In: Knorz, Gerhard; Kuhlen, Rainer (Hg.): Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposiums für Informationswissenschaft (ISI 2000), Darmstadt, 8. – 10. November 2000. Konstanz: UVK Verlagsgesellschaft mbH, 2000. S. 31 – 48

Effektivität von Recherchen im WWW

Vergleichende Evaluierung von Such- und Metasuchmaschinen

Christian Wolff
Universität Leipzig, Institut für Informatik,
Abt. Automatische Sprachverarbeitung
Augustusplatz 10/11, 04109 Leipzig
Wolff@informatik.uni-leipzig.de

Zusammenfassung

Der vorliegende Beitrag befasst sich mit der Informationssuche im World Wide Web und ihrer Evaluierung. Ausgehend von einer Zusammenschau der wesentlichen Merkmale des World Wide Web als heterogener Dokumentkollektion (Kap. 1) werden Aufbau und Merkmale von Suchmaschinen und Metasuchmaschinen eingeführt sowie die Problematik der Evaluierung von Suchmaschinen und eine Übersicht bisheriger Ergebnisse diskutiert (Kap. 2). In Kap. 3 werden Aufbau, Durchführung und Ergebnisse einer Vergleichstudie vorgestellt, bei der mit Hilfe eines *paper-and-pencil*-Experiments ausgewählte Such- und Metasuchmaschinen evaluiert wurden. Schließlich zieht Kap. 4 Schlussfolgerungen aus dieser Studie und gibt einen Ausblick auf Optimierungsmöglichkeiten für Suchmaschinen.

Abstract

This paper is concerned with the evaluation of web search engines. In particular, an empirical comparison of search engines with meta search engines is described. Starting from some introductory notes on basic characteristics of the web as a document collection (Ch. 1), the main features of search and meta search engines are given and relevant literature on search engine evaluation is presented (Ch. 2). Design, execution, and results of this empirical study follow in Ch. 3. Finally, Ch. 4 discusses major implications from these results and gives some hints for further study and possible optimisation strategies for search engines.

1. Einleitung: Charakteristika des World Wide Web

Das World Wide Web als größter elektronisch verfügbarer Informationsbestand stellt neue Herausforderungen für die Informationserschließung: Suchmaschinen als Information Retrieval-Systeme stellen neben *Web Directories* wie *Yahoo,* die auf intellektuell erstellten thematischen Hierarchien aufbauen, das wichtigste Werkzeug für die Informationssuche dar und werden auch in vielen Fällen als WWW-Portale, d. h. als primäre Zugangsseiten genutzt. Man kann Suchmaschinen zwar mit traditionellen IR-Systemen vergleichen, sie unterscheiden sich von diesen aber in besonderer Weise durch die Heterogenität und Variabilität der von ihnen indexierten Datenbestände im World Wide Web. Diese sind durch folgende Merkmale gekennzeichnet:

- Anzahl und Umfang der Informationseinheiten wachsen sehr dynamisch; gleichzeitig sind die Daten relativ unbeständig, d. h. ihre Verfügbarkeit über einen längeren Zeitraum ist nicht gesichert (vgl. OCLC 1999).

- Die Daten ein sehr hohes Volumen, derzeit ca. 10^9 Dokumente.

- Die Daten sind in der Regel semi- oder unstrukturiert, d. h. sie können deklaratives Markup in gewissem Umfang enthalten (HTML-Dateien); es existiert eine Vielfalt von Medientypen und Formaten.

- Die Ressourcen liegen in unterschiedlichen Sprachen vor, bei eindeutiger Dominanz des Englischen: Dessen Anteil liegt bei etwa 50 – 80 %, gefolgt von Japanisch und Deutsch mit je etwa 4 %, vgl. PIMIENTA et al. 1998.

- Durch die allgemein verfügbaren Publikationsmöglichkeiten im World Wide Web, bei denen in vielen Fällen keine Qualitätskontrolle stattfindet, sind die Inhalte unterschiedlicher (inhaltlicher, formaler) Qualität.

Eine einheitliche Methodologie für die Bestimmung der wesentlichen Merkmale des World Wide Web beginnt sich – unter dem Einfluss der *Web Characterization Activity* des *World Wide Web Consortium* – erst herauszubilden (vgl. OCLC 1999, LAVOIE & FRYSTYK NIELSEN 1999). PITKOW 1998 gibt einen Literaturüberblick zur den bisherigen Ansätzen der Kenngrößenerfassung und Charakterisierung des World Wide Web.

2. Suchmaschinen

Die Informationserschließung im World Wide Web kann grundsätzlich auf unterschiedliche Arten erfolgen:

- Durch direkte Anwahl einer bekannten Adresse (URL), ggf. in Verbindung mit einer lokalen Suche auf einem Webserver,

- durch die Verwendung eines Verzeichnisdienstes wie etwa *Yahoo* (vgl. *http://www.yahoo.com, http://www.yahoo.de),* der Ressourcen aus dem World Wide Web klassifiziert und ordnet,

- durch die Verwendung einer Suchmaschine, die über einen Index eines Teils der im World Wide Web verfügbaren Ressourcen verfügt und auf Anfrage Nachweise (d. h. Adressen und Kurzbeschreibungen) von Dokumenten im World Wide Web liefern kann oder

- durch Metasuchmaschinen, die zwar selbst keinen Index verwalten, aber die Ergebnismengen einer Mehrzahl von Suchmaschinen bündeln.

Es besteht dabei das grundsätzliche Problem, dass die (leistungsfähigeren) Suchmaschinen im World Wide Web kommerzielle Softwaresysteme sind, über deren inneren Aufbau nur wenige Details bekannt sind. In Kenntnis der Verfahren des Information Retrieval und durch Analyse der von den Suchmaschinen erzielten Ergebnisse und der dabei verwendeten Anfragelogik lassen sich aber Rückschlüsse bezüglich ihrer Funktionsweise ziehen.

2.1 Aufbau von Such- und Metasuchmaschinen

Eine Suchmaschine für das World Wide Web ist in der Regel aus folgenden Komponenten aufgebaut:

1. Einem oder mehreren Erfassungsagenten (spider, crawler), die Dateien aus dem World Wide Web zur Indexierung laden,

2. einer Indexierungskomponente, die die Daten auswertet und den Index aufbaut und verwaltet,

3. der *Abfrageschnittstelle,* über die der Benutzer seine Anfrage an die Suchmaschine formulieren kann, und

4. der *Anfrageverarbeitung,* die die Anfrage mit dem Index abgleicht (Retrievalfunktion) und die Ergebnisse an die Benutzerschnittstelle der Suchmaschine weiterleitet.

Im Unterschied zu „einfachen" Suchmaschinen verwalten Metasuchmaschinen keinen eigenen Index. An die Stelle der Crawler-Komponente tritt bei ihnen die Schnittstelle zu verschiedenen Suchmaschinen, die von der Metasuchmaschine abgefragt wird.
Die verschiedenen Suchergebnisse werden anschließend von der Metasuchmaschine unifiziert. Das Ziel ist die Optimierung der Retrievaleffektivität, da anzunehmen ist, dass aufgrund der geringen Überlappungsraten in den Inhalten einzelner Suchmaschinen eine solche Kombination von Ergebnissen die

Verbesserung der Suchresultate nach sich zieht. Dazu fehlen allerdings bisher empirische Studien. Eine Metasuchmaschine verfügt über Suchagenten für jede Suchmaschine und übersetzt die Anfrage des Benutzers in die jeweilige Abfragesprache der Einzelsuchmaschine, was zur Folge hat, dass Metasuchmaschinen oft nur eine im Vergleich zu einzelnen Suchmaschinen eingeschränkte Mächtigkeit der Anfragesprache aufweisen. Zu den bekannten Metasuchmaschinen zählen *SavvySearch (http://www. savvysearch.com,* vgl. DREILINGER 1996), *MetaCrawler (http://www.metacrawler.com)* und *MetaGer (http://www.metager.de,* für den deutschsprachigen Bereich).

2.2 Indexierung und Abdeckungsraten

Bei der Indexierung des World Wide Web durch Suchmaschinen entsteht analog zur Architektur eines Information-Retrieval-Systems eine invertierte Datei (Index). Die Analyse erfolgt auf der Basis des Volltexts jeder Webseite; zusätzlich können Strukturmerkmale (HTML-Marken für Schlagwörter, Überschriften, Hervorhebungen) bei der Indexierung berücksichtigt sein. Die invertierte Datei enthält als Beschreibungsmerkmale für jedes Dokument in der Regel nicht nur die Adresse (URL) selbst, sondern darüber hinaus eine Kurzbeschreibung des Inhalts (Textanfang, Überschrift, Textextrakt) und formale Angaben (z. B. letztes Aktualisierungsdatum). Zu den Leistungsmerkmalen von Suchmaschinen gehören ihre

- Datenbankgröße und ihr Abdeckungsgrad hinsichtlich des WWW,
- die Qualität und Aktualität der gespeicherten Daten,
- die Anfragemerkmale und das Retrievalmodell sowie
- die Art der Ergebnisaufbereitung.

	Anzahl der indexierten Webseiten			
	Search Engine Showdown		*Search Engine Watch*	
Suchmaschine	*XI/99*	*VII/00*	*XI/99*	*VI/00*
Northern Light	*200*	*282*	*189*	*265*
Fast Search	*192*	*327*	*200*	*340*
AltaVista	191	331	*250*	*350*
Google!	126	*355*	*85*	*560*
Anzwers	*79*	108	—	
j Won	*78*	*356*	—	—
Excite	*71*	*159*	150	*250*
Snap	*50*	*278*	—	—
HotBot	*39*	*280*	—	—

Tabelle 1: Datenbankumfang bekannter Suchmaschinen (Millionen URLs)

Aktuelle Statistiken zeigen, dass die umfangreichsten Indizes von Suchmaschinen eine Größenordnung von mehreren hundert Millionen Einträgen erreichen. Tab. 1 stellt dabei die Daten aus zwei bekannten Datenquellen für den Suchmaschinenvergleich (NOTESS 1999f, SULLIVAN 19990 für den Zeitraum des Testdesigns (XI/99) sowie aktuelle Daten für Mitte 2000 gegenüber.

Untersuchungen von NOTESS 1999f und LAWRENCE & GILES 1998 ergeben, dass diese Indizes *weitgehend disjunkt* sind, d. h. der indexierte Datenbestand differiert stark zwischen den verschiedenen Suchmaschinen. LAWRENCE & GILES 1998, 1999 kommen zu folgenden Ergebnissen: Derzeit deckt keine Suchmaschine mehr als 16 % des World Wide Web ab; generell werden Sites, auf die viele externe Verknüpfungen verweisen, bei der Indexierung berücksichtigt. Hinsichtlich der Domänen gilt dies auch für kommerzielle Sites (. com) sowie in den USA betriebene Webserver. Eine ebenfalls hohe Schwankungsbreite existiert bei den „toten Links", d. h. nachgewiesenen HTML-Seiten, die aufgrund des Alters des Index bereits nicht mehr existieren. Je nach Suchmaschine ist von einem Anteil zwischen 3 % und 20 % auszugehen (vgl. NOTESS 1999f, LEIGHTON & SRIVASTAVA 1999: 880 f., Appendix B, C). Mangelnde Abdeckungsraten bei gleichzeitiger Verschiedenartigkeit der indexierten Dokumentmengen ergeben eine starke Grundplausibilität für das Potential von Metasuchmaschinen hinsichtlich der Optimierung der Rechercheeffektivität im WWW. Diese Beobachtung ist daher die primäre Motivation für die in Kap. 3 beschriebene Evaluierungsstudie.

2.3 Retrievalmodelle und Anfragesprachen

Die von den Suchmaschinen verwandten Retrievalmodelle bauen auf Booleschen und statistischen Verfahren auf; in der Regel besteht für den
Benutzer sowohl die Möglichkeit, Operatoren der Booleschen Logik bei der Anfrage zu verwenden, als auch eine natürlichsprachliche Eingabe (mit praktisch beliebiger Länge) ohne Verwendung von Suchoperatoren zu verwenden. Tabelle 2 fasst Merkmale der Anfragesprachen und ihre Ausprägungen zusammen (nach NOTESS 1999f vgl. auch SCHWARTZ 1998: 975 f.).

Merkmal	*Mögliche Ausprägungen*
Defaultstrategie	Konjunktion, Disjunktion
Boolesche Operatoren	AND, OR, NOT,(), sowie die vereinfachte Syntax: +, -
Abstandsoperatoren	NEAR, exakte Phrase
Trunkierung	Rechtstrunkierung, Einzelzeichenmaskierung
Morphologische Expansion	v. a. automatische Pluralbildung im Englischen (selten)
Suche in Feldern	Titel, URL, Website
Inhaltliche und sprachliche	Dokumenttyp, Sprache, Datum (Sprachheuristik i. d. R. über *top level domains)*
Stoppworteliminierung	ja/nein

Tabelle 2: Recherchemöglichkeiten und Anfragesprachen von Suchmaschinen

Betrachtet man die Ausprägungen der in Tab. 2 genannten Merkmale, so fällt auf, dass sich bisher kein Konsens hinsichtlich der Interpretation von Benutzeranfragen *ohne Operatorenverwendung* herausgebildet hat: Die Suchmaschinen verwenden unterschiedliche Default-Strategien bei der Anfrageinterpretation. Dies ist vor allem deswegen bemerkenswert, weil Benutzer in der Regel kurze Anfragen ohne Operatoren stellen. Soweit verschiedene Suchmaschinen dies logisch unterschiedlich interpretieren, kann sich kein allgemein akzeptiertes Interpretationsmodell für Suchanfragen herausbilden, mit der Folge, dass Operatoren vielfach falsch eingesetzt werden (vgl. JANSEN, SARACEVIC & SPINK 2000).

2.4 Evaluierung von Suchmaschinen: state-of-the-art

Die Mehrzahl der bisher vorliegenden Studien zur Retrievaleffektivität von Suchmaschinen beschränkt sich auf einfache Maße wie die Anzahl der nachgewiesenen Treffer hinsichtlich einer Anfrage, ohne aber eine

Effektivitätsbewertung im engeren Sinn durchzuführen.[1] Recherchen im World Wide Web stellen im Kontext des IR insofern ein Novum dar, als ein wesentlich größerer Nutzerkreis ohne technische Vorkenntnisse mit Suchmaschinen als Information Retrieval-Systemen interagiert, vgl. dazu KIRSCH 1998, JANSEN et al. 1998 und SILVERSTEIN 1999. KIRSCH 1998: 4

berichtet beispielsweise, dass für die Suchmaschine *Infoseek* nur 1 % aller Anfragen mit Hilfe von *advanced search features* erfolgen. Intelligente Anfrageunterstützungsmechanismen, wie sie zeitweilig von *AltaVista* oder *Lycos* angeboten worden waren, sind mittlerweile wieder aus dem Angebot verschwunden, vgl. SCHWARTZ 1998: 977, Abb. 1: *Lycos Pro Power Panel* und 980: Abb. *2: AltaVista Refine.* Nach ersten Studien lässt sich ein „typischer Suchmaschinennutzer" wie folgt charakterisieren:

- Es werden kurze Anfragen gebildet (weniger als drei Suchbegriffe im Mittel, vgl. JANSEN et al. 1998, JANSEN, SPINK & SARACEVIC 2000).

- BOOLEsche Operatoren werden kaum verwendet (weniger als 10 % der Anfragen), positive und negative Auftretensoperatoren (+/-) in derselben Größenordnung. Zudem enthalten zahlreiche Anfragen Fehler hinsichtlich der Verwendung von Operatoren (JANSEN, SPINK & SARACEVIC 2000: 217).

- Benutzer betrachten nur selten mehr als zwei Ergebnisseiten (ein *cut-off-Wert* bei der Bewertung der Suchergebnisse von weniger als 30, meist wird nur die erste Ergebnisseite betrachtet). SILVERSTEIN et al. 1999: 10 analysieren mehrere hundert Millionen Anfragen und berichten , dass bei etwa für 85% aller Anfragen nur die erste Ergebnisseite betrachtet wird, bei JANSEN, SPINK & SARACEVIC 2000: 215 liegt dieser Wert bei „nur" 58%.

Eine umfassende benutzerbezogene Studie zur Retrievaleffektivität haben GORDON & PATHAK 1999 vorgelegt. Sie führen in einem Experiment mit 33 Testpersonen eine vermittelte Recherche an acht verschiedenen Suchmaschinen durch, wobei die Suchergebnisse von den Testpersonen nach einer vierstufigen Skala hinsichtlich ihrer Relevanz bewertet werden. Die Ergebnisse zeigen eine hohe Schwankungsbreite der Suchmaschinen sowohl hinsichtlich der *precision* als auch des *recall.* Bei einem cut-off-Wert von 20 betrachteten Dokumenten erreicht die qualitativ beste Suchmaschine (AltaVista) einen *recall* von nur 15 %, selbst bei 200 betrachteten Dokumenten steigt der Wert nicht über 23 % an (GORDON & PATHAK 1999: 160 ff.).

Auch in dieser Studie wird die bereits erwähnte Beobachtung geringer Überlappung der Ergebnismengen verschiedener Suchmaschinen bestätigt: Von insgesamt 160 Treffern (die ersten 20 Treffer von acht Suchmaschinen) wurden 150 Dokumente jeweils nur von einer Suchmaschine nachgewiesen (GORDON &

[1] Ein Überblick über bisherige Evaluierungsstudien für Suchmaschinen im World Wide Web findet sich bei GORDON & PATHAK 1999: 145 ff., insb. 148: Tabelle

PATHAK 1999: 170 ff.). Die bisherigen Ergebnisse zur Bewertung von Suchmaschinen und ihrer Retrievaleffektivität geben eine starke Anfangsplausibilität für die Verwendung von Metasuchmaschinen sowie für die Weiterentwicklung der Retrievalfunktionalität hinsichtlich besserer Benutzerschnittstellen und der Auswertung weiterer Strukturmerkmale der indexierten Texte. Die nachfolgende vergleichende Evaluierung widmet sich daher u. a. der Frage, ob mit Metasuchmaschinen bessere Effektivitätswerte erreicht werden können als mit "einfachen" Suchmaschinen.

3. Evaluierung von Such- und Metasuchmaschinen

Ausgehend von den bisherigen Ergebnissen zur Evaluierung von Suchmaschinen wurde im WS 1999 / 2000 im Rahmen einer Vorlesung zum Thema Information Retrieval am Institut für Informatik der Universität Leipzig eine empirische Untersuchung zur Evaluierung von Suchmaschinen im World Wide Web durchgeführt. Die Versuchspersonen waren Teilnehmer dieser Lehrveranstaltung, d. h. Informatik-Studenten im Hauptstudium mit vertieftem Hintergrundwissen zum Information Retrieval, insbesondere auch zur Problematik der Such- und Metasuchmaschinen. Die Kenntnisse im IR-Bereich umfassen dabei auch traditionelle IR-Systeme und deren Abfragesprachen (z. B. *stn / INSPEC / messenger),* allerdings kaum als Erfahrung aus *aktivem Umgang* mit solchen Systemen.

3.1 Aufbau und Durchführung der Untersuchung

Ein wesentliches Erkenntnisziel der Untersuchung war der Vergleich der *Retrievaleffektivität* von Such- und Metasuchmaschinen. Ausgehend von den Beobachtungen zur Abdeckungsrate von Suchmaschinen (vgl. Kap. 2.2), ist es naheliegend, zu untersuchen, ob Metasuchmaschinen durch Verbindung verschiedener Suchmaschinen — und damit deutlich unterschiedlicher indexierter Untermengen des World Wide Web — zu besseren Retrievalergebnissen kommen als Suchmaschinen.

Eine zweite Fragestellung ist die Unterscheidung von Anfragetypen nach fachspezifischen Anfragen (hier: mit Bezug zu Fragestellungen aus der Informatik) und nach Fragen von allgemeinem persönlichen Interesse der Testpersonen. Da es bisher keine differenzierte Theorie der Typisierung von Informationsbedürfnissen im World Wide Web gibt, erscheint eine solch einfache Aufteilung zulässig. Unterscheidungen wie sie BAEZA-YATES und & RIBEIRO-NETO 1999: 91 f. treffen *(specific queries, broad queries, vague queries),* haben kaum eine ausreichende Trennschärfe. Erste Benutzerstudien für das Web Retrieval, wie GORDON & PATHAK 1999, nehmen ebenfalls keine Typisierung der Anfragen vor, sondern zeigen lediglich die thematische Bandbreite ihrer Testanfragen auf (vgl. GORDON & PATHAK 1999: 150, Tabelle 2 und 177 (Beispielanfrage) und LEIGHTON & SRIVASTAVA 1999).

3.1.1 Testkonzeption

Die Evaluierungsstudie ist als zweistufiges paper-and-pencil-Experiment angelegt: Ähnlich der von GORDON & PATHAK 1999 vorgelegten Studie orientiert sich das Versuchskonzept dabei an der — für die Recherche im World Wide Web an sich nicht üblichen — vermittelten Recherchesituation (vgl. auch MIELKE 2000: 108ff): Den Versuchspersonen lag dabei eine aus BERGMANN et al. 1999:3f entlehnte Übersicht von Hinweisen zur Formulierung von Suchanfragen für WWW-Suchmaschinen vor.

In einem ersten Schritt wurden die Versuchspersonen gebeten, unter Anwendung eines vorgegebenen Operatoreninventars (+, -, *, „ ") Suchanfragen schriftlich zu formulieren. Dabei war je eine Anfrage von *fachlichem Interesse (z.* B. mit Bezug zu einer gerade besuchten Lehrveranstaltung) sowie eine Anfrage *mit persönlichem Interessenschwerpunkt* zu formulieren. Neben der Formulierung von Suchanfragen mit Standardoperatoren für Suchmaschinen wurden die Versuchspersonen auch gebeten, Suchformulierungen als natürlichsprachlichen Text sowie als Boolesche Anfrageausdrücke (für den Test erweiterter Recherchemodi der Suchmaschinen vorzunehmen). Die ausformulierte Variante der Anfragen diente dabei der inhaltlichen Kontrolle der formalisierten Anfragefassungen, d. h. für Korrekturen bei der Testdurchführung. Zusätzlich zu den Suchanfragen wurden auf einem Fragebogen einige allgemeine Parameter erfasst (Erfahrungen im Umgang mit Suchmaschinen, bevorzugte *search engines,* durchschnittliche wöchentliche Online-Zeit, etc., s. u. Kap. 3.2.1).

3.1.2 Recherchedurchführung

Für die Evaluierung wurden vier Suchmaschinen ausgewählt, je zwei einfache Suchmaschinen und zwei Metasuchmaschinen, von denen je eine seit längerem verfügbar ist bzw. jüngeren Datums ist. Für die Auswahl der Suchmaschinen wurden folgende Kriterien herangezogen:

- Abdeckungsrate bzw. Umfang des Index (vgl. oben Tab. 1)
- Kompatibilität der für die Suchmaschine verwendbaren Operatoren
- Anzahl der angesprochenen Suchmaschinen (nur bei Metasuchmaschinen).

Nach diesen Kriterien wurden folgende Suchmaschinen ausgewählt:

Suchmaschinen	*Metasuchmaschinen*
AltaVista, www.altavista.com	*MetaCrawler, www.metacrawler.com*
Northern Light, www.northernlight.com	*C4, www.c4.com*

Tabelle 3: Für den Test ausgewählte (Meta-)Suchmaschinen

Bei der Konfiguration der Metasuchmaschinen vor Durchführung der Recherchen wurde die maximale Zahl einzelner Suchmaschinen berücksichtigt. Dabei waren

jeweils auch die im Test als „Einzelsuchmaschinen" berücksichtigten Recherchedienste *AltaVista* und *Northern Light* enthalten.

Die formalisierten Anfragen der Testpersonen wurden vom Testleiter — ggf. unter Bereinigung offensichtlicher Fehler[2] — für die Abfrage von Suchmaschinen verwendet. Auf eine gesonderte Evaluierung der Anfragen mit erweiterter Boolescher Logik musste verzichtet werden, da diese nicht in hinreichender Zahl vorlagen. Von den Ergebnismengen der einzelnen Suchmaschinen gingen jeweils nur bis zu 30 Suchergebnisse in qualitätsorientierter Sortierung *(ranking)* in die Auswertung ein. Der *cut-off*-Wert 30 ist für vergleichbare Studien üblich (vgl. MIELKE 2000: 132ff) und läßt sich auch mit Blick auf bisherige Studien zur Suchmaschinennutzung begründen, nach denen selten mehr als zwei Ergebnisseiten zu je 10 Treffern betrachtet werden (vgl. oben Kap. 2.3).

Während bei den Suchmaschinen angenommen werden kann, dass die Ergebnismengen bezüglich einer Anfrage über einen gewissen Zeitraum hin stabil sind, hängt die Ergebnismenge einer Metasuchmaschine deutlich stärker von kurzfristig variablen Parametern wie Übertragungsgeschwindigkeit, Verfügbarkeit der *(primären)* Suchmaschinen und eingestellten *time-out*-Werten ab. Daher wurde für die Metasuchmaschinen jede Anfrage solange wiederholt, bis sich ein Maximum der Treffermenge für eine Anfrage beobachten ließ, d. h. jede Anfrage wurde bis zu 15 mal wiederholt, bis die jeweils höchste Trefferzahl mehrfach aufgetreten war. Von dieser umfangreichsten Treffermenge wurden anschließend die ersten 30 Dokumente zur Auswertung herangezogen.

3.1.3 Effektivitätsbewertung

Die Ergebnismengen der Recherchen, also die Trefferlisten der einzelnen Suchmaschinen, wurden jeweils für die einzelnen Anfragen zusammengefasst, einheitlich formatiert und den Testpersonen als Ausdruck zur Auswertung vorgelegt. Grundlage der Bewertung war dabei *nicht* das nachgewiesene Dokument selbst, sondern die auf den Ergebnisseiten der Suchmaschinen angegebene *Metainformation (i.* d. R. eine URL, Dokumenttitel sowie z. T. Textausschnitte aus dem Trefferdokument). Diese nicht unerhebliche Einschränkung der Auswertungssituation rechtfertigt sich zum einen aus dem großen Aufwand, der mit Ausdrucken und Verteilen von bis zu 4400 Webseiten verbunden gewesen wäre, zum anderen ist die Fragestellung, welche Effektivitätsergebnisse sich aus der Bewertung dieser Metainformation ableiten lassen, für sich genommen legitim.

Als Konsequenz aus dieser Einschränkung wurde nicht die traditionelle binäre Bewertung *relevant (rel.) — nicht relevant (n. rel.)* für die Effektivitätsbewertung verwendet, sondern als dritte Kategorie die Bewertung als *vielleicht relevant (v. rel.)* hinzugenommen, vgl. dazu SALTON & McGILL 1987: 151f. Letztere Kategorie sollte dabei in denjenigen Fällen angewandt werden, in denen sich die

[2] Z. B. Korrektur von Rechtschreibfehlern, Eliminierung redundanter Klammerung und Korrektur von Operatoren. Zur Motivation solcher Eingriffe diente dabei der natürlichsprachliche „Volltext" der Anfragen, der die intendierte Interpretation deutlich werden läßt.

Testpersonen zu einer sicheren Relevanzbeurteilung in Ermangelung des Trefferdokuments nicht in der Lage sahen, aber ggf. zu der entsprechenden Seite im WWW navigieren würden. Die nachfolgenden Ergebnisse sind auf der Basis dieser Vorgehensweise konservativ zu interpretieren, da anzunehmen ist, dass sich bei Bewertung der Dokumente selbst die Relevanzbewertungen verschlechtern würden.

3.2 Testergebnisse

Die Ergebnisse der Evaluierung umfassen zunächst die Daten zur Selbsteinschätzung der Versuchspersonen sowie die Relevanzbewertungen für die einzelnen Suchanfragen.

3.2.1 Fragebogenauswertung

Tab. 4 fasst die Auswertung des Fragebogens zu Vorkenntnissen und Erfahrungen der Versuchspersonen zusammen:

Parameter	*Ergebnis*
Anzahl Versuchspersonen	25
Mittlere wöchentliche Online-Zeit	0 10h, Minimum: 1h, Maximum 30h
Nutzung des WWW (0=nie ... 5=täglich)	0 4,14
Nutzung von Suchmaschinen (0=nie ... 5=täglich)	0 2,98
Bevorzugte Suchmaschinen (mehr als einmal genannt, Anzahl der Nennungen)	Altavista (21), Yahoo (15), Fireball (6), Google (4), Lycos (4), Metac-rawler (3), Excite (2), HotBot (2), Me-tager (2), Web.de (2)
Kenntnisse von HTML (0=keine ... 5=sehr gute)	0 3,04
Eigene Homepage	14 (= 56 %)

Tabelle 4: Ergebnisse der Auswertung des Fragebogens

Die Ergebnisse des Fragebogens bestätigen die Annahme, dass es sich bei der ausgewählten Personengruppe um relativ erfahrene Benutzer von Suchmaschinen handelt, denen deren Funktionsweise gut vertraut ist. Nimmt man hinzu, dass die Versuchspersonen gezielt in die Arbeitsweise von Suchmaschinen und die verfügbaren Suchoperatoren eingeführt wurden, so sind die nachfolgenden Ergebnisse dahin zu bewerten, dass man deutlich schlechtere Effektivitätswerte für „durchschnittliche Websurfer" erwarten kann. Dafür spricht auch die Analyse

der Suchanfragen hinsichtlich der Anzahl der verwendeten Konzepte[3] und Operatoren (Tab. 5): Sie unterscheidet sich deutlich von den Ergebnissen der Auswertung umfangreicher *query logs* von Suchmaschinen, wie sie etwa SILVERSTEIN et al. 1999 vorgelegt haben; dort kommen Operatoren in weit weniger als 10% aller Anfragen zum Einsatz.

Mittlere Anzahl Suchkonzepte	3,96
Anwendung des +-Operators	93%
Anwendung der Rechtstrunkierung (*)	41%
Anwendung des —Operators	20%
Kennzeichnung von Phrasen	20%

Tabelle 5: Ergebnisse der Analyse der Anfragen

3.2.2 Auswertung der Relevanzbewertungen

Bei der Auswertung der Relevanzbewertungen konnten 41 Anfragen berücksichtigt werden, davon 21 fachlich motivierte Anfragen und 20 mit persönlichen Interessenschwerpunkt. Sie enthalten insgesamt 4409 Einzelbewertungen *(rel. / v. rel. / n. rel.).* Aus ihnen lassen sich *precision* und *recall* als elementare Effektivitätskennzahlen errechnen. Aufgrund der relativ

schwach[en] Datenbasis kommt dabei der Abschätzung des *recall* eine untergeordnete Bedeutung zu. Tab. 6 a/b zeigt jeweils Werte für die *precision* als Anteil ermittelter relevanter an allen nachgewiesenen Dokumenten. [4]

	AltaVista			*C4*			*MetaCrawler*			*Northern Light*		
	Anz.	%	%	Anz.	%	%	Anz.	%	%		%	%
n. rel.	693	63,23	63,23	836	69,10	69,10	882	77,64	77,64	700	69,17	69,17
v. rel.	261	**23,81**	**(v + r)**	248	**20,5**	**(v + r)**	156	**13,73**	**(v + r)**	198	**19,57**	**(v + r)**
rel.	142	**12,96**	**36,77**	126	**10,41**	**30,91**	98	**8,63**	**22,36**	114	**11,26**	**30,83**
Σ	1096	100	100	1210	100,01	100	1136	100	100	1012	100	100

Tabelle 6a: Gesamtergebnisse precision nach Suchmaschinen

[3] Als *Suchkonzept* werden hier *inhaltlich unterschiedliche* Suchbegriffe innerhalb einer Anfrage gewertet, nicht aber die Verwendung von Synonymen oder unterschiedlicher flektierter Formen eines Suchbegriffs.

[4] Es werden recall und precision als die klassischen Maße der Effektivitätsbewertung [ver]wendet, vgl. dazu TAGUE 1981:66ff, WOMSER-HACKER 1989:34ff.

	Alle Suchmaschinen		
	Anzahl	%	%
nicht relevant (n. rel.)	3111	*70,56*	*70,56*
vielleicht relevant (v rel.)	818	***18,55***	***(v* + *r)*** **29,44**
relevant (rel.)	*480*	**10,89**	
Summe	4409	100	_100

Tabelle 6b: Gesamtergebnisse precision (addiert)

Tabelle 6 a/b zeigt keinen klaren Vorteil für Metasuchmaschinen; das mit Abstand beste Ergebnis liefert mit 13% bzw. *37% AltaVista,* d. h. bei weiter Interpretation werden zwei von fünf Dokumentennachweisen als *relevant* oder *vielleicht relevant* eingestuft. Angesichts der Tatsache, dass die ausgewählten Metasuchmaschinen auch jeweils beide im Test verwendeten „Einzelsuchmaschinen" abfragen, deutet dies darauf hin, dass es den Metasuchmaschinen bei der Zusammenführung der Einzelergebnislisten nicht gelingt, über das Qualitätsniveau der Einzellisten hinaus zu kommen.

Die Tabellen 7 und 8 zeigen die Differenzierung der Auswertung zur *precision* nach Interessenschwerpunkt der Anfrage und bei cut-off-Wert 10, d. h. bei Betrachtung jeweils nur der obersten 10 Dokumente jeder Ergebnismenge. Bei der Differenzierung nach Interessenschwerpunkte zeigt sich kein wesentlicher Unterschied, was — auch gestützt durch den für die Inhalte des WWW günstigen fachlichen Hintergrund der Versuchspersonen — wenigstens andeutungsweise den Schluß zuläßt, dass die im WWW ʳecherchierbaren Inhalte tatsächliche für eine sehr weite Bandbreite ⁱnhaltlicher Fragestellungen geeignetes Material bereithalten. Dafür spricht auch die Tatsache, dass zu *allen* Anfragen mindestens ein wenigstens *vielleicht relevanter* Nachweis ermittelt werden konnte; bei drei Anfragen konnte keine Suchmaschine ein als *relevant* eingeschätztes Dokument nachweisen.

	Gesamt			*fachliches Interesse*			*privates Interesse*		
	Anzahl	%	%	Anzahl	%	%		%	%
n. rel.	3111	70,56	70,56	1575	71,66	69,17	1536	69,47	69,47
v. rel.	818	**18,55**	**(v+r) 29,44**	419	**19,06**	**(v+r) 28,34**	399	**18,05**	**(v+r) 30,53**
rel.	480	**10,89**		*204*	**9,28**		276	**12,48**	
Summe	4409	100	100	2198	100	100	2211	100	100

Tabelle 7: precision-Werte für fachliche / private Interessenschwerpunkte der Anfragen

Der Anstieg der *precision* bei *cut-off* 10 um etwa 50 % gegenüber dem Wert bei *cut-off* 30 belegt einerseits das Funktionieren des *ranking* (relevante Dokumente befinden sich an den vorderen Plätzen in der Ergebnismenge), erreicht aber insgesamt dennoch keinen zufriedenstellenden Wert (Tab. 8).

	Alle Anfragen cut-off 10		
		%	*%*
nicht relevant (n. rel.)	1005	64,46	69,47
vielleicht relevant (v rel.)	319	**20,46**	**35 53**
relevant (rel.)	235	**15,07**	
Summe	1559	100	100

Tabelle 8: precision-Werte für cut-off-Wert Wert 10

Die Berechnung des mittleren *recall* als Anteil der ***nachgewiesenen*** *relevanten* in Bezug auf ***alle*** *relevanten* Dokumente der Kollektion, d. h. des WWW, ist nicht direkt möglich, da die Anzahl aller relevanten Dokumente zu einer Anfrage nicht bekannt ist.[5] Daher wird nach einer pooling-Methode vorgegangen, bei der die Gesamtzahl der zu einer Anfrage ermittelten relevanten Dokumente als Bezugsgröße verwendet wird (vgl. HARTER 1996: 37).[6] Aufgrund der geringen Anzahl untersuchter Systeme pro Anfrage ist diese Versuchsgröße vermutlich noch relativ weit von der tatsächlichen Gesamtmenge der relevanten Dokumente entfernt. Die ermittelten *recall*-Werte können daher ausschließlich für den Vergleich der Systeme *untereinander* herangezogen werden.

Wie Tab. 9 zeigt, ist hinsichtlich der Unterscheidung von Suchmaschinen und Metasuchmaschinen auch mit Hinsicht auf den *recall* keine klare Tendenz zu erkennen: Wie bei den Ergebnissen für die *precision* liefern je eine Such- und eine Metasuchmaschine *(AltaVista / C4)* die jeweils besten Ergebnissen. Die recall-Werte stehen zudem in Einklang mit den in anderen Evaluierungsstudien beobachteten Werten.[7]

Suchmaschine	*recall (nur rel.)*	*recall (rel. + v. relevant)*
AltaVista	0,235	0,298
Northern Light	0,231	0,212
C4	0,256	0,296
MetaCrawler	0,204	0,193
Gesamt	0,233	0,251

Tabelle 9: recall-Werte bei cut-off-Wert 30

[5] Dies wird durch die Tatsache gestützt, daß die von einzelnen Suchmaschinen erfassten Teilmengen des WWW weitgehend disjunkt sind, vgl. oben Kap. 2.2.

[6] Eine Feindifferenzierung jeweils nach einzelnen Anfragen wäre hier wünschenswert.

[7] Die Werteberechnung für den *recall* erfolgt durch Makromittelung, vgl. WOMSER-HACKER 1988: 67ff, MIELKE 2000: 135. s

4. Fazit & Ausblick

Die Ergebnisse der vorliegenden Untersuchung decken sich von ihrer Tendenz her mit bisherigen Evaluierungsstudien zur Retrievaleffektivität von Suchmaschinen im World Wide Web und machen deutlich, dass sich diese Ergebnisse auch auf die Betrachtung von Metasuchmaschinen übertragen lassen. Dabei hat in der vorliegenden Studie auch die verstärkte Verwendung von Suchoperatoren keinen wesentlichen Beitrag zur Effektivitätserhöhung bewirkt.[8] Die Annahme, durch Verwendung von Metasuchmaschinen zu besseren Ergebnissen zu gelangen, ließ sich gleichfalls nicht bestätigen. Dies bedeutet, dass die Verfahren der Zusammenführung einer Mehrzahl von Ergebnismengen, wie sie Metasuchmaschinen durchführen, zwar theoretisch den Zugriff auf eine größere Teilmenge der im WWW verfügbaren Daten möglich machen, für überschaubare Ergebnismengen (cut-off 30) aber zu keiner effektiven Verbesserung führen.

Dabei ist allerdings zu beachten, dass die schwerwiegenden methodischen Einschränkungen dieser Untersuchung (paper-and-pencil-Experiment, Bewertung der Rechercheergebnisse nur auf der Basis von Metainformation, geringe Zahl von Anfragen, Testpersonen und untersuchten Systemen) die Testergebnisse nur als erstes Indiz für die Effektivität von Metasuchmaschinen erscheinen lassen. In weiteren Evaluierungen auf einer breiteren empirischen Basis wäre neben den hier vorgestellten Ergebnissen auch zu untersuchen, wie stark sich die Ergebnismengen der Suchmaschinen überlappen und ob Metasuchmaschinen hier ein anderes Verhalten aufweisen. Wünschenswert wären dabei auch Untersuchungen mit Testpersonen ohne entsprechenden fachlichen Hintergrund.

Die insgesamt relativ schlechten Effektivitätswerte bei einer angenommenen hohen "Dunkelziffer" zusätzlicher relevanter Dokumente, die mit weiteren Suchmaschinen nachzuweisen wären, lassen es erforderlich erscheinen, die Arbeitsweise sowie die Benutzerschnittstellen von Suchmaschinen und Metasuchmaschinen zu verbessern. Denkbare Ansätze hierfür sind

- der verstärkte Einsatz von Visualisierungstechniken, z. B. um Zusammenhänge zwischen Suchbegriffen darzustellen,
- die verstärkte Individualisierung der Recherche durch Adaption des Suchwerkzeugs an typische Informationsbedürfnisse des Benutzers und
- der Einsatz von *relevance feedback-Verfahren,* wie sie von einigen Suchmaschinen bereits angeboten werden.

[8] Hier ist darauf hinzuweisen, dass Studien zur Operatorenverwendung i. d. R. auf Auswertung großer *query logs* ohne *empirische Evaluierung* beruhen, d. h. ein Zusammenhang zwischen Operatorenverwendung und Effektivität ist bisher nicht untersucht.

5. Literatur

BAEzA-YATES, Ricardo; RIBEIRO-NETO, Berthier (1999). Modern Information Retrieval. Harlow et al.: The ACM Press/The MIT Press.

BERGMAN, Michael et al. (1999). Search Tutorial: Guide to Effective Searching of the Internet. Vermillion/SD: VisualMetrics Corporation, Juli 1999, *http: //thewebtools. com/tutorial/index. htm.*

CAGLAYAN, Alper K.; HARRISON, Colin G. (1998). Intelligente Software-Agenten. Grundlagen, Technik und praktische Anwendung im Unternehmen. München & Wien: Hanser.

DREILINGER, Daniel E. (1996). "Description and Evaluation of a Meta-Search Agent." M. Sc. Thesis, Colorado State University, Department of Computer Science, Fort Collins/CO, Oktober 1996.

GORDON, Michael; PATHAK, Praveen (1999). „Einding Information on the World Wide Web: the Retrieval Effectiveness of Search Engines." In: Information Processing & Management 35 (1999), 141-180.

HARTER, Stephen P. (1996). „Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness." In: Journal of the American Society for Information Science 47 (1996), 37-49.

JANSEN, Bernard J. et al. (1998). „Real Life Information Retrieval: A Study of User Queries On the Web." In: SIGIR Forum 32(1) (1998), 5-17.

JANSEN, Bernard J.; SPINK, Amanda; SARACEVIC, Tevko (2000). "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web." In: Information Processing & Management 36(2) (2000), 207-227.

KIRSCH, Steve (1998). „Infoseek's Experiences Searching the Internet." In: SIGIR Forum 32(2) (1998), 3-7.

LAVOIE, Brian; FRYSTYK NIELSEN, Henrik (1999). Web Characterization Terminology & Definitions Sheet. World Wide Consortium Working Draft, Mai 1999, *http://www.w3.org/1999/05/WCA-terms/.*

LAWRENCE, Steve; GILEs, C. Lee (1998). „Searching the World Wide Web." In: Science 280 (1998), 98-100.

LAWRENCE, Steve; GILEs, C. Lee (1999). „Accessibility and Distribution of Information on the Web." In: Nature 400 (1999), 107-109 *[http://www.wwwmetrics. com/] .*

LEIGHTON, H. Vernon; SRIVASTAVA, Jaideep (1999). „First 20 Precision among World Wide Web Search Engines." In: Journal of the American Society for Information Science 50(10) (1999), 870-881.

MIELKE, Bettina (2000). Evaluierung juristischer Informationssysteme. Köln et al.: Heymanns [= Ius Informationis Bd. 11].

NOTESS, Greg (19990. Search Engine Showdown. The Users' Guide to Web Searching. Dezember 1999 / September 2000, *http://www.notess.com/search/.*

OCLC (1999). June 1999 Web Statistics. Online Computer Library Center, Inc, Web Characterization Project, *http://www. ocic. org/ocic/research/projects/webstats/statistics. htm.*

PIMIENTA, Daniel et al. (1998). L4. The Fourth Study on Languages and the Internet. The Place of Latin Languages and Cultures on the Internet. September 1998. Santo Domingo: Funredes Networks and Development Foundation, *http: //funredes. org/LC/ engl ish/L4prologue. html.*

PITKOW, James E. (1998). "Summary of WWW Characterizations." In: Proc. Seventh World Wide Web Conference, Brisbane, April 1998, *http://www 7. scu. edu.au/ program m e/fullpapers/18 77/com 18 77. htm*

SALTON, Gerard; MCGILL, Michael J. (1987). Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg et al.: McGraw-Hill.

SCHWARTZ, Candy (1998). „Web Search Engines." In: Journal of the American Society for Information Science 49(11) (1998), 973-982.

SILVERSTEIN, Craig et al. (1999). "Analysis of a Very Large Web Search Engine Query Log." In: SIGIR Forum 33(1) (1999), 6-12.

SULLIVAN, Danny (1999). Search Engine Watch. Dezember 1999 / September 2000, *http://www. searchenginewatch.com.*

TAGUE, Jean (1981). „The Pragmatics of Information Retrieval Experimentation." In: SPARCK JONES, Karen (ed.) (1981). Information Retrieval Experiment. London et al.: Butterworths, 59-102.

WOMsER-HACKER, Christa (1989). Der PADOK-Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Verfahren. Hildesheim: Olms [= Sprache und Computer Bd. 10].