Vergleich der Relevanz von Treffern bei algorithmischen Suchmaschinen, Social Bookmarking-Seiten und Frage-Antwort-Diensten

OLGA GAMMER
olga.gammer@web.de

HEIDI MEISSNER
h_meissner@lycos.de

MAGDALENA PRECKEL
m_preckel@lycos.de

ROBERT OEHLERT
r.oehlert@gmail.com

In diesem Bericht soll die Relevanz der Treffer von algorithmischen Suchmaschinen, Social Bookmarking-Diensten und Frage-Antwort-Diensten untersucht werden. Dazu wird eine Untersuchung mit Probanden durchgeführt, die die Relevanz und einige weitere Kriterien von Treffern bewertet. Die gewonnenen Daten werden statistisch ausgewertet. Sie zeigen, dass alle drei Arten von Suchdiensten eine niedrige Precision aufweisen. Google liefert die relevantesten Treffer, hat jedoch zugleich die meisten kommerziellen Angebote. Social Bookmarking-Dienste liefern besonders viele irrelevante Treffer oder tote Links. Frage-Antwort-Dienste hingegen erzielen eine überraschend gute Relevanz. Es gibt viele Überschneidungen bei den algorithmischen Suchmaschinen und bei den Social Bookmarking-Diensten. Abschließend zeigt die Untersuchung, dass die Suche bei Lycos IQ verbesserungswürdig ist.

1. EINLEITUNG

Um festzustellen, inwiefern die in Lycos iQ gesammelten Informationen zur Verbesserung der Ergebnisrelevanz in die algorithmische Suche eingebunden werden können, wurde beschlossen, zunächst die Relevanz verschiedener Suchdienste untereinander zu vergleichen. Als Grundlage für diese Entscheidung wurde eine empirische Evaluation der Relevanz von Treffern bei verschiedenen Suchdiensten durchgeführt und die daraus resultierenden Daten wurden unter verschiedenen Gesichtspunkten miteinander verglichen. Um die Vielfalt an Suchdiensten zu vergrößern, wurden in die Untersuchung auch Social Bookmarking-Dienste als eine weitere Community-gestützte Alternative zu klassischen Suchmaschinen aufgenommen.

Als problematisch bei dieser Untersuchung stellten sich die seit 40 Jahren in der Informationswissenschaft umstrittene Definition von Relevanz und die Abgrenzung dieser gegen den Begriff Pertinenz [Mizz97] [Stoc07] dar. Da diese Kontroverse

jedoch im Rahmen des Projektes nicht gelöst werden konnte, orientierte sich die Untersuchung an der Definition:

> „[R]elevance may be defined as a criterion reflecting the effectiveness of interactive exchange of information between people (or between people and objects potentially conveying information) in communicative relation, all within a context" [Sara96].

Neben der Precision, welche sich als Quotient aus der Anzahl der gefundenen relevanten Datensätze und der Gesamtzahl der gefundenen Datensätze ergibt [Stoc07], erfasste die Untersuchung verschiedene Aspekte, die helfen können, die Relevanz eines Dokuments zu bestimmen.

Für die Auswertung wurden die Daten der verschiedenen Suchdienste sowohl innerhalb ihrer Kategorie, also nach algorithmischen Suchmaschinen, Frage-Antwort-Diensten oder Social Bookmarking-Diensten, als auch kategorienübergreifend miteinander verglichen, um ein möglichst differenziertes Ergebnis zu erhalten.

Zu Beginn der Untersuchung wurde erwartet, dass gerade die Social Bookmarking-Dienste eine recht hohe Precision erreichen würden, da dort die User selber die Inhalte sammeln. So wie Maaß dachten wir, einen „interessante[n] Ansatz darin zu sehen, die Ergebnisse sozialer Bookmarking[-D]ienste in [...] Suchergebnislisten [von Suchmaschinen] zu integrieren, um die Qualität der Suchergebnisse zu verbessern" [MaGr07]. Frage-Antwort-Dienste dagegen schätzten wir als keine gute Ergänzung zur algorithmischen Suche ein.

2. METHODIK

Wir haben unseren Relevanztest an drei Suchmaschinen, zwei Social Bookmarking-Diensten und zwei Frage-Antwort-Diensten durchgeführt. Bei den Suchmaschinen entschieden wir uns für Google, Yahoo! und MSN, da sie die drei größten Suchmaschinen mit eigenem Index sind. Bei den Social Bookmarking-Diensten wählten wir Mr. Wong aus, da dies der bekannteste Anbieter dieses Services im deutschsprachigen Raum ist, und del.icio.us, da dieser der bekannteste Anbieter für Social Bookmarking weltweit ist. Aufgrund der engen Zusammenarbeit mit Lycos Europe bezüglich unseres Projektes, war es sinnvoll, deren Frage-Antwort-Dienst Lycos iQ zu in die Analyse mit aufzunehmen. Um vergleichbare Ergebnisse zu erzielen, wählten wir Yahoo! Clever als zweiten Frage-Antwort-Dienst aus. Vom Aufbau und der Grundidee unterscheiden sich diese beiden Dienste kaum.

2.1 Vorüberlegung

Bevor die Untersuchung begonnen hat, galt es, einige Vorüberlegungen zu treffen. Um die Relevanz der Treffer definieren und aufzeigen zu können, musste eine Kriterienliste erstellt werden. Danach folgte die sorgfältige Auswahl der Suchanfragen.

2.1.1 Kriterienliste

Zunächst wurde ein Kriterienkatalog mit fünf Fragen erstellt, die die Relevanz der gelieferten Treffer beurteilen sollten. Dabei war die Frage zu klären, was sich hinter Relevanz verbirgt.

Unter dem Begriff Relevanz lässt sich sowohl die Aktualität der Internetseiten, also der Treffer, als auch deren Verständlichkeit verstehen. Wenn eine Seite unverständlich ist, z. B. aufgrund einer fremden Sprache, wurdu sie in der Untersuchung als unverständlich beurteilt. Ein weiteres Kriterium für Relevanz ist der Inhalt der Webseite, d. h. ob ausreichende Informationen enthalten sind, um die Suchanfrage vollständig zu beantworten.

Außerdem ist für den Nutzer die Vertrauenswürdigkeit eines angebotenen Treffers von Interesse, z. B. bei dem Wunsch der Weiterverwendung der Quelle für wissenschaftliche Arbeiten. Relevante weiterführende Links bieten Aufschluss über weitere Informationen. Wie jedoch im Kapitel 2.1.2 beschrieben wird, wurden diese bewusst ausgespart.

Zusammengefasst ergaben sich für jeden Treffer die folgende Bewertungsfragen:

1. Ist der Treffer deiner Meinung nach relevant? Skala 1-6 (Schulnotensystem)
2. Wie schätzt du die Verständlichkeit der Seite ein? Ja/nein/irrelevant
3. Wie schätzt du die Aktualität der Seite ein? Ja/nein/irrelevant
4. Ist diese Seite vertraulich? Impressum/Zitierfähigkeit? Ja/nein/irrelevant
5. Sind weiterführende Links vorhanden? Ja/nein/irrelevant

2.1.2 Suchanfragen

Als weiterer Schritt folgt die generelle Auswahl der Suchanfragen. Broder [Lewa05] unterscheidet zwischen drei Arten von Suchanfragen: navigational (navigationsorientiert), informational (informationsorientiert) und transactional (transaktionsorientiert). Mit navigationsorientierten Anfragen sind Webseiten gemeint, die der Suchende bildlich bereits kennt, z. B. wenn er nach populären Unternehmen oder Personen sucht. Informationsorientierte Anfragen bilden die zweite Gruppe, welche über die Informationsvermittlung anhand nur eines Dokumentes hinaus geht. Nach Broder zielen informationsorientierte Anfragen auf jeden Fall auf statische Dokumente, d. h. es ist nach dem Aufruf des Dokuments keine weitere Interaktion auf der Website nötig, um an die gewünschten Informationen zu gelangen [Lewa06]. Hinzu kommen die transaktionsorientierten Anfragen. Transaktion bedeutet hier z. B. ein Produktkauf, ein Datei-Download oder eine Datenbankrecherche.

In der gesamten Untersuchung lag die Konzentration ausschließlich auf den informationsorientierten Suchanfragen.

2.2 Voruntersuchung (Pre-Test)

Vor dem Haupttest galt es, Vorüberlegungen für einen Pre-Test anzustellen. Die daraus resultierenden Ergebnisse sollten in der Hauptuntersuchung berücksichtigt werden.

2.2.1 Selektion der Fragen

Aus einem Logfile der T-Online-Suche mit über 10.000 beliebigen Fragen wurden fünf Fragen nach dem Zufallsprinzip ausgewählt. Ausgesucht wurden nur informationsorientierte Suchanfragen, da eine Relevanzbewertung von transaktionsorientierten Anfragen bei Frage-Antwort-Diensten zu keinem befriedigenden Ergebnis geführt hätte. Da nur bei informationsorientierten Fragen mehrere Dokumente sinnvoll bewertet werden können, ist hier eine zufriedenstellende Relevanzbewertung möglich.

Transaktionsorientierte Fragen dagegen zielen durch ihre kommerzielle Art auf Benutzer ab, die ggf. den Kauf eines Produktes oder die Inanspruchnahme einer Dienstleistung gegebenenfalls beabsichtigen. Es ist umöglich, mehrere Dokumente sinnvoll zu bewerten, da die abgegebenen Meinungen der Benutzer zu subjektiv sind. Folglich hat der Relevanzbegriff hier eine andere Bedeutung.

2.2.2 Vorgehensweise

Um ein repräsentatives Ergebnis zu erzielen, wurde innerhalb kürzester Zeit jede der fünf Fragen in jedem der sieben Suchdienste gestellt und anonymisiert. Da es sich sowohl um Ein- als auch Mehrwortanfragen handelt, werden die Anfragen vor der Suche in tatsächliche Fragen umgewandelt.

Das Ranking der Treffer sollten die späteren Probanden nicht kennen. Daher erfolgte schon im Pre-Test die Extraktion der Treffer-URLs in ein Excel-Dokument, die sogenannte Anonymisierung der Treffer.

Bei den Frage-Antwort-Diensten wurden auf die vorher gestellte Suchanfrage als „Treffer“ jeweils die ersten drei angezeigten Fragen ausgewählt, eine der dazugehörigen Antworten herauskopiert und der Ursprung unkenntlich gemacht. Dabei wurden nur geschlossene Fragen berücksichtigt. Bei den Antworten wurde die Antwort mit dem „Top-Sternchen“ ausgesucht. Gab es diese nicht, wurde die als „gut beantwortet“ gekenzeichnete Frage genommen. Wenn beide Bewertungen nicht vorhanden waren, blieb nur noch der Griff zum ersten Treffer von oben.

Bewertet werden sollten jeweils die ersten zehn Treffer bei algorithmischen Suchmaschinen und den Social Bookmarking-Diensten. Hinzu kamen drei Fragen pro Frage-Antwort-Dienst inklusive je einer Antwort. Doppelte URLs wurden vermerkt, jedoch ausgeblendet. Die fünf vorher ausgewählten Fragen inklusive anonymisierter Antworten (also Treffer) wurden auf vier Probanden verteilt und nach den Kriterien „Relevante Treffer“, „Verständlichkeit der Seite“, „Aktualität“, „Vertraulichkeit der Seite (Impressum/Zitierfähigkeit)“, „Reichen die Informationen aus?“ und „Sind weiterführende Links vorhanden“ ausgewertet.

2.2.3 Auswertung

Die beim Vortest ausgewerteten Daten waren statistisch nicht aussagekräftig, da sie aus einer viel zu kleinen Stichprobe stammten. Dennoch gaben sie uns einen ersten Eindruck davon, wie unsere Ergebnisse aussehen würden.

Unsere Erwartung, dass Google eine eindeutige Führungsposition einnehmen würde, schien in diesen ersten Ergebnissen bestätigt zu sein. Genauso verhielt es sich

mit unserer Erwartung, dass Frage-Antwort-Dienste eher schlecht abschneiden würden.

2.2.4 Ergebnisse

Erfahrungsgemäß zeichnete sich in den Suchdiensten schnell ab, wenn es keine Treffer zu einer Suchanfrage gab. Die Haupttestvorbereitung begann somit mit der Fragestellung in den Frage-Antwort-Diensten, gefolgt von den Social Bookmarking-Seiten. Gab es zu einer Suchanfrage in einem der ausgewählten Suchdiensten keinen einzigen Treffer, so wurde die Suchanfrage komplett herausgenommen und nicht berücksichtigt. Die Untersuchung war nur dann sinnvoll, wenn in jedem der drei Arten von Suchdiensten Treffer angezeigt wurden, da die Ergebnisse sonst verzerrt gewesen wären.

Wesentliche Erkenntnisse zeichneten sich bei dem Frage-Antwort-Tool Lycos iQ ab. Es erfolgte weder ein Ranking der Antworten noch eine Unterscheidung von Singular- und Pluralbegriffen. Das Fragenranking erschien völlig unklar. Zudem wurde keine Schreibkorrektur oder Phrasensuche angeboten. Vermutlich würde sich ein Ranking der Antworten, z. B. nach „Top-Sternchen“, bei einer zukünftigen Einbettung von Lycos iQ in die Suchmaschinen als tauglich erweisen.

Da Lycos iQ zeitweilig nicht mehr auf der Webseite von T-Online aufgeführt wird und um Komplikationen beim Haupttest zu vermeiden, sollte dieser direkt über die Startseiten der Frage-Antwort-Dienste Lycos iQ und Yahoo! Clever durchgeführt werden. Im Haupttest wurden bei den beiden Frage-Antwort-Diensten die ersten drei Fragen zu der jeweils gestellten Suchanfrage als „Treffer“ ausgewählt, in eine Exceltabelle kopiert und anonymisiert. Da bei den Frage-Antwort-Diensten lediglich die Fragen unsere „Treffer“ waren, blieben die Antworten unberücksichtigt und wurden weggelassen.

Wir hatten uns zu dieser Vorgehensweise aus mehreren Gründen entschlossen. Zum einen werden bei den Frage-Antwort-Diensten nicht die Antworten, sondern die Fragen indexiert. Wenn man eine Suchanfrage stellt, werden die Fragen nach möglichen Übereinstimmungen überprüft und nicht die Antworten, was auch logisch erscheint, denn sonst würde zu viel Informationsballast generiert. Zum anderen werden die Fragen als Trefferliste ausgegeben. Wenn die Frage nicht relevant zur Suchanfrage ist, können die Antworten logischerweise auch nicht relevant sein. Auch präsentiert sich ein Problem bei der Auswahl der Antworten, da es zur Zeit unserer Untersuchung kein Ranking der Antworten gab und mehrere Antworten als „Top Antwort“ gekennzeichnet wurden. Wenn man jedoch alle Antworten aufgenommen hätte, wäre es zu einem Information-Overload gekommen. Des Weiteren wird die Qualität der Antworten bei Lycos iQ von einer anderen Gruppe des Projektes noch einmal ausgiebig beleuchtet (siehe den Beitrag „Untersuchung der Qualität der Antworten bei Lycos iQ und deren Einbindung in die algorithmische Suche“ in diesem Band).

Bei den sieben Suchdiensten waren die unterschiedlichen Ranking-Maßnahmen (Anzeige mit oder ohne Nummerierung) und diverse Zusatzangebote wie beispielsweise kommerzielle Anzeigen, Verbesserungs- oder Tagvorschläge auffällig. Ein zusätzlicher Problempunkt findet sich bei den Social Bookmarking-Seiten wieder. Bei del.icio.us z. B. werden die Links verdeckt angezeigt, d. h. die eigentlichen URLs

sind nicht direkt sichtbar, sondern werden mit einem beliebigen Namen belegt, der im Web für die Nutzer unsichtbar ist und erst bei Anklicken durch das Öffnen der jeweiligen Internetseite im Browserfenster erscheint. Dies führt zu doppelten oder toten Links. Aus Usability-Sicht sind das Anklicken jedes Treffers und das Kopieren der eigentlichen URL in ein neues Browserfenster recht zeitaufwändig und umständlich. Tote Links wurden im Haupttest trotzdem mit aufgeführt und den Probanden gegeben, da die Seiten möglicherweise nur kurzfristig nicht aufrufbar waren.

Ein Problem stellte sich auch bei aufgeführten Unter-URLs, die eigentlich mit Main- oder Index-URLs verwandt sind – den sogenannten „related Links". Diese Links sind ähnlich strukturiert wie diejenigen auf den Hauptseiten. Die „Unter"-Webseiten beinhalten jedoch meist die eigentlichen Informationen, während die Hauptseiten oft nur Einführungs- bzw. Überblickseiten anbieten und eher zur Navigation als zur eigentlichen Informationsvermittlung dienen. Die Entscheidung, diese Links im Haupttest weiterhin aufzuführen, erschien aus objektivitäts- und datenmengentechnischen Gründen sinnvoll. Hätten wir diese Links aus unserer Untersuchung entfernt, hätte sich die Auswahl an möglichen Suchanfragen um ein vielfaches minimiert.

2.3 Haupttest (Main-Test)

Um die Relevanzbewertung möglichst objektiv zu gestalten, musste der Relevanztest von unabhängigen Testpersonen durchgeführt werden, die eine hohe Kompetenz im Bereich der Web- und Suchmaschinennutzung besitzen. Alle Probanden kamen demnach aus dem Bereich der Informationsvermittlung. Anhand von zufällig ausgewählten Suchanfragen aus dem Logfile der T-Online-Suche sollten sie die Relevanz von Treffern bei algorithmischen Suchmaschinen, Social Bookmarking-Seiten und Frage-Antwort-Diensten bewerten. Die dadurch gewonnenen Daten werden ausgewertet und interpretiert.

2.3.1 Suchanfragen

Es sollte eine möglichst große Anzahl an Suchanfragen, jedenfalls mindestens 50, ausgewertet werden, denn je höher die Anzahl der Anfragen, desto geringer ist die Fehlerwahrscheinlichkeit bei den Ergebnissen der Evaluation [Grie00]. Von circa 72 Fragen wurden 54 ausgewählt.

Jede Frage wurde in jedem der sieben Suchdienste einmal gestellt (siehe auch 2.2 Voruntersuchung (Pre-Test), 2.2.2 Vorgehensweise) und die jeweiligen Treffer (die Antworten) herausgefiltert. Pro Frage wurden dabei bei den drei algorithmischen Suchmaschinen und den zwei Social Bookmarking-Diensten jeweils die ersten zehn Treffer, bei den zwei Frage-Antwort-Diensten jeweils die ersten drei Treffer betrachtet.

Die Fragen für den Haupttest, die in einem der Suchdienste keinen Treffer erzielten, wurden nicht an die Probanden weitergegeben, sondern nur vermerkt. Eine Einschränkung auf bestimmte Themenbereiche fand bis auf pornographische Themen nicht statt. Die Herkunft der Treffer wurde unkenntlich gemacht, um eine Verzerrung

der Ergebnisse aufgrund von Vorlieben oder Abneigungen bei den Testpersonen zu bestimmten Suchdiensten zu vermeiden [GRB02].

Die Suchanfragen wurden schriftlich ausformuliert, um sicherzustellen, dass die informationsorientierte Seite der Frage tatsächlich im Vordergrund steht. Besonders die Suchanfragen, die nur aus einem Suchbegriff (Einwort-Suchanfragen) bestanden, liessen mehrere Interpretationen zu.

2.3.2 Vorgehensweise und Bewertung der Probanden

Insgesamt beschäftigten sich 20 Juroren mit jeweils ca 150 Webseiten und neun Fragen aus Frage-Antwort-Diensten, die sich aus Suchanfragen ergeben hatten. Der zeitliche Aufwand pro Suchanfrage wurde auf eine Stunde geschätzt. Jede Suchanfrage und jeder Treffer waren von nur einer Person zu bewerten, um die Eindeutigkeit der Bewertung zu garantieren.

Alle URLs der Treffer der verschiedenen Suchdienste und die Fragen der Frage-Antwort-Dienste wurden mit den entsprechenden Informationen („Von welchem Suchdienst stammt der Treffer?“, „Welchen Rankingplatz hat der Treffer inne?“, „Zu welcher Suchanfrage gehört er?“ und „Handelt es sich hierbei um einen doppelten Treffer?“) in einer Datenbank gespeichert.

Am Testtag wurden die URLs mit einer Identifikationsnummer und den zu bewertenden Kriterien in Excel-Tabellen ausgelesen. Für jede Suchanfrage ergaben sich zwei Tabellen: eine mit den URLs der Treffer und eine mit den kopierten Fragen der Frage-Antwort-Dienste. Dieser Vorgang hatte zum Ziel, den Prozess zu anonymisieren. Um die Treffer noch weiter zu anonymisieren, wurden die URLs alphabetisch geordnet und die Identifikationsnummern ausgeblendet. Die doppelten Treffer und Überschneidungen wurden schon beim Auslesen der Daten aus der Datenbank herausgefiltert, so dass jede Webseite nur einmal bewertet werden musste.

Die Probanden erhielten eine kurze mündliche Einweisung zum Thema Evaluation und ihren Aufgaben als Juroren. Danach bewerteten sie anhand der Bewertungsliste die anonymisierten Treffer. Während der Relevanzbeurteilung sollten die Probanden nicht beeinflusst werden. Die Leiter der Untersuchung standen jedoch für Rückfragen zu technischen Problemen oder für Verständnisfragen zur Verfügung [Grie00].

3. AUSWERTUNG DES HAUPTTESTS

In der Regel wird die Relevanz von Treffern mit dem klassischen Retrievalmaß Precision bewertet. Diese gibt den Anteil der gefundenen relevanten Treffer an der Gesamtzahl der gefundenen Treffer an [StSt00].

Dieses Maß ist verhältnismäßig leicht zu bestimmen, indem alle gefundenen Treffer von den Juroren bewertet werden und anschließend eine Auszählung der relevanten bzw. irrelevanten Treffer erfolgt. Ein Problem taucht aber auf, wenn man anstatt eines zweiwertigen Relevanzurteils (relevant/nicht relevant) eine Skala verwendet, die eine „zusätzliche Qualitätsabstufung möglich macht“ [LeHö07]. Aus diesem Grund wurden in der Untersuchung drei Abstufungen der Precision in „stark“, „mittel“ und „schwach“ unterschieden.

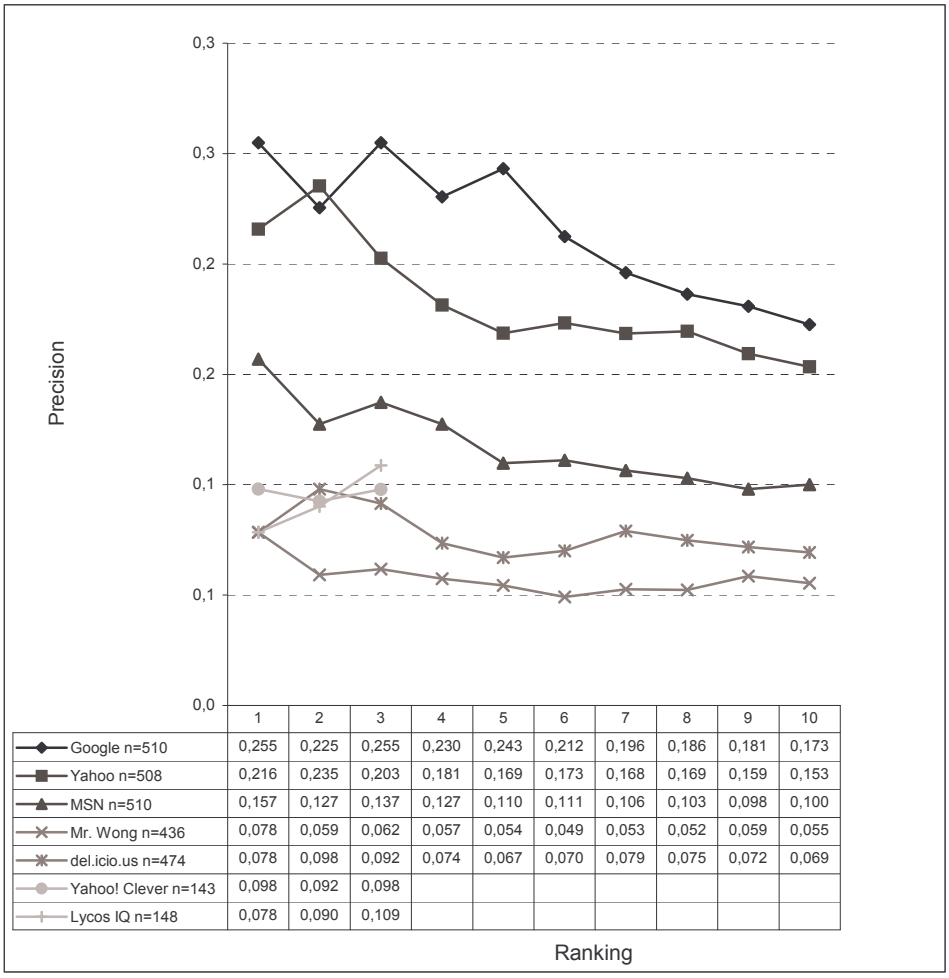
Su, Chen und Dong verwenden bei ihrer Evaluation von Suchmaschinen aus Anwendersicht ebenfalls eine Dreierskala (relevant, teilweise relevant, nicht relevant).

Anschließend rechnen sie sowohl die Precision der relevanten und teilweise relevanten Dokumente zusammen („schwache Precision“) als auch nur der relevanten Treffer („starke Precision“) [StSt00].

3.1 Top Ten Precision

In der Untersuchung der Relevanz von Treffern bei algorithmischen Suchmaschinen, Social Bookmarking-Seiten und Frage-Antwort-Diensten wurden circa 50 Suchanfragen verwendet, für die jeweils die ersten zehn Treffer jedes Suchdienstes (bei Frage-Antwort-Diensten jeweils die ersten drei Treffer) ausgewertet wurden. Die 20 Juroren bewerteten Treffer auf einer Skala von eins bis sechs (angelehnt an das Schulnotensystem: eins = Treffer ist relevant; sechs = Treffer ist nicht relevant).

Für die Auswertung wurde die „starke“, „mittlere“ und „schwache“ Precision berechnet. Allen Treffer, die mit eins bewertet wurden, wurde eine „starke“ Precision zugeschrieben. „Mittlere Precision“ haben diejenigen Treffer, die mit eins und zwei bewertet wurden. „Schwache“ Precision“ bezeichnet relevante, teilweise relevante und wenig relevante Treffer.



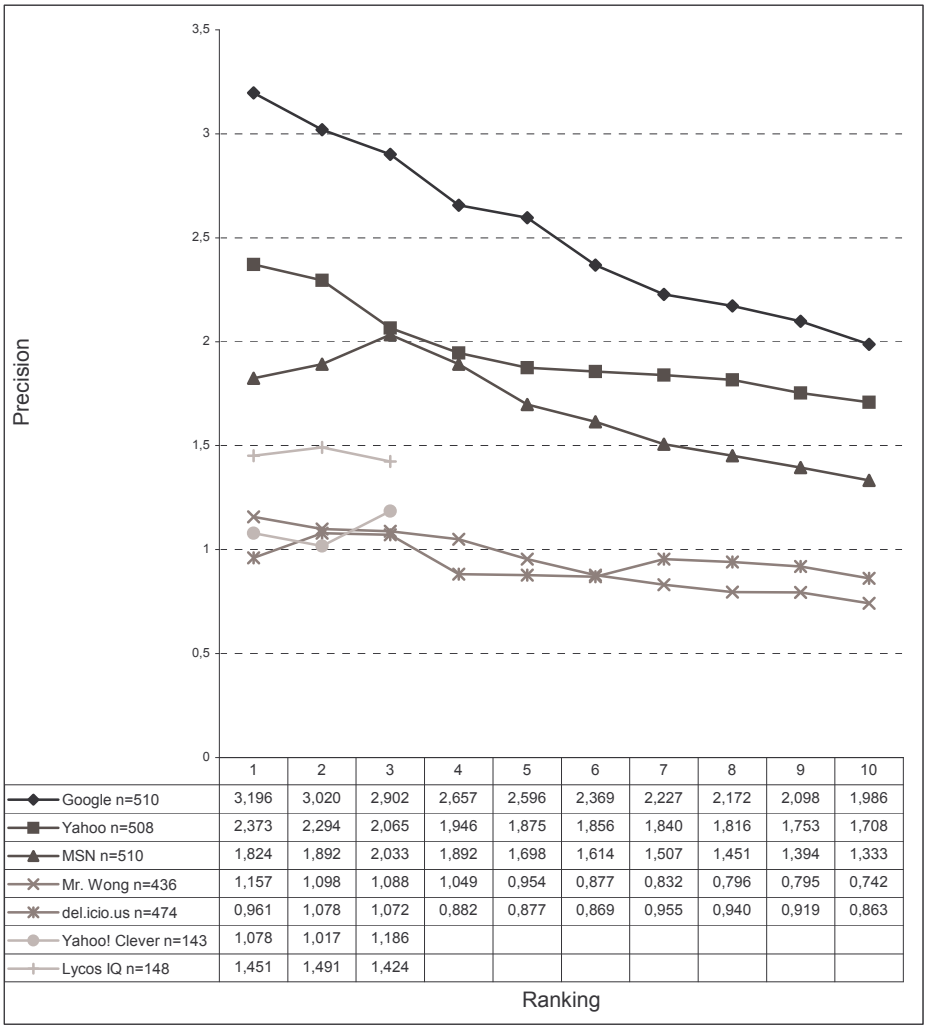
	1	2	3	4	5	6	7	8	9	10
Google n=510	0,255	0,225	0,255	0,230	0,243	0,212	0,196	0,186	0,181	0,173
Yahoo n=508	0,216	0,235	0,203	0,181	0,169	0,173	0,168	0,169	0,159	0,153
MSN n=510	0,157	0,127	0,137	0,127	0,110	0,111	0,106	0,103	0,098	0,100
Mr. Wong n=436	0,078	0,059	0,062	0,057	0,054	0,049	0,053	0,052	0,059	0,055
del.icio.us n=474	0,078	0,098	0,092	0,074	0,067	0,070	0,079	0,075	0,072	0,069
Yahoo! Clever n=143	0,098	0,092	0,098							
Lycos IQ n=148	0,078	0,090	0,109							

Abb. 3.1 Top Ten Precision (“starke” Precision)

Die Betrachtung der Top Ten Precision („starke“ Precision, Abb. 3.1) zeigt, dass Google fast bei allen Rangplätzen höhere Werte erreichte als die anderen Suchdienste. Eine Ausnahme bildet Rang zwei. Die Kurve überschneidet sich mit der von Yahoo!.

Vergleicht man die Kurven von del.icio.us und Mr. Wong, wird ersichtlich, dass del.icio.us besser abschnitt. Die Kurve von Lycos iQ steigt während der ersten drei Positionen an, während die Kurve von Yahoo!Clever gleich bleibt.

Im Falle der „starken“ Precision sind die Ergebnisse der einzelnen Suchdienste schwer zu vergleichen, weil nur wenige Treffer für die Auswertung vorhanden sind. Es waren lediglich elf Prozent Treffer, die ein Top-Ergebnis lieferten bzw. mit eins bewertet wurden. Besser lassen sich die Ergebnisse der Grafik „mittlere“ Precision auswerten. Insgesamt sind es 25 Prozent der Treffer, die mit eins und zwei bewertet wurden.



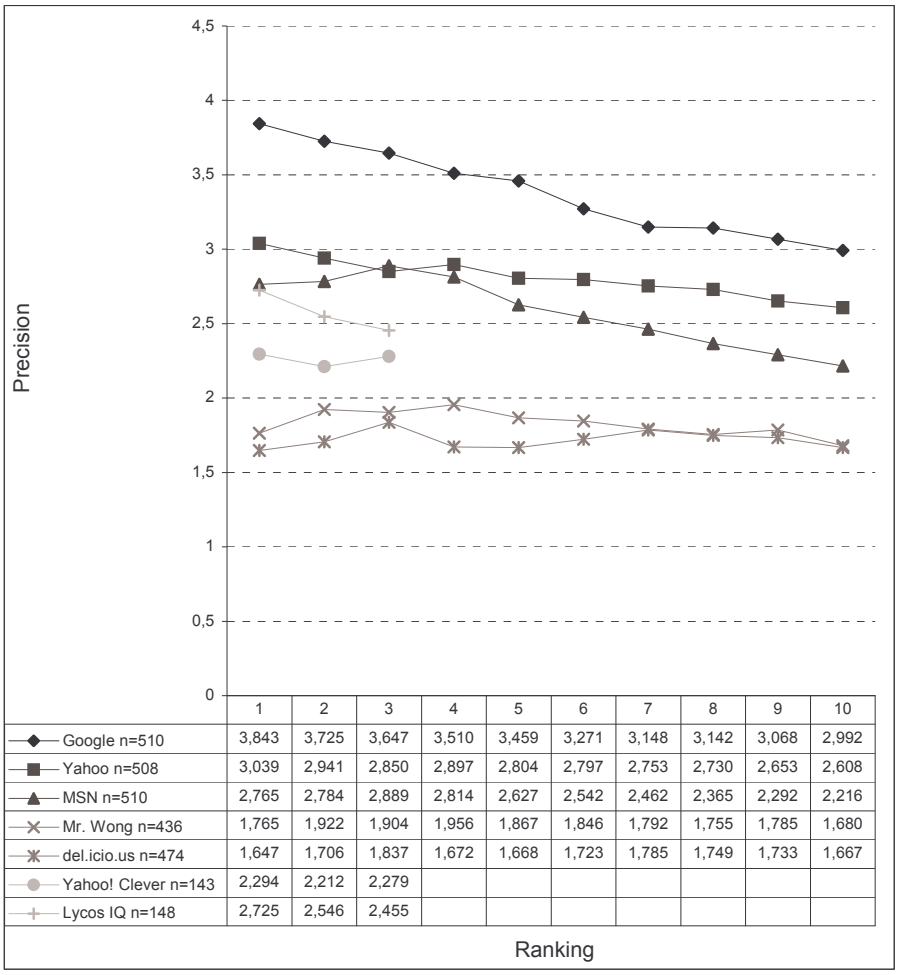
	1	2	3	4	5	6	7	8	9	10
Google n=510	3,196	3,020	2,902	2,657	2,596	2,369	2,227	2,172	2,098	1,986
Yahoo n=508	2,373	2,294	2,065	1,946	1,875	1,856	1,840	1,816	1,753	1,708
MSN n=510	1,824	1,892	2,033	1,892	1,698	1,614	1,507	1,451	1,394	1,333
Mr. Wong n=436	1,157	1,098	1,088	1,049	0,954	0,877	0,832	0,796	0,795	0,742
del.icio.us n=474	0,961	1,078	1,072	0,882	0,877	0,869	0,955	0,940	0,919	0,863
Yahoo! Clever n=143	1,078	1,017	1,186							
Lycos IQ n=148	1,451	1,491	1,424							

Abb. 3.2 Top Ten Precision (“mittlere" Precision)

Für die Auswertung der „mittleren“ und „schwachen“ Precision wurde das Gewicht der einzelnen Treffer berücksichtigt. Die Treffer, die mit eins bewertet wurden, erhielten sechs Punkte, da sie die besten Treffer waren. Die Treffer, die mit zwei bewertet wurden, erhielten fünf Punkte usw.

Der Verlauf der Kurven bei der Messung der mittleren Precision (Abb. 3.2) ist übersichtlicher und eindeutiger dargestellt (vgl. Abb. 3.2). Obwohl die Graphen sich bisweilen überschneiden, zeigt sich klar, welche Suchdienste die relevantesten Treffer lieferten. So erreichte Google bei jedem Rang höhere Precisionwerte als alle anderen Suchdienste. Die Graphen der Social Bookmarking-Diensten überschneiden sich mehrmals, so dass nicht eindeutig ersichtlich wird, welches der beiden Systeme eine

höhere Effektivität erreichte. Bei dem Vergleich der Frage-Antwort-Dienste ließ sich festhalten, dass Lycos IQ besser abschnitt als Yahoo! Clever. Auffällig ist hier der steigende Kurvenverlauf von Yahoo! Clever.



	1	2	3	4	5	6	7	8	9	10
Google n=510	3,843	3,725	3,647	3,510	3,459	3,271	3,148	3,142	3,068	2,992
Yahoo n=508	3,039	2,941	2,850	2,897	2,804	2,797	2,753	2,730	2,653	2,608
MSN n=510	2,765	2,784	2,889	2,814	2,627	2,542	2,462	2,365	2,292	2,216
Mr. Wong n=436	1,765	1,922	1,904	1,956	1,867	1,846	1,792	1,755	1,785	1,680
del.icio.us n=474	1,647	1,706	1,837	1,672	1,668	1,723	1,785	1,749	1,733	1,667
Yahoo! Clever n=143	2,294	2,212	2,279							
Lycos IQ n=148	2,725	2,546	2,455							

Abb. 3.3 Top Ten Precision (“schwache” Precision)

Mr. Wong erzielte bei der Auswertung der „schwachen“ Precision leicht höhere Werte als del.icio.us, allerdings waren die Unterschiede relativ gering. Die Kurven von Yahoo! und MSN überschneiden sich bei der Position drei, sonst erreichte Yahoo! bessere Precisionwerte als MSN. Bei dem Vergleich der Frage-Antwort-Dienste schneidet Lycos iQ eindeutig besser ab als Yahoo! Clever.

Die Auswertung der „schwachen“ Precision zeigt eine eindeutige Reihenfolge im Hinblick auf die Relevanz der Treffer. Die algorithmischen Suchmaschinen erwiesen sich als die effektivsten Retrievalsysteme, gefolgt von den Frage-Antwort-Diensten. An letzter Stelle stehen die Social Bookmarking-Dienste.

Das Gesamtergebnis über die Precision der ersten zehn Treffer ist in der nachfolgenden Tabelle zu sehen.

Suchdienst:	„starke“ Precision	„mittlere“ Precision	„schwache“ Precision
Google n=510	0,22	0,46	0,75
Yahoo! n=508	0,18	0,36	0,62
MSN n=510	0,12	0,30	0,57
Mr. Wong n=436	0,058	0,176	0,40
del.icio.us n=474	0,077	0,173	0,38
Yahoo! Clever n=143	0,096	0,20	0,50
Lycos iQ n=148	0,092	0,26	0,57

Abb. 3.4 Precision der ersten zehn Treffer

Die Ergebnisse zeigen, dass zwischen den Suchdiensten klare Unterschiede vorliegen. Ungeachtet des Precisionsgrades („stark“, “mittel“ oder „schwach“) belegte Google immer den ersten Platz. Mit circa 46 Prozent der relevanten Treffer bei „mittlerer“ Precision z.B. erzeugte Google ca. 15 Prozent mehr relevante Treffer als MSN.

Beim Vergleich zwischen den algorithmischen Suchmaschinen mit den Social Bookmarking-Diensten und den Frage-Antwort-Diensten lagen die Frage-Antwort-Dienste überraschenderweise an zweiter Stelle, wobei Lycos iQ um sechs Prozent besser abschnitt als Yahoo! Clever. Die letzte Position belegten die Social Bookmarking-Dienste, die im Vergleich zu Google weniger als die Hälfte an relevanten Treffern aufzeigen (ca. 17 Prozent).

Die Gründe, warum die Social-Bookmarking-Dienste so wenige relevante Treffer lieferten, sind:

- viele „tote“ URLs, die automatisch mit der Schulnote sechs bewertet wurden. Dies zeigt, dass nichtfunktionierende URLs aus dem Index der Social Bookmarking-Dienste nicht gelöscht werden. Die Ergebnisse zeigen, dass etwa fünf Prozent der ausgegebenen Treffer „tote“ URLs waren. Deswegen ist es erforderlich, „dass Anwender entsprechende Links in ihre Bookmark-Liste einpflegen“ [MaGr07].
- viele fremdsprachige Seiten (ausgeschlossen sind englischsprachige Seiten), die natürlich zwar relevant sein können, aber aufgrund des Nichtverstehens der Sprache von den Juroren meistens subjektiv als nicht relevant bewertet wurden. Von den 17 fremdsprachigen Seiten, die ausgewertet wurden, stammten 16 aus del.icio.us.

In der folgenden Auswertung wird auf nicht-relevante Treffer und die Gründe für ihre Irrelevanz, sowie auf die Verständlichkeit der Seiten eingegangen. Zur Berechnung des Gesamtergebnisses wurde die Grafik „mittlere“ Precision (Abb. 3.5) herangezogen.

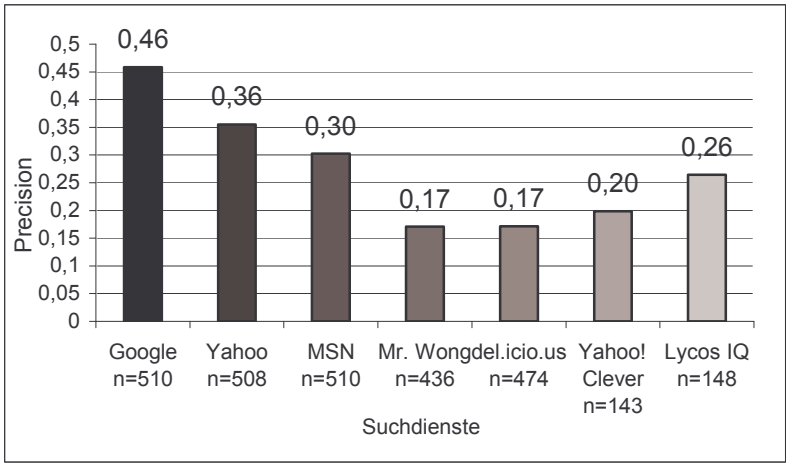


Abb. 3.5 Gesamtergebnis: „mittlere“ Precision

Bei den Ergebnissen der Precision fiel auf, dass der Anteil der relevanten Treffer niedrig war. Selbst der „Testsieger“ Google erreichte nur eine Precision von 0,46. Dies bedeutet, dass nur 46 Prozent der Treffer wirklich relevant sind.

Im Hinblick auf die „schwache“ Precision, bei der alle, sogar wenig relevante Treffer berücksichtigt wurden, betrug die Precision von Google 0,75. Es blieben noch 25 Prozent der in der Top Ten ausgegebenen Treffer, die nicht relevant bzw. schlecht waren.

Bestätigt wird dieser Befund durch eine Untersuchung von Griesbaum, der einen Retrievaltest an „deutschen“ Suchmaschinen (AltaVista.de, Fireball.de, Google.de und Lycos.de) durchführte. Die Ergebnisse zeigten,

> „dass etwa 45 Prozent der in den Top 20 ausgegebenen Treffer nicht relevant sind und auch auf kein relevantes Dokument verweisen. Werden diese Werte nicht in Relation zu den anderen, schlechter abschneidenden Suchmaschinen betrachtet, sondern nur in Hinblick auf den Anteil der relevanten Treffer, so ist das Ergebnis insgesamt als schlecht zu bezeichnen“ [Lewa05].

3.2 Weitere Bewertungskriterien

Neben Relevanz wurden im Haupttest die Bewertungskriterien für die Qualität der Treffer aus dem Vortest übernommen: Aktualität, Verständlichkeit, ausreichende Information und Vertrauenswürdigkeit. Im Folgenden wird auf jedes von ihnen einzeln eingegangen.

3.2.1 Verständlichkeit

Bezogen auf die gesamte Treffermenge waren 84 Prozent der Treffer für die Benutzer verständlich (vgl. Abb. 3.6).

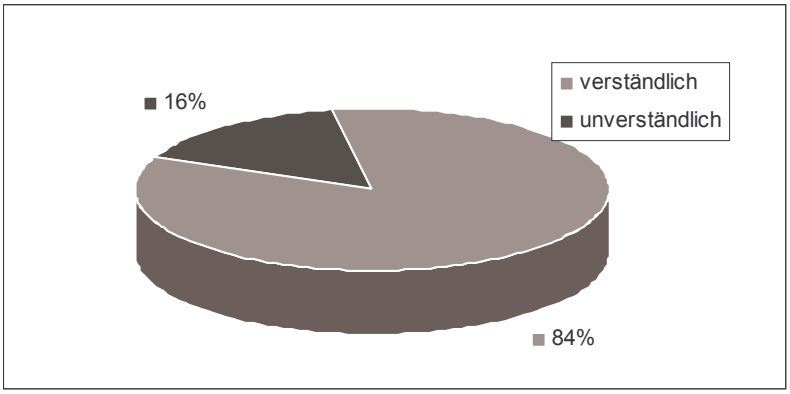


Abb. 3.6 Verständlichkeit der Treffer (n(gesamt)= 2729)

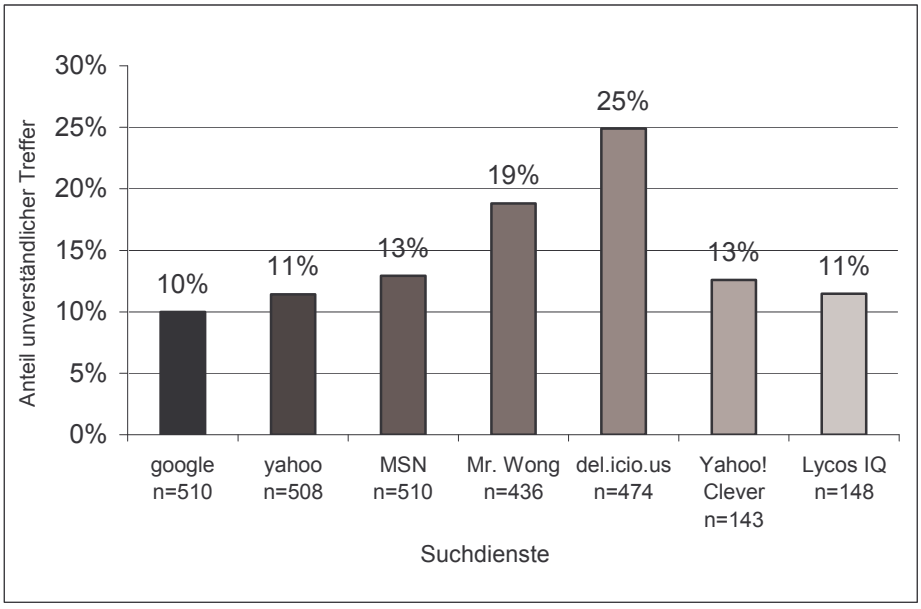


Abb. 3.7 Vergleich der Suchdienste nach Anteil der verständlichen Treffer

Beim Vergleich der Suchdienste nach dem Anteil der Treffer, die für Juroren unverständlich waren (Abb. 3.7), wurde erneut deutlich, dass Social Bookmarking-Dienste verstärkt diese Art Treffer aufweisen. Dabei fiel auf, dass bei del.icio.us 25 Prozent aller Treffer von den Juroren als unverständlich bewertet wurden. Die anderen Suchdienste dagegen hatten einen relativ geringen Anteil an unverständlichen Treffern.

3.2.2 Vertrauenswürdigkeit

Bei der Auswertung der Vertrauenswürdigkeit der Seiten wurden Frage-Antwort-Dienste nicht berücksichtigt, da sie nicht nach diesem Kriterium abgefragt wurden. Bezogen auf die gesamte Treffermenge wurden 49 Prozent der Treffer von den Probanden als vertrauenswürdig und 46 Prozent als nicht vertrauenswürdig eingestuft. Fünf Prozent der Treffer waren, wie bei der Auswertung der anderen Kriterien, irrelevant oder wurden von den Probanden nicht bewertet.

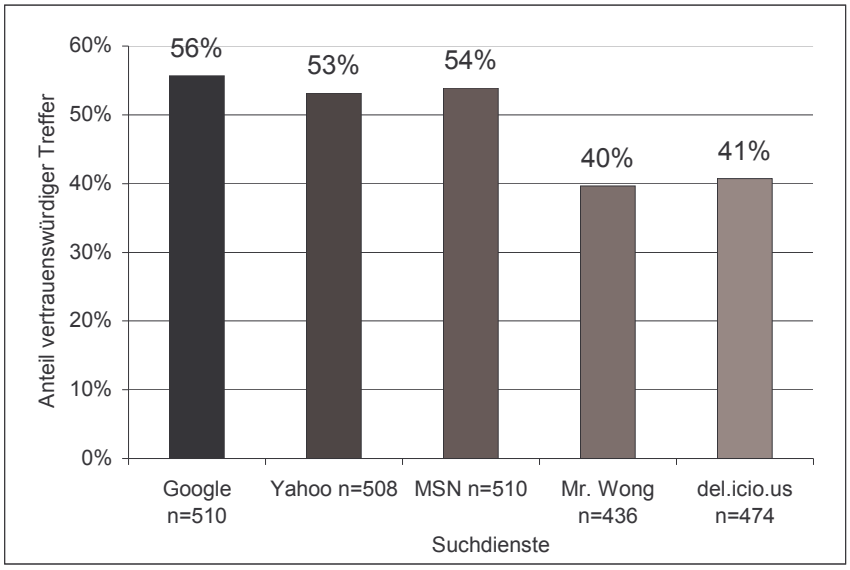


Abb. 3.8 Vergleich der Suchdienste mit Anteil der Treffer, die für Juroren vertrauenswürdig sind (n=2438)

Bei dem Vergleich des Trefferanteils der Suchdienste, die für Juroren vertrauenswürdig sind, führten die algorithmische Suchmaschinen. Die Social Bookmarking-Dienste erreichten nur ca. 40 Prozent der Treffer, die von den Nutzern als vertrauenswürdig beurteilt werden. Dies bedeutet, dass Social Bookmarking-Dienste nicht die von uns erwartete Qualität von Webseiten aufweisen. Wie in der einschlägigen Literatur diskutiert, tendierten auch wir zu der Annahme, dass die von Nutzern ausgewählten Webseiten eine höhere Qualität besitzen, als diejenigen, die von Suchmaschinen gefunden wurden [MaGr07].

3.2.3 Wikipedia-Seiten

Als Nebenprodukt des Relevanztests erschien es interessant, Wikipedia-Seiten zu untersuchen. Das Klischee, Wikipedia-Seiten seien nicht vertrauenswürdig, weil darin beliebig viele Personen Inhalte publizieren können, wurde von den Nutzern anders beurteilt. Obwohl bei Wikipedia-Seiten oft unklar ist, woher die Informationen stammen, wurden sie dennoch von den Nutzern als vertrauenswürdig angesehen.

Insgesamt waren fünf Prozent aller Treffer Wikipedia-Seiten, von denen 63 Prozent als vertrauenswürdig erachtet wurden.

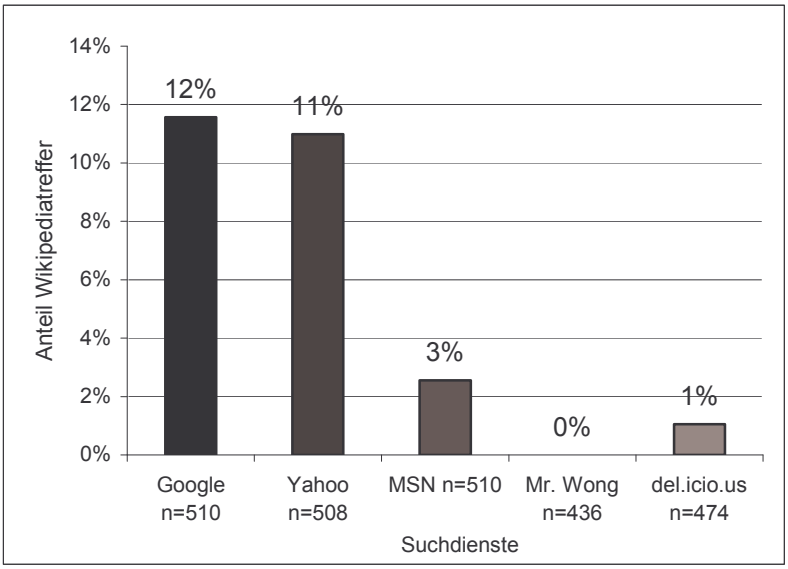


Abb. 3.9 Suchdienste (nur: algorithmische Suchmaschinen und Social-Bookmarking-Dienste (n=2438)) mit den meisten Wikipedia Treffern

Die meisten Wikipedia-Seiten tauchten bei den algorithmischen Suchmaschinen auf. Bei Social Bookmarking-Diensten hingegen wurde sehr wenig auf Wikipedia-Seiten verwiesen.

3.2.4 Ausreichende Information

Das Kriterium „Ausreichende Information“ ist für Frage-Antwort-Dienste nicht anwendbar, weil nur die Fragen dieser Suchdienste untersucht wurden. Das Gesamtergebnis zeigte, dass nur 24 Prozent aller ausgegebenen Treffer ausreichend Informationen boten. Bezogen auf die einzelnen Suchdienste ergab sich das folgende Bild.

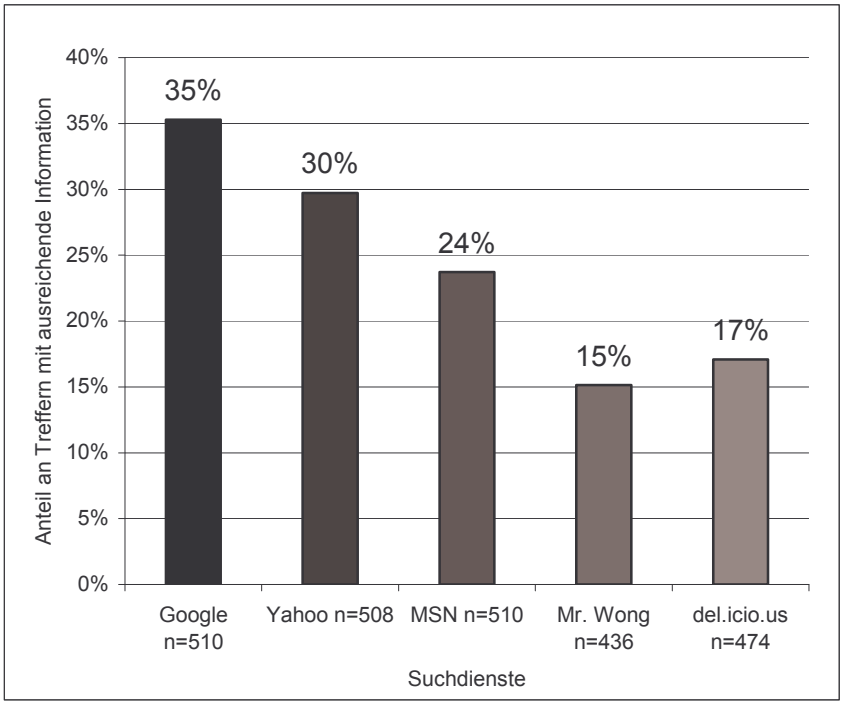


Abb. 3.10 Anteil an Treffern „ausreichende Information“ verteilt nach Suchdiensten

Google erzielte auch hier die höchsten Werte, Yahoo! lag an zweiter Stelle, gefolgt von MSN. Bei den Social Bookmarking-Diensten beinhalteten nur ca. 15 Prozent der ausgegebenen Treffer ausreichende Informationen. Das bedeutet, dass bei den Suchmaschinen häufig bereits beim ersten Treffer alle gesuchten Informationen gefunden werden. Wohingegen der Suchende mit großer Wahrscheinlichkeit bei den Social Bookmarking-Diensten mehrere Treffer ansehen muss, um die gewünschten Informationen zu erhalten.

3.3 Überschneidungen

Unter Überschneidungen versteht man Links aus der Trefferliste eines Suchdienstes, die ein anderer Suchdienst ebenfalls in seiner Trefferliste auflistet. Dabei handelt es sich nicht lediglich um die gleichen URLs, sondern auch um sich minimal unterscheidende Links, die zur gleichen Seite führen. Links, die mehrmals in derselben Trefferliste auftauchen, bezeichnet man als doppelte Treffer.

Da Frage-Antwort-Dienste keine Trefferliste mit Links vorzuweisen haben, sondern hier eine natürlich-sprachige Frage den Treffer ausmacht, tauchen Frage-Antwort-Dienste in diesem Teil der Untersuchung nicht auf.

Von den 2438 Treffern unserer Stichprobe, ohne Frage-Antwort-Dienste, waren 1020 Überschneidungen oder doppelte Treffer. Die meisten Überschneidungen lieferte Google, während die Trefferliste von Mr. Wong die wenigsten Übereinstimmungen mit den Trefferlisten der anderen Suchdienste aufwies. Hierbei

war es interessant zu untersuchen, welche Suchdienste die gleichen Treffer liefern. So stimmten zum Beispiel 34 Prozent der Treffer von Yahoo! und Google überein.

Bei den individuellen Prozentzahlen gab es geringfügige Abweichungen, da die Gesamtmenge der Treffer bei jedem untersuchten Suchdienst unterschiedlich groß war (Abb. 3.11). Diese Unregelmäßigkeiten ergaben sich aus der unterschiedlichen Anzahl von Treffern, die der jeweilige Suchdienst zu einigen der Suchanfragen aufgelistet hatte. Während die Suchmaschinen so gut wie immer mindestens zehn Treffer fanden, kam es bei den Social-Bookmarking-Diensten des Öfteren vor, dass nur fünf bis acht Treffer aufgelistet wurden.

Überschneidungen	Google	Yahoo!	MSN	del.icio.us	Mr. Wong
Google n=510	*0,98%*	34,06%	28,43%	13,08%	12,84%
Yahoo! n=508	33,92%	*1,57%*	26,67%	10,34%	10,09%
MSN n=510	28,43%	26,77%	*1,96%*	7,81%	8,94%
del.icio.us n=474	12,16%	9,65%	7,25%	*4,22%*	15,14%
Mr. Wong n=436	10,98%	8,66%	7,65%	13,92%	*2,06%*

Abb 3.11 Anteil der Übereinstimmung der verschiedenen Suchdienste

Wenn die Treffer zusammengerechnet werden, die sowohl von Google als auch von einem oder mehreren anderen Suchdiensten gefunden wurden, ergeben sich 50 Prozent. Das heißt, dass 50 Prozent der Treffer, die andere Suchdienste lieferten, ebenfalls bei Google gefunden werden. Die größte Übereinstimmung findet sich hierbei zwischen Yahoo! und MSN. Yahoo! findet 48 Prozent der Treffer, die auch die anderen Suchdienste aufgelistet haben und bei MSN gibt es eine Übereinstimmung von 40 Prozent. Del.icio.us hingegen teilt sich die meisten Überschneidungen mit Mr. Wong und umgekehrt.

Eine Besonderheit sind die doppelten Treffer, in Abb. 3.11 kursiv dargestellt (siehe auch Abb. 3.14), die sich nicht bei anderen Suchdiensten finden, sondern in der gleichen Trefferliste. Hierbei liegt del.icio.us mit rund vier Prozent eindeutig vorn. Das hat zu der Vermutung geführt, dass hier keine doppelten URLs erkannt und automatisch zusammengeführt werden, wenn ein Nutzer sie hinzufügt. Da hier in der Trefferliste selbst die URLs nicht sichtbar sind, könnte dies bei der Suche schnell zu Frustration bei den Nutzern führen.

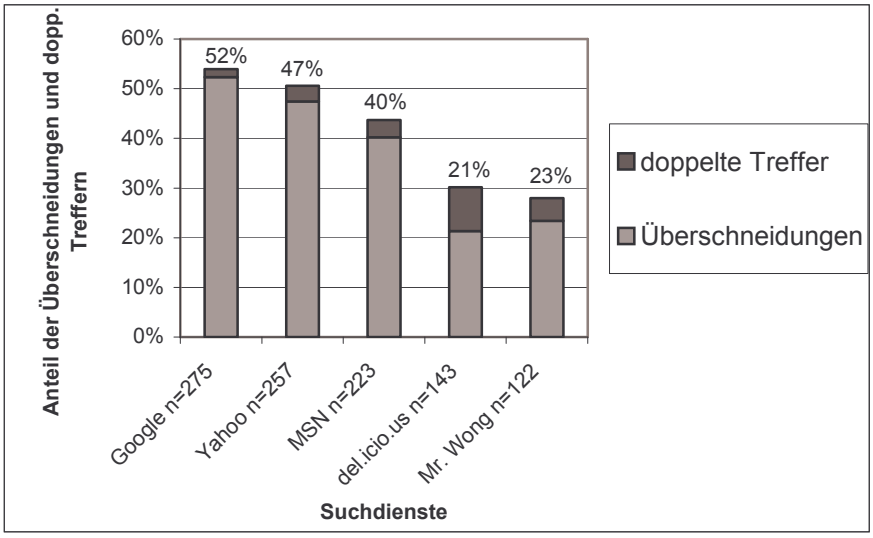


Abb. 3.12 Verteilung der Überschneidungen und doppelten Treffer (n=1020) auf die Suchdienste

Google und Yahoo! wiesen besonders bei Treffern mit „mittlerer" Precision, aber auch bei den irrelevanten Treffern, Überschneidungen mit anderen Suchdiensten auf.,. Die Ähnlichkeit der Verteilung dieser beiden Suchdienste ergab sich aus der großen Übereinstimmung ihrer Treffer. Auch MSN zeigte eher Überschneidungen bei Treffern mit „mittlerer" Precision. Im Gegensatz zu seinen Konkurrenten fanden sich hier jedoch vermehrt Überschneidungen bei weniger relevanten und irrelevanten Treffern.

Die Social Bookmarking-Dienste zeigten ebenfalls Überschneidungen bei relevanten Treffern, doch eine Häufung war eindeutig bei den irrelevanten Treffern zu sehen.

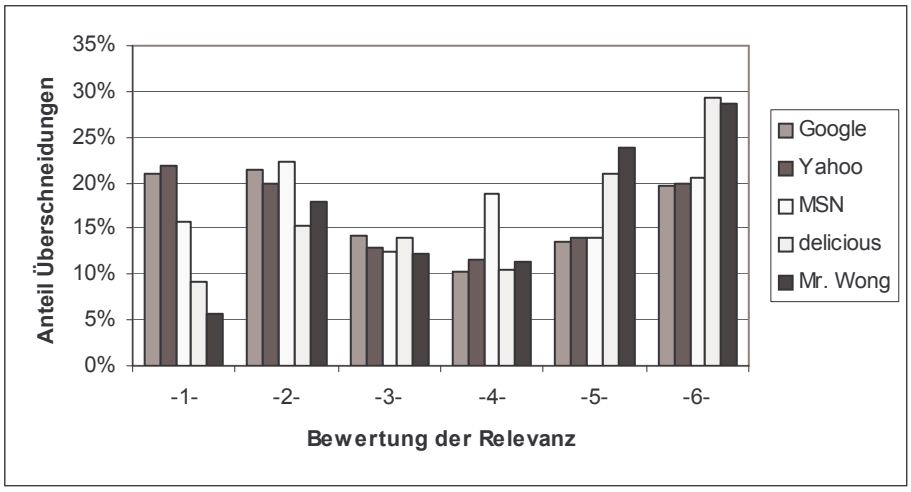


Abb. 3.13 Relevanz der Überschneidungen und doppelten Treffer (n=1020) bei den einzelnen Suchdiensten

Bei allen Suchdiensten fanden sich die meisten Überschneidungen und doppelten Treffer auf den Rängen eins bis fünf, wobei sie sich besonders auf den beiden ersten Plätzen häuften. Bei Mr. Wong zeigte sich eine besondere Konzentrierung auf die ersten beiden Ränge.

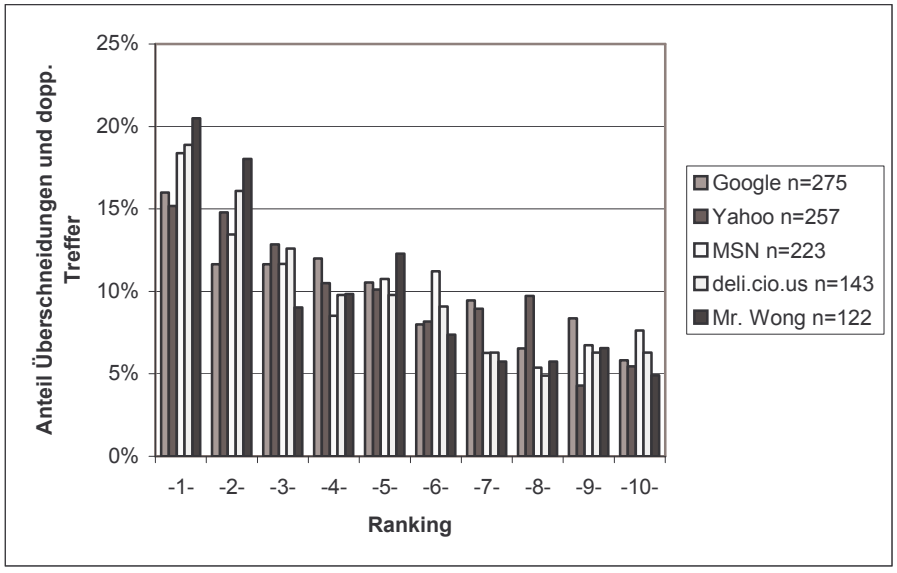


Abb. 3.14 Ranking der Überschneidungen und der doppelten Treffer (n=1020)

Die Verteilung der Überschneidungen und doppelten Treffer auf die Suchanfragen zeigte, dass hier nochmals eine Unterscheidung zwischen den sich überschneidenden Treffern stattfindet. Zum einen ging es um die identischen URLs, die die Suchdienste

auf die gestellten Suchanfragen ausgaben, und zum anderen um die Seiten, auf die diese URLs verwiesen. Bei „Eishockey“ und „Zeitumstellung“ wiesen die meisten doppelten Treffer, im Durchschnitt 3,2, auf die gleiche Seite. Hier waren also bei einer Suche in unterschiedlichen Suchdiensten immer wieder die gleichen Treffer erhältlich. Bei „dd“ hingegen lieferten die meisten Suchdienste unterschiedliche Treffer, was wahrscheinlich an der hohen Mehrdeutigkeit der Suchanfrage lag (siehe auch Anhang).

3.4 Irrelevante Treffer

Abb. 4.15 gibt die “negative” Precision an, also die Precision auf die Relevanzbewertung fünf und sechs. Dabei sticht hervor, dass die Social Bookmarking-Dienste am schlechtesten abschnitten. Auch die Frage-Antwort-Dienste wiesen keine befriedigenden Ergebnisse auf. Eine Ausnahme bildete allerdings der erste Treffer bei Lycos iQ, dessen Wert besser war als der der zweiten Treffer von Yahoo! und MSN. Der für Suchanfragen auf Deutsch am wenigsten geeignete Suchdienst war del.icio.us, dessen Precision-Wert auf die Bewertung fünf und sechs 0,5 nie unterschritt.

Bei den algorithmischen Suchmaschinen zeichnete sich erwartungsgemäß unter den ersten Treffern ein recht guter Wert ab. Diesen guten Wert hielt jedoch nur Google bis zum fünften Treffer. Die anderen Dienste lieferten zum Teil bereits ab dem zweiten Treffer negative Ergebnisse (vgl. Abb. 3.15). Dieses Ergebnis bestätigt die Aussagen der „normalen Precision“. Des Weiteren ist in der Tabelle die Verteilung der nicht relevanten Treffer auf den Plätzen der Rankinglisten angegeben. So wies Google auf Platz eins – über alle Anfragen hinweg – nur zwölf nicht relevante Treffer auf, Mr. Wong hingegen 32.

	Google		Yahoo!		MSN		Mr. Wong		del.icio.us		Yahoo! Clever		Lycos IQ	
Ranking	Trefferanzahl	Precision?	Trefferanzahl	Precision?	Trefferanzahl	Precision?	Trefferanzahl	Precision?	Trefferanzahl	Precision?	Trefferanzahl	Precision?	Trefferanzahl	Precision?
1	12,00	0,24	19,00	0,37	20,00	0,39	32,00	0,63	30,00	0,61	24,00	0,47	19,00	0,37
2	13,00	0,25	22,00	0,43	20,00	0,39	26,00	0,52	32,00	0,63	22,00	0,48	23,00	0,47
3	14,00	0,29	21,00	0,41	18,00	0,35	26,00	0,58	28,00	0,51	21,00	0,46	23,00	0,48
4	17,00	0,31	15,00	0,29	21,00	0,41	21,00	0,47	35,00	0,69				
5	16,00	0,31	23,00	0,45	28,00	0,55	30,00	0,64	31,00	0,61				
6	23,00	0,45	21,00	0,41	26,00	0,51	23,00	0,53	24,00	0,51				
7	24,00	0,47	24,00	0,48	26,00	0,51	27,00	0,66	22,00	0,53				
8	15,00	0,29	23,00	0,45	32,00	0,63	25,00	0,63	29,00	0,66				
9	23,00	0,45	27,00	0,53	30,00	0,59	19,00	0,51	26,00	0,60				
10	23,00	0,45	25,00	0,50	26,00	0,51	29,00	0,78	28,00	0,67				

Abb 3.15 „negative“ Precision

Begründung	Häufigkeit
Suchanfrage	6
Fremdsprache	14
Anderes	18
Subjektive Frage	21
Error	67
Zu wenig Information	73
Kommerziell	564
Inhalt passt nicht	563
Gesamtmenge "irrelevante Treffer"	1332

Tab. 3.16 Begründung für die Bewertung eines Treffers als „irrelevant“

Um einen begründeten Vergleich anzustellen, warum ein Treffer als irrelevant bewertet wurde, war es zunächst nötig, die freien Begründungen in verschiedene Kategorien zu unterteilen (vgl. Abb. 3.16), von denen vier einer näheren Erläuterung bedürfen.

Die Kategorie „kommerziell“ umfasst alle Treffer, die zwar zur Suchanfrage passen, aber rein kommerzielle Angebote sind. Dementsprechend finden sich in der Kategorie „Inhalt passt nicht“ alle Treffer, die nicht zum Thema passen – unabhängig davon, ob es sich um ein kommerzielles Angebot oder eine informationsorientierte Seite zu einem anderen Thema handelt.

Unter „Anderes“ sind die Bemerkungen zusammengefasst, die im einstelligen Bereich lagen, aber zu keiner der anderen Kategorien passten, wie z. B. „nur Links“ (sechs Nennungen) oder „Forum zum Thema“ (drei Nennungen).

„Subjektive Frage“ ist die einzige Kategorie, die nur für die Frage-Antwort-Dienste zählt. Sie erfasst jene Fragen, die auf die Meinung der Community abzielen und somit keine objektive Information generieren..

Die Auswertung der einzelnen Kategorien ergab vor allem bei den Social Bookmarking-Diensten einige Mängel. So entstammten 58,2 Prozent der Seiten, welche nicht geladen werden konnten, einer Anfrage an del.icio.us oder Mr. Wong. Von den 14 Treffern, die aufgrund ihrer Sprache nicht verstanden werden konnten, entstammten 13 del.icio.us, dem einzigen untersuchten Dienst ohne deutsche Lokalisierung. Ein weiterer Treffer wurde von MSN gefunden.

In den Kategorien „kommerziell“ und „Inhalt passt nicht“ wurden die Prozente der Kategorie aus der Gesamtmenge aller gefundenen Treffer innerhalb eines Suchdienstes errechnet. Hierbei fiel auf, dass bei allen algorithmischen Suchmaschinen die kommerziellen Seiten deutlich mehr Raum einnahmen als die unpassenden. (Google: 18,24 Prozent kommerziell, 13,14 Prozent unpassend; Yahoo!: 24,07 Prozent kommerziell, 15,88 Prozent unpassend; MSN: 28,04 Prozent kommerziell, 14,51 Prozent unpassend).

Dieses Ergebnis ist wahrscheinlich dem Versuch zuzuschreiben, möglichst alle interessanten Seiten zu finden, da die Maschinen nicht in der Lage sind, eine informationsorientierte Anfrage zu erkennen. Dem gegenüber stehen die Social Bookmarking-Dienste, deren Treffer eher unpassend als kommerziell sind. Dabei erscheinen die Zahlen für kommerzielle Treffer immer noch überraschend hoch, wenn man bedenkt, dass der Index dieser Dienste von einer Community aufgebaut wird

(Mr. Wong: 23,39 Prozent kommerziell, 28,21 Prozent unpassend; del.icio.us: 21,52 Prozent kommerziell, 28,27 Prozent unpassend).

Da die Frage-Antwort-Dienste bisher frei von kommerziellen Treffern sind, wurden hier nur die Werte für unpassende Treffer verglichen. Das Ergebnis dieses Vergleichs (Yahoo! Clever: 30,07 Prozent; Lycos iQ: 27,7 Prozent) zeigte aber wiederholt, dass das Ranking der Fragen verbessert werden müsste.

Beim Vergleich der Trefferanteile innerhalb der unpassenden Treffer zeigte sich, dass die algorithmischen Suchmaschinen im Mittelfeld lagen (Google: 12 Prozent, Yahoo!: 14 Prozent, MSN: 13 Prozent), während die Frage-Antwort-Dienste am wenigsten unpassende Treffer lieferten (Yahoo! Clever: 8 Prozent; Lycos iQ: 7 Prozent). Überraschend war die Tatsache, dass die Social Bookmarking-Dienste – trotz der geringeren Gesamttreffermenge – die meisten unpassenden Treffer fanden (Mr. Wong: 22 Prozent; del.icio.us: 24 Prozent) (vgl. Abb. 3.17).

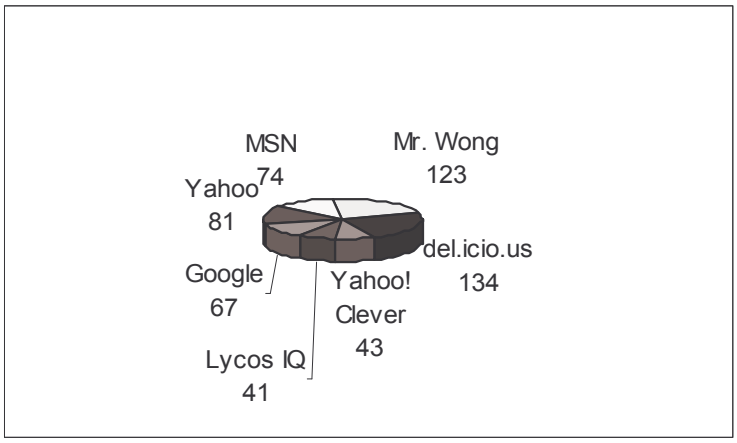


Abb. 3.17 Anteil an der Gesamtmenge „unpassender Treffer“ (n=563) verteilt nach Suchdienst

Beim Vergleich der kommerziellen Treffer pro Suchdienst lagen die Social Bookmarking-Dienste (je 18 Prozent) nur einen Prozentpunkt hinter Google (17 Prozent) (vgl. Abb. 4.14). Die anderen klassischen Suchmaschinen lagen bei 22 und 25 Prozent und produzierten somit fast die Hälfte der gesamten kommerziellen Treffer.

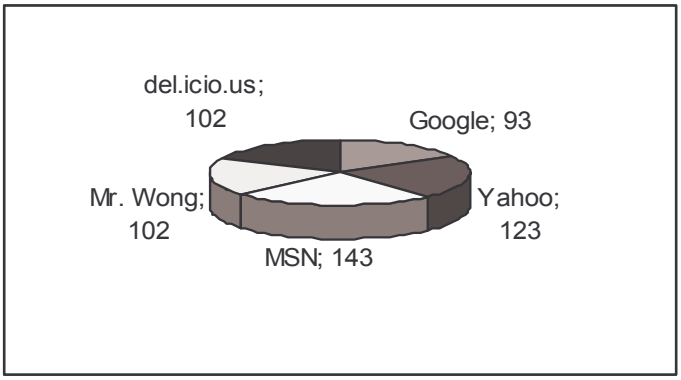


Abb. 3.18 Anteil an der Gesamtmenge „kommerzieller Treffer“ (n=564) - verteilt nach Suchdienst

Bei den als subjektiv bewerteten Fragen schneidete Yahoo! Clever mit 6,2 Prozent besser ab als Lycos IQ mit 8,1 Prozent. Der direkte Vergleich der absoluten Zahlen neun und zwölf zeigt, dass der Anteil von subjektiven Fragen bei beiden Diensten ähnlich war.

4. FAZIT

Aus den Ergebnissen wird ersichtlich, dass im deutschsprachigen Raum Suchmaschinen, hierbei allen voran Google, bei der Bereitstellung relevanter Treffer noch immer am besten abschneiden.

Die irrelevantesten Ergebnisse erzielten überaschenderweise die Social Bookmarking-Dienste. Die Erwartung, dass die von Nutzern generierten Linklisten eine höhere Precision als Frage-Antwort-Dienste aufweisen würden, erwies sich als falsch. Wir nahmen an, dass gute Links von vielen Personen gespeichert werden und somit eine hohe Position in der Trefferliste einnehmen würden [MaGr07]. Diese Vermutung hat sich durch die Untersuchung nicht bestätigt.

Stattdessen gab es besonders bei Social Bookmarking-Diensten sehr viele tote Links und doppelte Treffer in ein und derselben Trefferliste. Hier fehlt eine vernünftige Software, die den Bestand wartet. Der Nutzer scheint dazu nicht fähig oder hat kein Interesse daran. Somit ist unser Eindruck von diesen Diensten, dass sie zwar zur persönlichen Verwaltung von Lesezeichen und zum Austausch von Links innerhalb einer Gruppe sehr gut geeignet, aber als Suchplattform vorerst nicht weiter zu gebrauchen sind.

Entgegen den Ergebnissen bisheriger Studien zeigte Google große Übereinstimmung mit den anderen Suchdiensten. Es läßt sich behaupten, dass eine große Wahrscheinlichkeit besteht, dass Google den relevanten Link, den Yahoo! oder MSN einer suchenden Person ausgibt, ebenfalls findet.

Wie frühere Studien schon gezeigt haben, offeriert Google zwar insgesamt die relevantesten Links, liefert aber, für sich gesehen, gerade einmal 36 Prozent stark relevante Treffer. 35 Prozent der Treffer hingegen besitzen überhaupt keine Relevanz [GRB02] [Grie04] [Véro06]. Das ist kein befriedigendes Ergebnis.

Bei allen anderen Suchdiensten gab es mehr irrelevante als stark relevante Treffer. Bei der Einbindung der Fragen aus den Frage-Antwort-Diensten empfehlen wir, sie bei einer Google-Trefferliste auf den sechsten Rang zu platzieren, da ab hier bei Google die Precision stark abnimmt und somit die Frage aus dem Frage-Antwort-Dienst einen höheren Precision-Wert erreicht als der sich dort befindliche Treffer. Bei einer Trefferliste von Yahoo! würde sich das Einbinden nach dem dritten Treffer lohnen.

Weiterhin empfehlen wir Lycos, bei Lycos iQ ein besseres Ranking einzuführen. Dadurch würde sich die Precision vermutlich erhöhen. Bei unserem Test war es nicht ersichtlich, nach welchen Kriterien die Fragen geordnet wurden. Auch wäre eine Verbesserung der Suchfunktionen notwendig. Phrasenerkennung und eine automatische Berichtigung der Schreibweise sollten zum Standard gehören. Ein Ranking der Antworten, z. B. nach Top-Bewertung, würde bei einer zukünftigen Einbettung von Lycos iQ in eine Trefferliste ebenfalls hilfreich sein.

Verwendete Literatur

[GRB02] Griesbaum, Joachim/Rittberger, Marc/Bekavac, Bernard (2002): Deutsche Suchmaschinen im Vergleich: Alta Vista.de, Fireball.de, Google.de und Lycos.de. In: Womser-Hacker, Christa/Wolff, Christian/Hammwöhner, Reinhard (Hrsg.): Information und Mobilität: Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposium für Informationswissenschaft (ISI 2002). UVK: Konstanz.S. 201–223. Onlinedokument: http://www.inf-wiss.uni-konstanz.de/infwiss/download/isi2002/cc-isi2002-art14.pdf [Abruf am 16.01.2008]

[Grie00] Griesbaum, Joachim (2000): Evaluierung hybrider Suchsysteme im WWW.Diplomarbeit Universität Konstanz. Onlinedokument: http://www.inf.uni-konstanz.de/Prozent7Egriesbau/files/evaluierung_hybrider_suchsysteme_im_www.pdf [Abruf am 16.01.2008]

[Grie04] Griesbaum, Joachim (2004): Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. In: Information Research Bd. 9, H. 4, S. 189. Onlinedokument: http://informationr.net/ir/9-4/paper189.html [Abruf am 28.12.2007]

[LeHö07] Lewandowski, Dirk/Höchstötter, Nadine (2007): Qualitätsmessung bei Suchmaschinen: System- und nutzerbezogene Evaluationsmaße. In. Informatik Spektrum, Bd. 30, H. 3, S. 1-11. Onlinedokument: http://www.durchdenken.de/lewandowski/doc/IS_2007_Preprint.pdf [Abruf am 16.01.2008]

[Lewa05] Lewandowski, Dirk (2005): Web Information Retrieval: Technologien zur Informationssuche im Internet. Onlinedokument: http://www.durchdenken.de/lewandowski/web-ir/?25_Arten_von_Suchanfragen.html. [Abruf am 06.12.2007]

[Lewa06] Lewandowski, Dirk (2006): Themen und Typen der Suchanfragen an deutsche Web-Suchmaschinen.Onlinedokument: http://www.durchdenken.de/lewandowski/doc/mkwi2006.pdf. [Abruf am 26.12.2007]

[MaGr07] Maaß, Christian/Gräfe, Gernot (2007): Alternative Suchdienste: sieben Thesen zur Bedeutung des „Social Bookmarking“ In: Praxis der Wirtschaftsinformatik HMW (in Drucklegung) S. 1-12.

[Mizz97] Mizzaro, Stefano (1997): Relevance: the whole story. In: Journal of the American Society for Information Science, Bd, 48, S. 810-832.

[Sara96] Saracevic, Tefko (1996): Relevance reconsidered. In: Information Science: Integration in Perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2). Kopenhagen. S. 201-218.

[Stoc07] Stock, Wolfgang G. (2007): Infomation Retrieval: Informationen suchen und finden. Oldenbourg Wissenschaftsverlag: München.

[StSt00] Stock, Metchild/Stock, Wolfgang G. (2000): Internet – Suchwerkzeuge im Vergleich: Test 1: Retrievaltest mit Known Item Searches. In. Password, Bd. 15, H. 11, S. 23-31. Onlinedokument: http://philfak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/1078739217password_1.pdf [Abruf am 16.01.2008]

[Véro06] Véronis, Jean (2006): A comparative study of six search engines. Onlinedokument: http://sites.univ-provence.fr/veronis/pdf/2006-comparative-study.pdf [Abruf am 28.12.2007]

Weiterführende Literatur

1. Womser-Hacker, Chista (1989), Der PADOK-Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Systemen. Hildesheim et al. (Dissertation Universität Regensburg, Linguistische Informationswissenschaft)
2. Machill, Marcel/Welp, Carsten (Hrsg.) (2003): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. Bertelsmann Stiftug: Gütersloh.