet ist. In Abhängigkeit der Resultate des Pretests wird die Testanordnung gegebenenfalls angepasst bis dieses Ziel sichergestellt scheint. Sind diese Voraussetzungen erfüllt, werden die Tests durchgeführt, die Daten ausgewertet und analysiert.

Um einerseits die pragmatische Handlungsrelevanz der Ergebnisse im realen Nutzungskontext der untersuchten Suchdienste einschätzen zu können und andererseits Probleme bei der Durchführung der Untersuchung aufzuzeigen und – idealerweise – Optimierungspotenziale für künftige Retrievaltests zu erschließen, werden abschließend sowohl die Ergebnisse als auch die Untersuchung selbst kritisch hinterfragt.

Innerhalb dieser „theoretisch-normativen“ Rahmenstruktur wird bei der konkreten Ausgestaltung der einzelnen Testparameter versucht, webspezifische Eigenheiten des Information Retrieval hinsichtlich Datenbestand, Hypertextstrukturen und Nutzerverhalten möglichst realitätsnah nachzubilden. [Gordon & Pathak 1999] nennen sieben Kriterien zur Evaluierung von Suchmaschinen im Web als essentiell, die von [Hawking et al. 2001] kritisch hinterfragt wurden und zu Recht auf fünf Kriterien reduziert werden:

1. Reale Informationsbedürfnisse von Nutzern sollen abgebildet werden.
2. Bei der Einbindung von Informationsvermittlern soll das originäre Informationsbedürfnis sorgfältig mitgeteilt werden.
3. Es soll eine große Anzahl von Suchfragen genutzt werden.

[11] In Klammern stehen die englischen Originalbezeichnungen für die einzelnen Stufen eines Retrievaltests.

4. Die wichtigsten Suchmaschinen sollen involviert sein.
5. Die Untersuchung soll gut und sorgfältig aufgebaut und durchgeführt werden.

Diese fünf Kriterien, die bei einer Suchmaschinenevaluierung erfüllt sein sollen, werden von [Hawking et al. 2001] um das Kriterium, dass die Frageformulierung das Informationsbedürfnis möglichst gut treffen soll, ergänzt.

Wir werden im Rahmen der Systematik von [Tague-Sutcliffe 1992] die einzelnen Schritte unseres Retrievaltests im nächsten Abschnitt vorstellen. Ausreichend diskutiert erscheint uns dabei schon der erste Punkt der Motivation des Retrievaltests. Auch werden wir die Ergebnispräsentation ausführlicher in einem eigenen Abschnitt darstellen.

3 Durchführung des Tests

3.1 Testart – Bestimmung des grundsätzlichen Testverfahrens

Das gewählte Testverfahren ist ein Test anhand einer Testkollektion. Die Informationsbedürfnisse, Suchanfragen, Bewertungsmaße und Bewertungskriterien werden den Testpersonen (Juroren) somit extern von den Untersuchenden vorgegeben. Dieses dem Cranfield-Paradigma folgende, an die TREC ad hoc und Web trac Tasks [Voorhees & Harman 2001] angelehnte, insgesamt eher laborhaft ausgestaltete Testverfahren bildet die reale Nutzungssituation bei Suchdiensten zwar nur unvollkommen nach, ermöglicht aber gerade durch die „standardisierte“ Testanordnung eine Abstraktion von individuell wirkenden, schlecht zu kontrollierenden Einflussfaktoren und sichert damit die Vergleichbarkeit der einzelnen Testergebnisse [Griesbaum 2000], S.26, 58f.

3.2 Variablendefinition und –zuordnung

Die Variablen Relevanzeinstufung, Dokumentdarstellung, Bewertungsmaße, Suchanfragen und Informationsbedürfnisse, Testpersonen, Dokumentraum, und Relevanzeinstufung des Retrievaltests sowie die Handhabung der Variablen werden festgelegt.

Relevanzeinstufung
Als zentrales Konzept zur Beurteilung von Retrievalsystemen gilt die Relevanz der gelieferten Treffer [Robertson 1981], S.14. Dieses trefferbasierte Bewertungsmaß bildet trotz der Problematik der personengebundenen Subjektivität jedes Relevanzurteils, mangels Alternativen, auch in diesem Test die Grundlage der Bewertungsmaße [Warner 2000], S.77.

Die Relevanzeinstufung der Treffer wird von Testpersonen, die als Juroren fungieren, vorgenommen. Dies gewährleistet zumindest, dass die Bewertungen von Vorlieben bzw. Abneigungen der Untersuchenden unbeeinflusst bleiben [Robertson 1981], S.17. Um dies auch bei den Juroren selbst sicherzustellen, wird die Herkunft der Treffer unkenntlich gemacht. Um die Eindeutigkeit der Relevanzurteile abzusichern, wird jedes Dokument nur von einem Juror bewertet.

Durch die Hypertextstruktur des Web ist es denkbar, dass ein an sich irrelevantes Dokument über Verknüpfungen den direkten Zugriff auf relevante Seiten ermöglicht. Aus diesem Grund wird von einer binären Relevanzeinstufung in „relevant“ und „nicht relevant“ zunächst abgesehen und als dritte Bewertungsmöglichkeit „verweist auf relevante Seite(n)“ hinzugefügt.[12] Weil solche als „verweist auf relevante Seite(n)“ bewerteten Treffer letztendlich hilfreich sind, um Informationsbedürfnisse zu befriedigen, werden sie bei der Auswertung zu den relevanten Treffern gezählt. Dies reflektiert auch das tatsächliche Verhalten der Suchdienstenutzer, bei denen während der Nutzung der Suchdienste die Browsing-Perioden überwiegen [Körber 2000], S.41.

Problematisch ist, dass bei der Relevanzbewertung von einer Einzelbetrachtung der Dokumente ausgegangen wird. Damit wird beispielsweise ein gleichbleibender Wissensstand des Jurors impliziert. Die sogenannten „Grenzfälle der Relevanz“, vgl. [Griesbaum 2000], S.64, werden damit bis auf die Berücksichtigung von Dubletten, die nur beim ersten Auftreten als relevant gewertet werden können, nicht berücksichtigt.

Dokumentdarstellung
Die Relevanzeinstufung wird anhand der Ergebnisseiten selbst vorgenommen. Die Juroren haben eine Liste von Links im Browser vor sich, die auf die relevanten Seiten der von den Suchmaschinen gefundenen Treffer verweisen.

[12] Dies reflektiert auch das tatsächliche Verhalten der Suchdienstenutzer, bei denen während der Nutzung der Suchdienste die Browsing-Perioden überwiegen [Körber 2000], S.41.

Bewertungsmaße

Recall und Precision sind die Standardwerte zur Effektivitätsmessung von Retrievalsystemen [Lesk 1995]. Auf Recall wird in dieser Untersuchung verzichtet, da er zum einen im Web nicht oder nur unzureichend bestimmt werden kann [Oppenheim et al. 2000], S.190, und zum anderen der vollständige Nachweis aller relevanten Dokumente, gerade für Nutzer im Web häufig nur von geringem Interesse ist, da sie in der überwiegenden Zahl der Fälle nur die ersten zwei, selten die ersten drei Ergebnisseiten[13] der Suchmaschinen sichten.[14] Dieser Retrievaltest beschränkt sich deshalb auf die Effektivitätsbeurteilung anhand der Relevanzbeurteilung der ersten 20 Treffer, der sogenannten Top20 Precision.

Bei der Auswertung werden sowohl die Mikro- als auch die Makromethode verwendet. Bei der Makromethode werden zunächst die einzelnen Suchanfragen als Grundeinheit betrachtet, d.h. zuerst werden die Precisionwerte pro Suchanfrage berechnet und dann die Werte der Suchanfragen gemittelt, damit fließt jede Suchanfrage gleichgewichtig in die Bewertung ein. Bei der Mikromethode hingegen werden die Dokumente als Grundeinheit genommen, sie berechnet, unabhängig von der Trefferanzahl der einzelnen Suchanfragen, das Verhältnis der relevanten Treffer zu allen Treffern, dadurch fließt jedes Dokument gleichgewichtig in die Bewertung mit ein [Womser-Hacker 1989], S.66f .

Suchanfragen und Informationsbedürfnisse

Die Informationsbedürfnisse und Suchanfragen sind das zentrale Element jedes Retrievaltests. Die Thematik, Komplexität, Spezifität der Fragestellungen und die Art der Suchanfragenformulierung spezifizieren die inhaltliche Ausprägung der Untersuchung und determinieren so als wichtigste Inputfaktoren bei der Testdurchführung unmittelbar die Quantität und Qualität der Treffer der Suchmaschinen.[15]

---

[13] Gemeint sind hier die Trefferlisten.

[14] Laut AltaVista.com benutzen sogar weniger als 10% die zweite Ergebnisseite, vgl. [Körber 2000], S.33 und [Jansen et al. 2000].

[15] Zur Veranschaulichung folgendes Beispiel:
Zur Suchanfrage „hypothesengenerierende Untersuchungsverfahren“ in Phrasenform liefern Google.de, Lycos.de, Fireball.de und bei Altavista.de keinen Treffer. Bei der Suchanfrage „Star Wars Klonkriege Premiere Termin“ ohne Phrasenform referenziert Google.de 4 Treffer, Lycos.de 3 Treffer, Fireball.de 2 Treffer und Altavista.de 1 Treffer, die Suchanfrage „Autokauf“ liefert bei Google.de 93.800 Treffer, bei Lycos.de 23888 Treffer, bei Fireball.de 14996 Treffer und bei Altavista.de 17765 Treffer.
Anfragen durchgeführt am 30.04.02

Die Fragestellungen werden in dieser Untersuchung nicht den „typischen“ Fragestellungen im Web nachgebildet, sondern als normative Ausprägung des Retrievaltests festgelegt.

Der Themenbereich wird zunächst durch Abgrenzung unerwünschter Bereiche negativ spezifiziert. Explizit ausgeschlossen werden Fragestellungen, die auf materiellen Bedürfnissen beruhen und auf Produkt- oder Dienstleistungsangebote kommerzieller Anbieter zielen. Ebenso ausgegrenzt werden Fragestellungen, die direkt auf Freizeitaktivitäten oder Hobbies schließen lassen. Fragestellungen aus dem Porno/Erotik-Bereich werden ebenfalls ausgeklammert.

Die Anzahl der Suchanfragen wird mit 50, entsprechend der Standardvorgabe von TREC, hinreichend hoch gewählt, um verallgemeinerungsfähige Testergebnisse zu erhalten. [Buckley & Voorhees 2000], S.33. Damit ein Puffer zur Verfügung steht, werden 60 Suchanfragen vorbereitet.
Die Formulierung der Suchanfragen wird den typischen Nutzergewohnheiten angelehnt, die überwiegend Suchwörter ohne Operatoren oder Klammerung eingeben [Jansen et al. 2000].

Testpersonen
Zur Verfügung stehen die Teilnehmer der Information Retrieval Vorlesungen der Informationswissenschaft an den Universitäten in Düsseldorf und Konstanz. Alle Teilnehmer sind fortgeschrittene Studierende, die sowohl im Umgang mit dem Web als auch mit Suchmaschinen über viel Erfahrung verfügen.

Dokumentraum Internet
Das Internet ist im Rahmen dieses Retrievaltests als nicht zu beeinflussender Dokumentraum einzuordnen. Die Datenbestände des Internet sind sehr heterogen und dynamisch, bieten aber den Suchdiensten und ihren Indexierungskomponenten dieselben Bedingungen.

Relevanzeinstufung der Treffer
Suchdienste liefern in der Regel eine nach maschinell berechneter Relevanz sortierte Liste von Links (Treffern), die den Suchdienstenutzer direkt zu den Ergebnisseiten führen. Die Listen werden im Testumfeld so maskiert, dass die Juroren nicht erkennen können, von welcher Suchmaschine welche Treffer stammen.

3.3 Ausgewählte Suchmaschinen

Es werden vier Suchmaschinen miteinander verglichen. Die Auswahlkriterien sind Reichweite und Nutzungsgrad. Die Analyse einschlägiger statistischer Quellen[16] ergibt folgende Auswahl: Altavista.de, Fireball.de, Google.de und Lycos.de. Mitausschlaggebend für die Auswahl war die Untersuchung der Stiftung Warentest [o.V. 2001], die die genannten vier Suchmaschinen als einzige mit einem guten (Google.de) oder zumindest befriedigendem Abschneiden bei der Bewertung der Suchergebnisse versehen hat.

3.4 Erschließung der Informationsbedürfnisse und Suchanfragen

Das grundlegende Problem der „Repräsentativität“ von Suchanfragen und Informationsbedürfnissen kann im Rahmen dieses Tests nicht gelöst werden. Um aber bei der Auswahl eine thematische oder sonstige wie auch immer geartete Verengung zu vermeiden, werden bei der Erschließung der Suchanfragen verschiedene Quellen berücksichtigt.

Quelle	Anzahl Suchanfragen
Cross Language Track (Trec 7)	7
Cross Language Track (Trec 8)	5
GIRT (Trec 8)	17
Web Track (Trec 8)	4
Ask Jeeves	27

Tabelle 1: Die Quellen der Suchfragen

Anzahl offene Fragestellungen	Anzahl geschlossene Fragestellungen
11	49

Tabelle 2: Anzahl offener und geschlossener Suchfragen

Anzahl Suchbegriffe	Anzahl Suchfragen
1	1
2	30
3	18
4	8
5	1
6	1

Tabelle 3: Anzahl der Suchbegriffe

Deshalb werden zum einen Informationsbedürfnisse und Fragestellungen aus verschiedenen Teilkollektionen von Trec verwendet (Cross Language Topics, GIRT, Web Track), zum anderen aber auch neue „real existierende“ Fragestellungen durch Auswertung von Anfragen im Web erschlossen. Dabei

[16] Vgl. Jupiter MMXI, European Search Engine Ratings.
URL http://www.searchenginewatch.com/reports/mmxi-europe.html 05.04.02,
URL http://www.webhits.de/deutsch/index.shtml?/deutsch/webstats.html 03.04.02

wird, in Anlehnung an die Fragestellungen von Web Track, der Fragenprotokolldienst von Ask Jeeves ausgewählt.[17]

Tabelle 1 zeigt die Herkunft der Suchfragen an, Tabelle 2 die Anzahl der offenen und geschlossenen Suchfragen und Tabelle 3 die Anzahl der Suchbegriffe in den einzelnen Fragen. Demnach gibt es bspw. acht Suchfragen, die mit vier Suchbegriffen operieren.

3.5 Durchführung der Suchanfragen

Das Ziel, die Validität, Reliabilität und Effizienz des Retrievaltests im Vergleich zur vorangegangenen Untersuchung [Griesbaum 2000] zu erhöhen, soll durch eine Automatisierung bei der Durchführung der Suchanfragen erreicht werden.

In [Griesbaum 2000] wurden die Suchanfragen händisch durchgeführt. Dieser Prozess, insbesondere die zur Verfügungstellung der Suchergebnisse, war sehr zeitaufwändig – Dauer rund zwei Wochen – und die originalgetreue Nachbildung einiger, insbesondere dynamischer Ergebnisseiten gelang nur näherungsweise. Diese Probleme werden in der vorliegenden Evaluierung konzeptionell durch ein Abfrageskript gelöst, welches es gestattet, die Suchanfragen erst unmittelbar vor der Relevanzbeurteilung der Dokumente durchzuführen. Damit sollen die zeitlichen Verzerrungen zwischen Anfrage und Bewertungszeitpunkt minimiert werden, wodurch die Validität der Untersuchung erhöht wird. Da damit zugleich die fehlerbehaftete lokale Spiegelung der Ergebnisseiten überflüssig wird, werden weniger Ressourcen benötigt, was wiederum zu einer Effizienzsteigerung führt.

3.6 Testanordnung

Aufgrund der Menge der Suchanfragen und der Anzahl und Verteilung der Juroren in Konstanz und Düsseldorf wird entschieden, den Test innerhalb von zwei Tagen durchzuführen. Die Auswirkungen der zeitlichen Unterschiede werden als hinnehmbar betrachtet, solange keine der Suchmaschinen innerhalb des Testzeitraums größere Modifikationen vornimmt.[18] Die Juroren haben jeweils insgesamt zwei Suchanfragen zu beurteilen. Der zeitliche Aufwand wird pro Termin auf rund eine Stunde geschätzt, wobei eine breite

[17] http://www.askjeeves.com/docs/peek/ 05.05.02

[18] Etwa vergleichbar mit dem Relaunch von Fireball.de am 04.04.2002 http://www.tecchannel.de/news/20020404/thema20020404-7194.html 05.05.02

Streuung der benötigten Zeit von rund einer halben bis zu zwei Stunden erwartet wird.

Im Testablauf ist es die Aufgabe der Juroren, nacheinander die Treffer 1-20 der vier Suchdienste zu beurteilen. Diese Anordnung entspricht am ehesten einem „Repeated Measures Design“ [Tague-Sutcliffe 1992]. Die dabei auftretenden Lern- und Ermüdungseffekte sollen durch die Variation der Reihenfolge der Trefferlisten der Suchmaschinen kompensiert werden.

3.7 Datenerfassung

Es werden die Bewertungen und Anmerkungen zu den Ergebnisseiten aufgenommen und die persönlichen Angaben über die Juroren hinsichtlich Alter, Geschlecht, Qualifikation und Gemütsbefinden erfasst.

3.8 Datenauswertung

Zunächst werden die Relevanzeinstufungen „Relevant“ und „Verweist auf relevante Seite(n)“ zusammengeführt und als relevant bewertet, um zu einem binären Werturteil zu kommen, welches für die Berechnung der Retrievalmaße notwendig ist.

Die Retrievaleffektivität der Suchmaschinen wird konkret anhand der Maßzahl der Top20 Precision überprüft. Dabei interessieren primär zwei Ergebnissichten. Zum einen die Makroprecision, die zeigt, wie effektiv die Suchmaschinen die einzelnen Suchanfragen beantworten, und zum anderen die Mikroprecision, die aufzeigt, welche Effektivität die Suchdienste über alle Suchanfragen hinweg bei den Cut-off Werten 1-20 erreichen

Zur Ermittlung und Absicherung der statistischen Validität der Ergebnisse werden diese auf Signifikanz überprüft. Erst dadurch ist es möglich zu entscheiden, ob die ermittelten Unterschiede hinreichend sind, um verallgemeinernde Schlüsse zu ziehen.

4 Pretest

Der Pretest wird am 04.01.02 durchgeführt. Bei der Durchführung der Suchanfragen gibt es Schwierigkeiten. Einige Suchmaschinen, vor allem Google, liefern aus nicht nachvollziehbaren Gründen bei einigen Anfragen zunächst keine Treffer bzw. bei wiederholter Anfrage unterschiedliche Trefferzahlen.

Deshalb werden beim realen Test die Suchanfragen bereits einen Tag vor dem ersten Termin, an dem die Relevanzbeurteilung durch die Juroren vorgenommen wird, durchgeführt und nicht jeweils unmittelbar vor den Terminen.
Diese Entscheidung wirft Probleme auf. Denn die Trefferlisten werden zwar zum gleichen Zeitpunkt generiert, die Ergebnisseiten sind allerdings zum Zeitpunkt der Relevanzbeurteilung schon mindestens einen Tag „veraltet". Damit besteht die Gefahr bzw. die hohe Wahrscheinlichkeit, dass sich zumindest bei einigen Ergebnisseiten innerhalb dieses Zeitraums die Inhalte verändern oder gar nicht mehr auf die Seiten zugegriffen werden kann.[19]

Im Gegensatz zur Durchführung der Suchanfragen verläuft die Relevanzbeurteilung unproblematisch.[20] Der Test dauert eine Stunde und der Testjuror selbst bewertet die benötigte Zeitdauer und die persönliche Belastung als unproblematisch, die Aufgabenstellung als intuitiv verständlich und den Ablauf als klar gekennzeichnet.

Um zu überprüfen, ob das Verschleiern der Herkunft der Treffer trotz der Tatsache, dass ausschließlich Google PDF-Files als Treffer liefert und Katalogtreffer mit dem String „directory.google.com" kennzeichnet, auch bei Experten funktioniert, wird die Pretestsuchanfrage besonders ausgestaltet. Und zwar derart, dass die Trefferliste von Google bei 8 Treffern PDF-Dokumente als Ergebnisseiten liefert. Der Juror antwortet nach dem Pretest auf die Frage, ob er die Trefferlisten einzelnen Suchmaschinen zuordnen konnte, mit nein. Dies ist ein starker Hinweis darauf, dass die Präsenz von PDF-Dokumenten in den Trefferlisten nicht dazu führt, dass die Juroren Google als Quelle dieser Treffer assoziieren.

5 Testdurchführung

Die Suchanfragen werden am 14.01. in Konstanz durchgeführt. Die Probleme unterschiedlicher Treffermengen treten bei einigen Suchanfragen wieder auf und werden durch wiederholte Durchführung der entsprechenden Anfragen kompensiert. Die Relevanzbeurteilungen werden am 15.01. und 16.01. durchgeführt. Die Testpersonen sind die Teilnehmer der Retrievalkurse in

[19] Beispielsweise bei URL not found (404) Seiten, die gelöscht oder verschoben wurden.
[20] Die speziell für den Pretest vorbereitete Suchanfrage findet sich unter http://www.inf-wiss.uni-konstanz.de/CURR/winter0102/IR/uebungen_ir_ws0102/suchanfragen/63_ghmb_o.html 07.05.02

Konstanz und Düsseldorf. In Konstanz werden zusätzlich noch einige Mitarbeiter im Fachbereich rekrutiert. Insgesamt können die Suchtreffer von 56 der 60 vorbereiteten Suchanfragen auf Relevanz bewertet werden.

Die Relevanzbeurteilungen verlaufen unproblematisch. Hinsichtlich der getesteten Suchmaschinen gibt kein Juror einen Hinweis dahingehend, dass er die Trefferlisten einzelnen Suchdiensten zuordnen kann. Vielmehr werden die Untersuchenden von den Juroren nach dem Test oft explizit gebeten, doch mitzuteilen, welche Trefferliste welcher Suchmaschine zuzuordnen ist.

6 Ergebnisanalyse

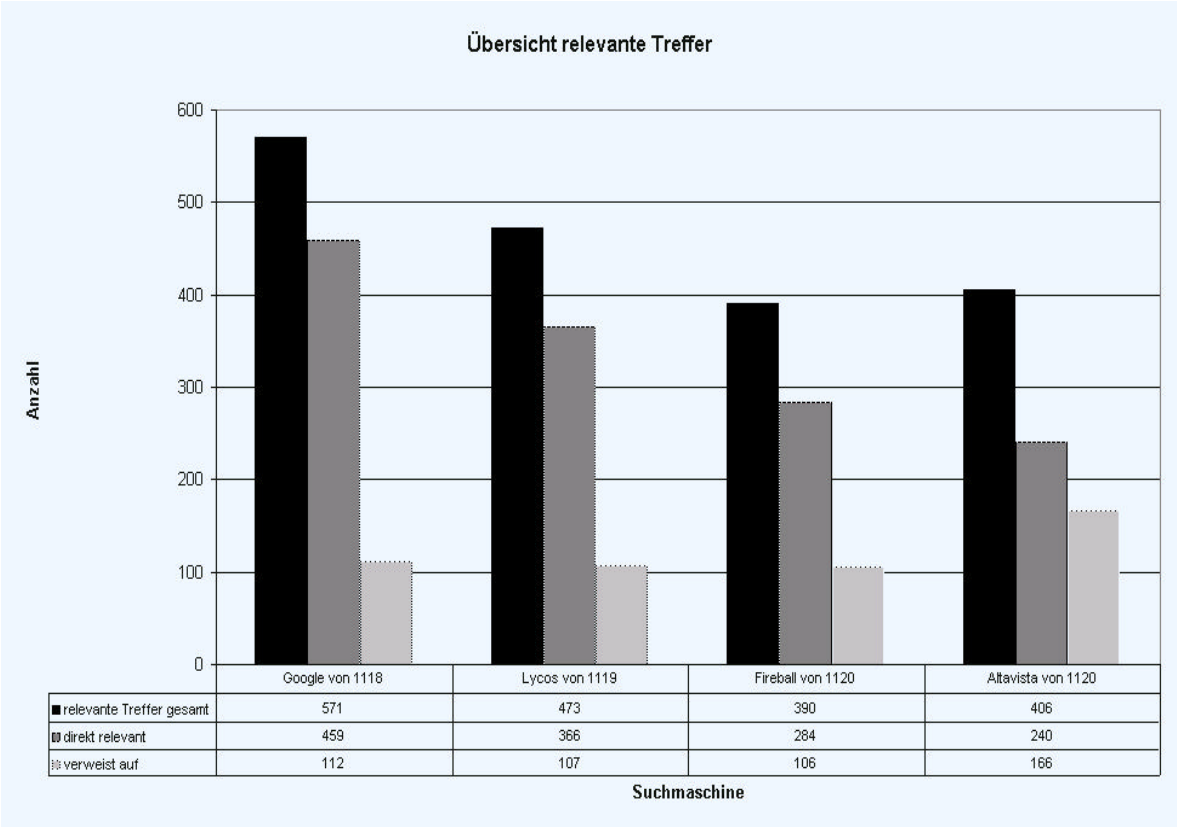
6.1 Übersicht Anzahl relevanter Treffer

Die Übersicht über die relevanten Treffer in Abbildung 1 zeigt an, wie viele Treffer von der pro Suchmaschine maximal erreichbaren Trefferzahl von 1120 Treffern als relevant beurteilt wurden.[21] Während Google über 50% relevante Treffer liefert, erreichen Altavista und Fireball "nur" gut ein Drittel, Lycos liegt dazwischen. Im Vergleich der Anzahl relevanter Dokumente schneidet Google.de somit am besten ab.

6.2 Top20 MeanAverage Precision

Der Vergleich Top20 (Mikro) Precision mittels Recall-Precision Graph in Abbildung 2 sagt aus, welche Suchmaschine in der Lage ist, die größte Anzahl relevanter Treffer auf den kumulierten Rangplätzen (Cut-off Werte 1-20) über alle Suchanfragen zurückzugegeben.

[21] Das bedeutet nicht, dass jede Suchmaschine so viele Treffer zurücklieferte, tatsächlich ist dies nur bei Altavista (wahrscheinlich aufgrund der disjunktiven Verknüpfung der Suchbegriffe) der Fall. Google referenziert zu allen Suchanfragen auf den ersten 20 Rangplätzen 1084 Treffer, Lycos 1082 und Fireball 1024 Treffer. Würde man also die Precision auf die tatsächliche Trefferanzahl innerhalb der ersten zwanzig Rangplätze beziehen und nicht auf die ersten 20 Rangplätze, so würde diese bei den letztgenannten Maschinen leicht höher ausfallen.

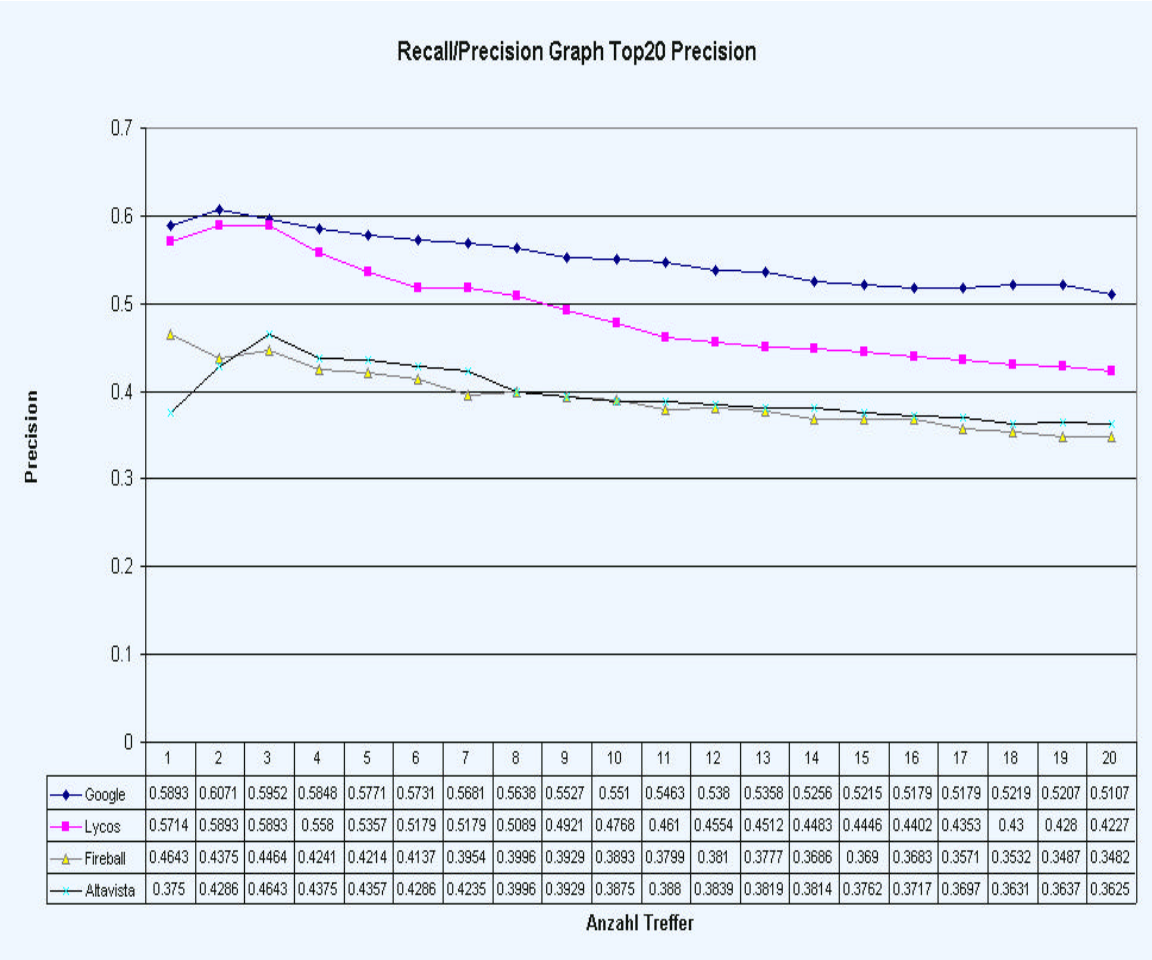


	Google von 1118	Lycos von 1119	Fireball von 1120	Altavista von 1120
relevante Treffer gesamt	571	473	390	406
direkt relevant	459	366	284	240
verweist auf	112	107	106	166

Abbildung 1: Anzahl der relevanten Treffer im Verhältnis zu allen Treffern für die Suchmaschinen google.de, lycos.de, fireball.de, altavista.de

Abbildung 2 ist wie folgt zu deuten. Die Werte auf Rangplatz eins sagen z.B. aus, dass Google über alle Suchanfragen hinweg auf der ersten Position in 58,93% der Fälle relevante Treffer liefert, Lycos 57,14% usw. Die Werte auf Rangplatz zwei sind der Anteil der relevanten Dokumente bis einschließlich Rangplatz zwei. Dies sind z.B. bei Altavista auf Rangplatz zwei 42,86%. Der Graph kumuliert also immer die Anzahl der relevanten Dokumente bis einschließlich der jeweiligen Position.

Google erreicht die höchste Effektivität, die Top1 Precision beträgt 58, 93% und fällt bis auf Platz 20 auf die oben schon genannte Gesamt-, also die Top20 Precision von 51%. Es lässt sich die Aussage treffen, dass Google über alle Suchanfragen bei jedem Cut-off Wert mehr relevante Treffer liefert, als jede andere Maschine. An zweiter Stelle steht Lycos mit einer Top1 Precision von 57% die bis zur Top20 Precision auf 42% fällt. Lycos liefert wiederum zu jedem Cut-off Wert mehr relevante Treffer als Altavista und Fireball. Vergleicht man Altavista und Fireball, so ist nicht unmittelbar klar, welche der beiden Maschinen einen höheren Recall – Precision Graph aufweist. Bei den Rangplätzen 1 und 2 ist Fireball besser, aber ab dem Cut-off Wert von 3 erreicht Altavista mit Ausnahme der Positionen 9 und 10 leicht bessere Werte.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Google	0.5893	0.6071	0.5952	0.5848	0.5771	0.5731	0.5681	0.5638	0.5527	0.551	0.5463	0.538	0.5358	0.5256	0.5215	0.5179	0.5179	0.5219	0.5207	0.5107
Lycos	0.5714	0.5893	0.5893	0.558	0.5357	0.5179	0.5179	0.5089	0.4921	0.4768	0.461	0.4554	0.4512	0.4483	0.4446	0.4402	0.4353	0.43	0.428	0.4227
Fireball	0.4643	0.4375	0.4464	0.4241	0.4214	0.4137	0.3954	0.3996	0.3929	0.3893	0.3799	0.381	0.3777	0.3686	0.369	0.3683	0.3571	0.3532	0.3487	0.3482
Altavista	0.375	0.4286	0.4643	0.4375	0.4357	0.4286	0.4235	0.3996	0.3929	0.3875	0.388	0.3839	0.3819	0.3814	0.3762	0.3717	0.3697	0.3631	0.3637	0.3625

Abbildung 2: Top20 (Mikro-) Precision für google.de, lycos.de, fireball.de und altavista.de

Nimmt man als Mean Average Precision[22] den Durchschnitt der pro Rangplatz erreichten Precisionwerte, um die Resultate der Top1 bis Top20 Precision der verschiedenen Maschinen in einem aussagekräftigen Datum miteinander zu vergleichen, so ergibt sich das Bild in Tabelle 4.

Suchmaschine	Mean Average Precision
Google	0.551
Lycos	0.488
Fireball	0.391
Altavista	0.396

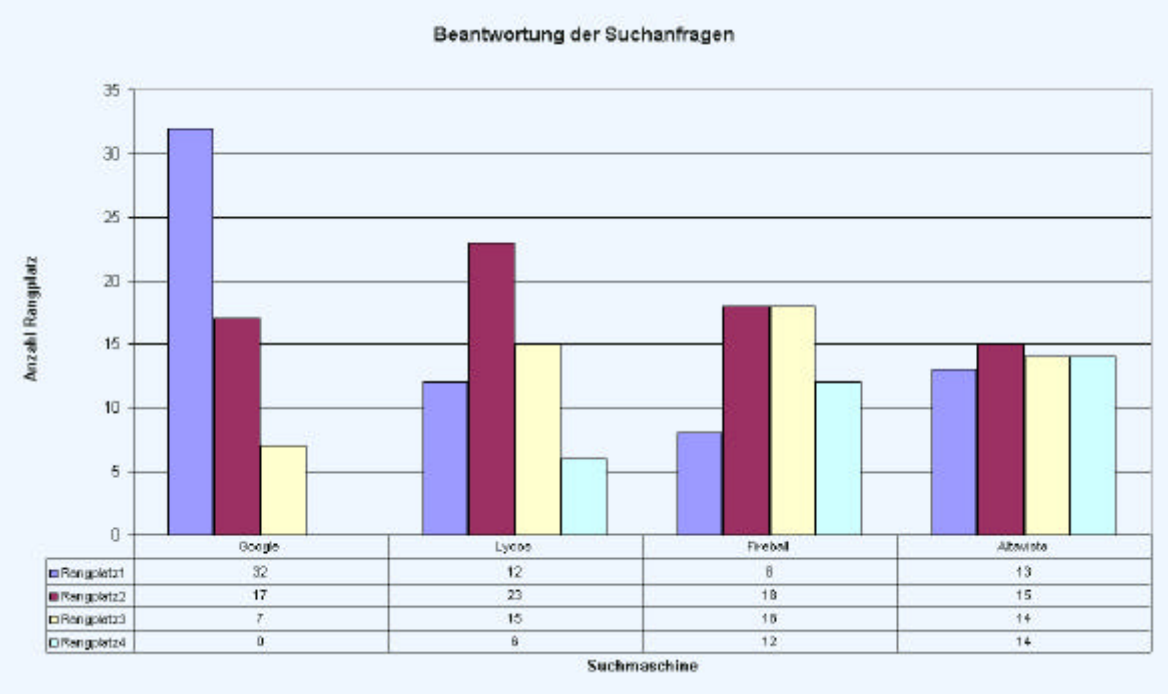
Tabelle 4: Mean Average Precision der Suchmaschinen google.de, lycos.de, fireball.de und altavista.de

[22] http://www-nlpir.nist.gov/works/presentations/spie99/tsld016.htm 12.04.02

Die Überprüfung ergibt, dass Google signifikant besser ist als die anderen Maschinen und Lycos signifikant besser als Altavista und Fireball. Im Vergleich Altavista Fireball lässt sich aber keine klare Aussage treffen.[23]

6.3 Beantwortung der Suchanfragen

Die Retrievaleffektivität der Suchmaschinen bei den einzelnen Suchanfragen (Makroprecison) beschreibt, wie effektiv die Suchmaschinen Informationsbedürfnisse befriedigen.



	Google	Lycos	Fireball	Altavista
Rangplatz1	32	12	8	13
Rangplatz2	17	23	18	15
Rangplatz3	7	15	18	14
Rangplatz4	0	6	12	14

Abbildung 3: Rangplätze in Bezug auf die 56 Suchanfragen der Suchmaschinen google.de, lycos.de, fireball.de und altavista.de

Abbildung 3 zeigt, wie häufig welche Maschine im Vergleich welchen Rangplatz, bezogen auf die Precision bei der jeweiligen Suchanfrage, erreicht. Rangplatz 1 bedeutet, die Maschine erreicht bei der jeweiligen Suchanfrage

[23] Der Vergleich der kumulierten Precision mittels Vorzeichentest zur Überprüfung der Signifikanz ist in diesem Fall methodisch problematisch, da die Werte der paarweisen Einzelvergleiche nicht unabhängig voneinander sind (z.B. ist der Wert von Top2 zu 50% durch den Wert von Top1 bestimmt). Eine simple alternative Lösung ist aber nicht bekannt, da die Position der Treffer primär durch die Suchmaschinen determiniert wird (Ranking - Treffer auf Platz zwei ist deshalb auf Platz 2, weil er laut Maschine schlechter als der auf Platz 1 und besser als der auf Platz 3 ist). Ein denkbarer anderer Vergleich, etwa nur der Treffer auf den jeweiligen Positionen (z.B Vergleich der Treffer auf Platz 20 von AV und Google) wäre deshalb grundlegend falsch. Es ist deshalb fraglich, ob die Standardmethodik zur Signifikanzbestimmung abhängiger Messgrößen, Zeitreihenanalysen methodisch angemessen oder nicht noch "falscher" wäre. Zumal sie nur für eine höhere Anzahl von Vergleichsfällen geeignet ist.

den höchsten Precisionwert von allen vier Maschinen, Rangplatz 2 den zweithöchsten usw.. Beispiel: bei der Suchanfrage "arbeitsrecht kündigung" erreichte Google eine Precision von 70%, Lycos von 65%, AltaVista von 15% und Fireball eine Precison von 35%. Google erzielt somit Rangplatz 1, Lycos Rangplatz 2, Altavista Rangplatz 4 und Fireball Rangplatz 3.

Google ist auch bei dieser Sichtweise die Maschine mit der höchsten Effektivität. Die Suchmaschine erreicht bei fast der Hälfte der Suchanfragen den höchsten Precisionwert (Rangplatz 1) und belegt bei keiner Suchanfrage Rangplatz 4. Lycos erreicht die zweite Position, da sie häufiger Rangplatz 1 und 2 erzielt als Fireball und Altavista. Die Interpretation ist aber nicht ganz eindeutig, denn Altavista erreicht z.B. häufiger Rangplatz 1 als Lycos. Ein intuitiver Vergleich zwischen Fireball und Altavista scheint aufgrund der Ergebnisgrafik nicht sinnvoll, da Altavista alle vier Rangplätze ungefähr gleichhäufig belegt und bei Fireball "das Mittelfeld", also Rangplatz 2 und Rangplatz 3, dominiert.

Die statistische Überprüfung zeigt auch hier, dass Google signifikant besser ist als die anderen Maschinen. Die Unterschiede zwischen Lycos, Altavista und Fireball halten der statistischen Überprüfung hingegen nicht stand.

Insgesamt betrachtet, ist dieses durch den Vergleich der Rangplätze ermittelte Ergebnis eher skeptisch zu bewerten, wie das folgende Beispiel deutlich macht. Bei der Suchanfrage 3 "eheschließungen entwicklung weltweit" erreicht Google eine Precison von 0%, Lycos 10%, Fireball 10% und Altavista 20%. Die vergebenen Rangplätze sind: Google Rangplatz 3, Lycos Rangplatz 2, Fireball Rangplatz 2 und Altavista Rangplatz 1.

Das bedeutet zunächst, dass die Ergebnisse nummerisch leicht positiv verfälscht werden, da Rangplätze bei gleich hoher Precision mehrfach vergeben und somit vordere Rangplätze häufiger belegt werden als hintere. Suchanfrage 3 zeigt aber nicht nur diese eher "kosmetische Verzerrung", sondern insbesondere auch ein inhaltlich bedeutendes, weil qualitativ schwerwiegendes Problem auf. Google erzielt bei Suchanfrage 3 eine Precision von 0 und erreicht damit Rangplatz 3. Dabei weist Google zu Suchanfrage 3 keinen relevanten Treffer nach, kann die Suchanfrage also überhaupt nicht beantworten, während die anderen Maschinen zumindest 2 relevante Treffer referenzieren, das Informationsbedürfnis also zumindest graduell befriedigen.

Diese rein quantitative Ergebnisbetrachtung reflektiert also in keiner Weise den grundlegenden qualitativen Gegensatz zwischen „liefert die wenigsten

Treffer“ und „kann die Suchanfrage nicht beantworten“. Dabei ist es ein erheblicher Unterschied, ob eine Suchmaschine eine Suchanfrage, wenn auch eher schlecht, beantworten kann oder nicht.

Wie die Problematik dieser sogenannten Nullantworten[24] aufzulösen ist, bleibt ungeklärt. Da dieser Unterschied nicht adäquat quantifiziert werden kann, wird hier darauf verzichtet, diesen, etwa durch die Vergabe von “Strafpunkten“, in die Berechnung mit einzubeziehen.

Stellt man die Frage, wieviele Suchanfragen die Suchdienste mit zumindest einem relevanten Treffer beantworten können, ergibt sich das Ergebnis aus Tabelle 5. Altavista ist die Maschine, welche die höchste Zahl an Suchanfragen mit zumindest einem relevanten Treffer beantwortet. Google folgt auf Rang zwei und vermag zwei Anfragen weniger zu beantworten.

Suchmaschine	Anzahl beantworteter Suchanfragen (max. 56)	Nicht beantwortet und mehr als 20 Treffer
Google	53	2
Lycos	50	2
Fireball	51	2
Altavista	55	1

Tabelle 5: Anzahl beantworteter Suchfragen bzw. ohne relevante Treffer

7 Bewertung der Ergebnisse

In einem Satz zusammengefasst lautet das Ergebnis dieser Untersuchung: Google ist effektiver als die anderen Suchmaschinen. Zwischen den anderen Maschinen lässt sich kein Unterschied bei der Retrievaleffektivität feststellen. Zu beachten bleibt, dass dieses Ergebnis sich nur auf die zugrundeliegenden Zahlenwerte stützt, d.h. davon ausgeht, dass diese objektiv ermittelt werden konnten. Die Verzerrungen, die sich durch die zeitliche Divergenz zwischen Anfrage- und Beurteilungszeitpunkt sowie die Nulltreffermengen ergeben, bleiben unberücksichtigt.

Die Problematik der Nulltreffersuchanfragen, zeigt sehr deutlich, dass die je nach Bewertungssicht und Bewertungsmaß variierenden Effektivitätswerte immer von den Faktoren präjustiert werden, die dem jeweiligen Retrievaltest immanent sind.[25]

---

[24] Zur Problematik der Nullantworten siehe [Womser-Hacker 1989], S.151-156, vgl. auch http://www.uni-hildesheim.de/~einf_iw/S10_Eval_IRS.pdf 16.04.02

[25] Untersuchungsziel, Bewertungsmaße, Kriterien der Relevanzbeurteilung, Art und Ausgestaltung der Informationsbedürfnisse, Durchführung der Suchanfragen.

Wie sind die Ergebnisse dieses Testes also einzuordnen? Den normativen Rahmen dieser Untersuchung bilden primär die Thematik, Komplexität und Spezifität der Fragestellungen. Suchanfragen und Informationsbedürfnisse, die freizeitbezogen sind, Produkt- und Dienstleistungsangebote kommerzieller Anbieter zum Ziel haben oder erotischer/pornographischer Natur sind, werden ausgeschlossen. Die Suchanfragen dieser Untersuchung sind thematisch eher dem (sozial)wissenschaftlichen, politischen Umfeld zuzuordnen. Die Ergebnisse dieser Untersuchung gelten nur für solche Informationsbedürfnisse und Suchanfragen.

Diese Ergebnisse beruhen dabei ausschließlich auf der Beurteilung der Relevanz der ersten 20 Ergebnisseiten. Mehrwerte, wie die Hilfe zur Vorab-Relevanzbeurteilung durch die Metainformation der Trefferliste der Suchmaschinen, das Browsing in thematisch passenden Rubriken[26] oder der Zugriff auf nicht mehr vorhandene Webseiten mit Hilfe des Google Archivs bleiben in der Testanordnung unberücksichtigt. Die Suchmaschine Google bietet dabei mehrere solcher Mehrwerte als Alleinstellungsmerkmale. Beispiele hierfür sind u.a. Zugriff auf PDF- und Postscript-Dokumente oder der Nachweis nicht mehr verfügbarer Dokumente über das Google Archiv. Da diese Alleinstellungsmerkmale den Rechercheerfolg bei Google im Vergleich zu anderen Maschinen zumindest tendenziell begünstigen, ist davon auszugehen, dass Google für den Nutzer in weitaus größerem Maße die beste Alternative unter den untersuchten Suchmaschinen darstellt, als die Ergebniswerte allein vermuten lassen.

8 Einschätzung des Retrievaltests

Die Zielsetzung dieses Retrievaltests ist es, die Retrievaleffektivität der untersuchten Suchmaschinen zu ermitteln und das in [Griesbaum 2000] konzipierte Evaluationsverfahren zu optimieren und weiterzuentwickeln. Zunächst kann festgehalten werden, dass die auf [Tague-Sutcliffe 1992] beruhende methodische Konzeption geeignet erscheint, den Retrievaltest strukturiert zu konzipieren und bei auftauchenden Schwierigkeiten Problemlösungsstrategien vor der Testdurchführung zu entwickeln. Die von [Hawking et al. 2001] genannten Rahmenbedingungen für die Evaluation von Suchmaschinen können eingehalten werden. Es werden reale Informationsbedürfnisse eingebunden, die durch eine große Anzahl von Suchfragen repräsentiert wurden. Ebenso werden die wichtigsten Suchmaschinen für den deutschen Sprachraum in der Untersuchung berücksichtigt. Inwieweit Suchfrageformulierungen ein Informationsbedürfnis

[26] Beispielsweise bei den Katalogtreffern von Lycos und Google.

gut repräsentieren, bleibt natürlich immer schwer zu beantworten. Aber durch die meist kurze Frageformulierung ist zumindest die Länge der Suchfragen den Gegebenheiten im Web angepasst.

Die Optimierung des Evaluationsverfahrens soll in dieser Untersuchung durch die Entwicklung und Verwendung eines automatischen Abfrageskripts umgesetzt werden, das im Vergleich zur „händischen" Durchführung der Suchanfragen in der genannten Untersuchung, durch die Minimierung der zeitlichen Verzerrungen zwischen Anfrage- und Bewertungszeitpunkt, zugleich Validitäts- und Effizienzsteigerungen bewirken soll. Dieses Ziel wird nicht vollständig erreicht, da der Aufwand, ein solches Skript zuverlässig auszugestalten und persistent am Laufen zu halten, sehr hoch ist. Da die Suchmaschinen zu jedem Zeitpunkt Modifikationen durchführen, z.B. die Abfrage-URL ändern können, kann letztlich keine Garantie dafür gegeben werden, dass es zum benötigten Zeitpunkt überhaupt funktioniert. Aus diesem Grund werden bei der Testdurchführung aus Sicherheitserwägungen die Suchanfragen im Voraus durchgeführt und somit im Maximalfall zwei Tage zeitlicher Verzögerung zwischen Suchanfragendurchführung und Relevanzbeurteilung bewusst in Kauf genommen.

Die Nulltrefferproblematik zeigt, dass die Validität der Untersuchungsergebnisse auch von den Ergebniswerten selbst abhängig ist. Hätten beispielweise alle Suchmaschinen zu allen Anfragen mindestens einen relevanten Treffer geliefert, so wäre dieses Problem gar nicht aufgetreten. Prinzipiell ist zu fragen, inwieweit das Ziel, die Testparameter realitätsnah auszugestalten, durch die Annäherung an sogenannte typische Verhaltensmuster von Suchdienstenutzern erreicht werden kann, wenn diese typischen Verhaltensmuster durch die quantitative Generalisierung qualitativ unspezifizierter Interaktionsdaten, z.B. durch Logfileanalyse, identifiziert werden und gerade deshalb nicht zutreffen müssen. Hier stellt sich wieder das für Retrievaltests zentrale Problem der qualitativen Ausgestaltung der Testparameter, für das keine theoretische Lösung vorliegt. Eine einfache Lösung ist nicht abzusehen, umso dringender ist der Bedarf, Verfahren und Methoden zu entwickeln, die solche qualitativen Probleme bei scheinbar objektivierbaren Kriterien aufdecken.

Hinzu kommt ein weiteres Problem, welches die Annahme, dass die Herkunft der Treffer mit vertretbarem Aufwand verschleiert werden kann, grundsätzlich in Frage stellt. Die Alleinstellungsmerkmale von Google.de hinsichtlich indexierter Dokumentformate (PDF, Postscript) werden zwar in die Testanordnung mit integriert, aber beim Google Archiv gelingt das nicht. Hier stellt sich also die Frage, ob die grundlegende Testanforderung, die

Herkunft der Ergebnisse unkenntlich zu machen, weiterhin durchzuhalten ist oder durch Weiterentwicklungen der Suchmaschinen bei der Ergebnispräsentation künftig nicht obsolet wird.[27]

Fasst man diese Probleme und Schwierigkeiten zusammen, lässt sich Folgendes festhalten:

Das bei diesem Retrievaltest verfolgte Ziel, die Validität und Effizienz der Testanordnung durch die Entwicklung eines automatischen Anfrageskript zu erhöhen, wird aus technischen Gründen nicht erreicht, die Effizienz jedoch erhöht. Im Laufe des Untersuchung zeigt sich, dass sich die realitätsnahe Ausgestaltung der Testparameter weit schwieriger gestaltet, als in der Vorgängeruntersuchung und weitaus weniger umgesetzt werden kann. Dies liegt zum einen in Ergebniswerten begründet, die qualitative Annahmen des Testdesigns in Zweifel ziehen (Top 20 Grenze), hat zum anderen aber auch ihre Ursache in den Merkmalen von Suchmaschinen (Google Cache), die Grundanforderungen der Testanordnung (Unkenntlichmachung der Treffer) grundlegend in Frage stellen.

Aus den Problembereichen der Untersuchung wird ein breites Optimierungspotenzial bei künftigen Untersuchungen, insbesondere bezüglich der Durchführung der Suchanfragen, der Anzahl der bewerteten Dokumente und der Ergebnisdarstellung ersichtlich. Daraus lässt sich schlussfolgern, dass bei künftigen Untersuchungen mehr Ressourcen für die realitätsnahe Ausgestaltung der Testparameter bereit gestellt werden sollen.[28]

9 Literatur

Bager, J. & Schulzki-Haddouti, C. (2001). Alle gegen Google. Wisenut, Teoma, Vivisimo - neue Konkurrenz für die populäre Suchmaschine. c't Magazin für Computer Technik Nr.19, 104-108

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. Journal of the American Society for Information Science and Technology, 53 Nr.4, 308-319

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

[27] Vgl. die Visualisierung der Suchergebnisse bei dem Metadienst KartOO URL http://www.kartoo.com 11.05.02.

[28] Die Autoren bedanken sich bei den Studierenden der Informationswissenschaft in Konstanz und Düsseldorf, die an dem Retrievaltest mit viel Interesse und Engagement teilgenommen haben. Besonderer Dank gebührt Caroline Berns, Marion Herb und Birgit Scherer, die im Rahmen studentischer Projekte an der Durchführung und Auswertung beteiligt waren.

Dennis, S., Bruza, P., und McArthur, R. (2002). Web searching: a process-oriented experimental study of three interactive search paradigms. Journal of the American Society for Information Science and Technology, 53 Nr.2, 120-133

Dresel, R., Hörnig, D, Kaluza, H., Peter, A., Roßmann, N., und Sieber, W. (2001). Evaluation deutscher Web-Suchwerkzeuge. Ein vergleichender Retrievaltest. nfd Information - Wissenschaft und Praxis, 52 Nr.6, 381-392

Ellis, D. (1992). Paradigms and proto-paradigms in information retrieval research. In: Conceptions of library and information science: historical, empirical, and theoretical perspectives. Vakkari, P. and Cronin, B. (eds.); London: Taylor Graham, 165-186

Ford, N., Miller, D., und Moss, N. (2001). The role of individual differences in Internet searching: an empirical study. Journal of the American Society for Information Science, 52 Nr.12, 1049-1066

Gordon, M. & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing & Management, 35 Nr.2, 141-180

Griesbaum, J. (2000). Evaluierung hybrider Suchsysteme im WWW. http://www.inf.uni-konstanz.de/~griesbau/files/evaluierung_hybrider_suchsysteme_im_www.pdf 22.04.2002

Hawking, D., Craswell, N., Bailey, P., und Griffiths, K. (2001). Measuring search engine quality. Information Retrieval, 4 Nr.1, 33-59

Jansen, B. J., Spink, A., und Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. Information Processing & Management, 36 Nr.2, 207-227

Körber, S. (2000). Suchmuster erfahrener und unerfahrener Suchmaschinennutzer im deutschsprachigen World Wide Web. Ein Experiment. http://kommunix.uni-muenster.de/IfK/examen/koerber/

Leighton, H. V. & Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). Journal of the American Society for Information Science, 50 Nr.10, 870-881

Lesk, M. (1995). The seven ages of information retrieval. In: Conference for the 50th anniversary of As We May Think. MIT, 12-14

Mettrop, W. (2001). Internet search engines-fluctuations in document accessibility. Journal of Documentation, 57 Nr.5, 623-651

o.V. (2001). Internet-Suchmaschinen. Google trifft am besten. Test.Stiftung Warentest Nr.9, http://www.warentest.de

Oppenheim, C., Morris, A., McKnight, C., und Lowley, S. (2000). Progress in documentation the evaluation of WWW search engines. Journal of Documentation, 56 Nr.2, 190-211

Robertson, S. E. (1981). The methodology of information retrieval experiments. In: Information Retrieval Experiment. Jones, K. Sparck (ed.); London: Butterworth, 9-31

Stock, M & Stock, W. (2000). Internet-Suchwerkzeuge im Vergleich. Retrievaltest mit Known Item Searches. Password, 11, 23-31

Su, L. T. & Chen, H.-l. (1999). Evaluation of Web Search Engines by Undergraduate Students. In: Knowledge: Creation, Organization and Use. ASIS '99 - Proceedings of the 62nd ASIS Annual Meeting. Woods, L. (ed.); Information Today, Inc., 98-114

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. Information Processing & Management, 28 Nr.4, 467-490

Voorhees, E. M. & Harman, D. (2001). Overview of TREC 2001. 2001, http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf, 01.05.02

Wang, H., Xie, M., und Goh, T. N. (1999). Service quality of Internet search engines. Journal of Information Science, 25 Nr.6, 499-507

Warner, J. (2000). In the catalogue ye go for men: evaluation criteria for information retrieval systems. ASLIB Proceedings, 52 Nr.2, 76-82

Wolff, C. (2000). Effektivität von Recherchen im WWW. Vergleichende Evaluierung von Such-und Metasuchmaschinen. In: Informationskompetenz - Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposiums für Informationswissenschaft. Knorz, G. & Kuhlen, R. (eds.); Universitätsverlag Konstanz, 31-48

Womser-Hacker, C. (1989). Der PADOK Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Systemen. Hildesheim: Georg Olms