Datumsbeschränkung bei WWW-Suchanfragen

Eine Untersuchung der Möglichkeiten der zeitlichen Einschränkung von Suchanfragen in den Suchmaschinen Google, Teoma und Yahoo

Dirk Lewandowski

Heinrich-Heine-Universität Düsseldorf
Institut für Sprache und Information
Abt. Informationswissenschaft
Universitätsstrasse 1
D – 40225 Düsseldorf
dirk.lewandowski@uni-duesseldorf.de

Zusammenfassung:
Suchmaschinen bieten als erweiterte Suchfunktion die Einschränkung der Suchanfrage auf Dokumente aus einem bestimmten Zeitraum an. Die Feststellung des tatsächlichen Datums einer inhaltlichen Änderung von Web-Dokumenten bereitet jedoch Probleme. Die Funktionen zur Datumsbeschränkung werden bei den Suchmaschinen Google, Teoma und Yahoo untersucht. Dabei wird festgestellt, dass die Datumsbeschränkung von diesen Suchmaschinen nur unzureichend beherrscht wird. Abschließend wird ein Ausblick auf Möglichkeiten gegeben, wie sich das Aktualisierungsdatum von Web-Dokumenten besser bestimmen ließe.

1 Einleitung

Verschiedenen Untersuchungen haben gezeigt, dass Suchmaschinen-Nutzer die angebotenen erweiterten Suchmöglichkeiten[1] nicht oder nur selten nutzen (aktuell u.a. Spink 2003; Machill et al. 2003, 166ff.). Wird nach den Gründen der Nicht-Nutzung gefragt, so wird deutlich, dass „die meisten Nutzer nicht willens [sind], bei der Formulierung ihres Suchziels allzu viel kognitive und zeitliche Energie aufzuwenden“ (Machill et al. 2003, 169). Dies erklärt die

[1] Einen Überblick über die Suchmöglichkeiten bei den wichtigsten Suchmaschinen bietet Lewandowski (2004).

seltene Nutzung etwa von Booleschen Operatoren und der Klammersetzung innerhalb von Suchanfragen. Bei intuitiv „einleuchtenden“ Möglichkeiten wie der Beschränkung der Suche auf aktuelle Dokumente reicht diese Erklärung jedoch nicht aus. Eine solche Einschränkung ließe sich ohne großen Aufwand realisieren und wäre geradezu ein Kandidat, um von der erweiterten Suchmaske in das reguläre Suchformular überführt zu werden.

Allerdings ist die Beschränkung der Suche auf einen bestimmten Zeitraum bei Web-Dokumenten mit verschiedenen Schwierigkeiten verbunden. Diese sollen im ersten Teil dieses Texts dargestellt werden. Auf der anderen Seite gibt es – vor allem aus dem professionellen Umfeld – auch viele Klagen darüber, dass die von den Suchmaschinen angebotenen Datumsbeschränkungen nicht oder nur eingeschränkt funktionieren (Price, Tyburski 2003). Eine systematische Untersuchung dieser Funktionen lag allerdings nach Kenntnis des Verfassers bislang nicht vor.

Im Hauptteil des Texts soll deshalb eine Untersuchung vorgestellt werden, die verschiedene populäre Suchmaschinen hinsichtlich ihrer Fähigkeit, dem Nutzer relevante aktuelle Ergebnisse zu liefern und damit die im ersten Teil dargestellten Probleme zu lösen, testet.

2 Ermittlung von Datumsangaben aus Web-Dokumenten

Im Folgenden soll unter dem Datum eines Dokuments der Aktualisierungszeitpunkt inhaltlicher Elemente des Dokuments (also in der Regel des *Texts*) verstanden werden; andere Aktualisierungen wie beispielsweise die Anpassung des Layouts oder die Aktualisierung des Copyright-Vermerks sollen nicht als Aktualisierung gewertet werden.

Die Textauszeichnungssprache HTML bietet kein eigenes Tag, in dem das Datum eines Dokuments angegeben wird. Deshalb sind Suchmaschinen bei der Aufnahme von Dokumenten in ihren Index auf andere Indikatoren für die Ermittlung des Datums eines Dokuments angewiesen. Prinzipiell existieren vier Möglichkeiten der Bestimmung des Datums eines HTML-Dokuments:

- Auswertung der Angaben des Servers, auf dem das Dokument abgelegt ist
- Verwendung des Datums des ersten Auffinden des Dokuments durch die Suchmaschine
- Auswertung der Angaben in den Metadaten des Dokuments

- Auswertung des Inhalts des Dokuments in Hinblick auf eventuell vorkommende Datumsangaben

Bei der Auswertung der Server-Angaben besteht das Problem, dass hier das Datum angegeben wird, an dem die *Datei* zum letzten Mal aktualisiert wurde. Dies bedeutet jedoch nicht, dass das Dokument entsprechend inhaltlich verändert wurde. Dazu kommt, dass bei Dokumenten, die mit Content-Management-Systemen generiert wurden, oft das jeweils tagesaktuelle Datum zurückgegeben wird. Ist das Datum im Server falsch eingestellt, überträgt sich dies auf die zurückgegebenen Aktualisierungsdaten.

Das Datum des ersten Auffindens des Dokuments kann ein Anhaltspunkt für eine von der Suchmaschine zu erfassende Datumsangabe sein. Veränderungen des Dokuments müssen natürlich weiterhin als Aktualisierungen gewertet werden. Hierbei muß wieder klar zwischen einer Veränderung des Texts und einer Veränderung anderer Elemente des Dokuments unterschieden werden.

Ein spezifisches Problem dieses Ansatzes ist die Menge all der Dokumente, die vor dem Start der jeweiligen Suchmaschine erstellt und nicht mehr verändert wurden. Diesen kann nur das Datum des Beginns der Indexierung durch die Suchmaschine zugeordnet werden. Weitere Probleme ergeben sich, wenn die Indexgröße der Suchmaschine beschränkt ist (was in der Regel der Fall ist) und diese erweitert werden soll.

Datumsangaben in den Metainformationen eines Dokuments wären eine gute Möglichkeit, das tatsächliche Datum des Dokuments zu ermitteln. Eine Angabe ist sowohl in den „regulären“ Metadaten als auch in speziellen Metadaten-Sets wie z.B. Dublin Core vorgesehen. Zusätzlich besteht bei den Metadaten eine klare Vorgabe, in welchem Format die Angaben zu machen sind.

Bei einer Voruntersuchung zu dieser Untersuchung am 4.2.2004 wurde allerdings festgestellt, dass nur ein verschwindend geringer Anteil der ausgewerteten Dokumente eine Datumsangabe in den Metaangaben enthielt; bei insgesamt etwa 500 betrachteten Dokumenten waren bei nur vier Dokumenten solche Angaben vorhanden. Deshalb wurde im Weiteren auf die Auswertung dieser Angaben verzichtet.

Die letzte Methode, das Datum eines HTML-Dokuments zu ermitteln, ist die Auswertung seines Inhalts. Datumsangaben haben ein bestimmtes Format (wenn dieses auch variieren kann; z.B. europäisches vs. us-amerikanisches

Datumsformat) und können daher maschinell gefunden und ausgewertet werden. Des weiteren werden Datumsangaben, die sich auf das Erstellungs- bzw. Aktualisierungsdatum des Dokuments beziehen, in der Regel an bestimmten Stellen des Dokuments vorkommen (meist am Anfang oder am Ende), so dass das Auffinden dieser Angaben erleichtert wird. Teilweise werden die Datumsangaben auf den Seiten allerdings automatisch generiert und immer das aktuelle Datum eingesetzt. Als einziger Ausweg ist hier der Vergleich des Inhalts des Texts in seiner alten und seiner neuen Version zu sehen, welcher allerdings einen gewissen Aufwand erfordert.

Heutige Suchmaschinen werten die Angaben, die vom Server zurückgegeben werden und teils auch das Datum des ersten Auffinden des Dokuments und dessen Veränderungsfrequenz aus. Die Auswertung von Metadaten scheitert aufgrund dessen, dass diese von den Autoren der Dokumente nur selten angegeben werden. Datumsangaben innerhalb des Dokumententexts werden bisher nicht ausgewertet. Dies wird durch die Ergebnisse der hier vorgestellten Studie bestätigt.

In den erweiterten Suchformularen der Suchmaschinen wird die Datumsbeschränkung in der Regel durch „Aktualisierung“ umschrieben. Dabei wird der Sachverhalt verschleiert, da die Nutzer davon ausgehen, dass mit dem Aktualisierungsdatum dasjenige des auf der jeweiligen Seite präsentierten Textes gemeint ist und nicht die Aktualisierung von Layout-Elementen oder schlicht das Neu-Aufspielens des Dokuments auf den Server.

3 Zielsetzung

Nachdem die grundsätzliche Problematik der Datumsbeschränkung bei Suchmaschinen bekannt war, sollte in einer Untersuchung überprüft werden, in wie weit Suchmaschinen überhaupt in der Lage sind, solche Einschränkungen korrekt vorzunehmen. Dafür wurden 50 Suchanfragen ausgewählt und an vier verschiedene Suchmaschinen gestellt; einmal ohne Datumsbeschränkung, einmal mit der Einschränkung auf Dokumente, die innerhalb der letzten sechs Monate erstellt wurden. Der Zeitraum von sechs Monaten wurde gewählt, da dieser von allen untersuchten Suchmaschinen unterstützt wird. Google und Yahoo lassen keine genaue Datumsbeschränkung zu, sondern nur Aktualisierungszeiträume. Der Test wurde am 3. April 2004 durchgeführt.

Für die ersten 20 ausgegebenen Treffer sollten Aktualitätsquoten berechnet werden, die den Anteil derjenigen Dokumente, die aus dem letzten halben

Jahr stammen, wiedergeben. Damit sollte festgestellt werden, ob sich die Datumseinschränkung für einen Rechercheur "lohnt", d.h. ob es gelingt, mit dieser Einschränkung tatsächlich nur aktuelle Dokumente zu finden und entsprechend inaktuelle Dokumente auszuschließen. Letztlich sollte bestimmt werden, welche Suchmaschine am Geeignetsten für datumsbeschränkte Suchanfragen ist.

In der Untersuchung sollten alle gefundenen Seiten auf ein Aktualisierungsdatum hin untersucht werden. War ein solches vorhanden, wurde es notiert und ging in die Auswertung mit ein. Wenn kein Aktualisierungsdatum vorhanden war oder dieses nicht eindeutig war, ging der entsprechende Treffer nicht in die Auswertung mit ein.

4 Methodik

4.1 Auswahl der Suchmaschinen

Für diese Untersuchung wurden die Suchmaschinen Google, Yahoo und Teoma ausgewählt. Dies sind diejenigen Suchmaschinen, die die weltweit größten und am meisten benutzten Indizes anbieten (vgl. Sullivan 2003). Die früher bedeutenden Suchmaschinen All the Web und Alta Vista arbeiten seit April 2004 nicht mehr mit eigenen Datenbeständen, sondern basieren mittlerweile auf dem Yahoo-Index. Dieser wiederum ist als Nachfolger der Inktomi-Datenbank anzusehen, weshalb auch die typischen „Inktomi-Suchmaschinen“ wie bspw. HotBot außer Acht gelassen wurden. Auch bei diesen Suchmaschinen ist mit einer baldigen Umstellung auf die Yahoo-Datenbank zu rechnen. Ebenso unberücksichtigt blieben speziell auf einen Sprachraum oder ein Thema beschränkte Suchmaschinen.

4.2 Auswahl der Suchanfragen

Die Auswahl der Testfragen sollte zufällig erfolgen. Die Suchanfragen für diese Untersuchung wurden über die "Live-Suche" von Fireball[2] ausgewählt, in der Suchanfragen angezeigt werden, die jeweils aktuell an Fireball gestellt werden. Diese Vorgehensweise gewährleistet die zufällige Auswahl der

[2] http://www.fireball.de/livesuche.csp [8.4.2004]

Suchanfragen und die Orientierung am tatsächlichen Suchverhalten der Nutzer. Aufgrund der Zielsetzung von Fireball, das deutschsprachige Web zu erschließen und entsprechende Suchanfragen zu beantworten, waren diese größtenteils deutschsprachig.

Die Anfragen wurden am 15.3.2004 ermittelt; ausgeschlossen wurden Anfragen in der Bildersuche und im internationalen Index. Beide werden in der Live-Suche gesondert angegeben, so dass die Auswahl der Anfragen an den deutschsprachigen Index als zuverlässig anzusehen ist. Weiterhin ausgeschlossen wurden Suchanfragen, die auf ein pornographisches Interesse hindeuteten. Schließlich wurden die gefundenen Suchanfragen von Dubletten gereinigt.
Mittels dieser Methode wurden insgesamt 50 Anfragen ausgewählt, die für die weitere Untersuchung genutzt wurden. Für eventuell auftauchende Problemfälle wie z.B. einem Ergebnis von null Treffern für eine Suchanfrage wurden weitere Suchanfragen vorbereitet, die als Ersatz verwendet werden konnten.

4.3 Testaufbau

Für die Untersuchung wurden die 50 ausgewählten Suchanfragen an die unterschiedlichen Suchmaschinen gerichtet. Ausgewertet wurden die ersten 20 Plätze der Trefferlisten jeweils in der Standardsuche und in der Suche nach Dokumenten der letzten sechs Monate.

Die Standardeinstellungen der Suchmaschinen wurden beibehalten, so dass Dokumente in einer beliebigen Sprache gefunden wurden. Bei Yahoo wurde jeweils die „weltweite Suche“ manuell ausgewählt.

Für die Auswertung wurden die jeweils 20 höchstplatzierten Treffer aus den Trefferlisten entnommen. Wurden 20 oder weniger Treffer ausgegeben, so wurde die Trefferliste vollständig ausgewertet.

Bei der Auswertung der Treffer wurde keinerlei Überprüfung der Relevanz der Treffer vorgenommen. Das einzige Kriterium der Auswertung war das Vorkommen einer Datumsangabe im Dokument. Berücksichtigt wurden alle ausgegebenen Dateitypen. Bei der Durchführung des Tests wurden allerdings nur Ergebnisse in den Formaten HTML und PDF gefunden.

Wenn in den Trefferlisten tote Links auftauchten, so wurden diese ignoriert. Die Trefferliste wurde stets so weit ausgewertet, bis der Schwellenwert von 20 abrufbaren Dokumenten erreicht wurde. Bezahlte Treffer ("sponsored listings" etc.), die über, unter oder neben den Trefferlisten angezeigt wurden, wurden in der Auswertung ignoriert.

4.4 Auffälligkeiten bei einzelnen Suchmaschinen

Schon bei einem ersten Stichprobentest im Vorfeld der Untersuchung fiel auf, dass bei Google die Datumsbeschränkung in der erweiterten Suche vollkommen wirkungslos ist. Das heißt: gleichgültig, ob das Datum eingeschränkt wird oder nicht, bleiben die Ergebnisse und deren Anordnung gleich. Warum diese Funktion überhaupt noch in der erweiterten Suche vorhanden ist, ist rätselhaft. Es handelt sich auch nicht um einen temporären „Bug"; dieser Fehler besteht seit mindestens November 2003. Allerdings besteht eine Möglichkeit, die Suche doch noch erfolgreich über das Datum einzuschränken; dazu muss die Datumsangabe jedoch in Form eines Befehls eingegeben werden.[3] Allerdings verwendet Google intern julianische Datumsangaben[4] (Calishain, Dornfest 2003, 35). Alle Suchanfragen müssen also erst in dieses Format übersetzt werden. Da dies manuell nicht zu leisten ist, gibt es Interfaces wie beispielsweise das „Google Ultimate Interface"[5], die eine einfache Suche nach dem Datum ermöglichen. Allerdings ist vom durchschnittlichen Suchmaschinen-Nutzer nicht zu erwarten, dass er entsprechende Tricks kennt, um die Suchmaschine trotz ihres offensichtlichen Fehlers zu „überlisten".[6]

4.5 Auswertung der Datumsangaben

Die den Test durchführenden Personen wurden gebeten, auf den gefundenen Webseiten nach Datumsangaben zu suchen. Wenn ein *Aktualisierungs*datum identifiziert werden konnte, sollte dies auf einem Erhebungsbogen notiert werden. Folgende Regeln wurden angewendet:

---

[3] Dieser lautet *daterange:{Startdatum}-{Enddatum}*

[4] Das julianische Datum wird in Tagen seit dem 1. Januar 4713 vor unserer Zeit gemessen. Der Tag beginnt jeweils um 12 Uhr mittags. Das julianische Datum für den 8. April 2004 nachmittags lautet beispielsweise 2453104.

[5] http://www.faganfinder.com/google.html [8.4.2004]

[6] Das „Google Ultimate Interface" wird von Google nicht offiziell unterstützt. Allerdings setzt es der Suchanfrage nur eine entsprechende Ergänzung um den Daterange-Befehl hinzu und schickt die Anfrage direkt an Google. Die ausgegebene Trefferliste kommt direkt von Google und läuft nicht mehr über das „Ultimate Interface", so dass Manipulationen ausgeschlossen werden können. Der Daterange-Befehl wird im Google-API („Application Programming Interface") ausdrücklich unterstützt.

- Wenn das Dokument ein explizites Änderungsdatum im Text enthielt, wurde dieses gewertet. Ein solches Änderungsdatum konnte beispielsweise durch einen Hinweis am Seitenanfang oder -ende wie "last modified: ...." ausgedrückt werden. Auch bestimmte Texttypen wie Nachrichtenmeldungen, die in der Regel datiert sind, konnten entsprechend ausgewertet werden.
  Allerdings enthalten einige Seiten automatisch generierte Datumsangaben, die keine echte Aktualisierung anzeigen. Ausgeschieden wurden solche Seiten, die neben dem aktuellen Datum auch die aktuelle Uhrzeit enthielten. Weiterhin ausgeschieden wurden Seiten mit einer Datumsangabe, die aufgrund des Inhalts eindeutig als automatisch generiert identifiziert werden konnten. Seiten mit automatischer Datumsangabe wurden gesondert gezählt.
- Enthielt das untersuchte Dokument einen Copyright-Hinweise, so bestand dieser in nahezu allen Fällen lediglich aus einer Jahreszahl. In vielen Fällen wird dieser Hinweis automatisch generiert und für alle Dokumente einer Site auf das aktuelle Jahr gesetzt. Copyright-Hinweise mit der Jahresangabe 2004 oder 2003 wurden daher nicht in die Auswertung mit einbezogen; lautete der entsprechende Hinweis jedoch 2002 oder älter, so wurde dies als Zeichen für die Inaktualität der Seite gewertet und ging in die Wertung mit ein.
- Teils wurden auf den Seiten auch Datumsangaben gefunden, die in der Zukunft lagen. Solche Angaben wurden ignoriert.
- Die Testdurchführenden wurden darum gebeten, die in den europäischen und us-amerikanischen Datumsangaben bestehenden Unterschiede (Reihenfolge von Tag und Monat) zu beachten.

Mit dieser Methode konnte festgestellt werden, dass zwischen 28 und 33 Prozent der untersuchten Seiten eine Datumsangabe beinhalten (vgl. Tabellen 1 und 2). Die Unterschiede zwischen der Betrachtung derjenigen Seiten, die bei der uneingeschränkten Suche gefunden wurden, und der, die bei der eingeschränkten Suche gefunden wurden, sind nicht signifikant.

Mit etwa 30 Prozent der gefundenen Seiten, die eine Datumsangabe enthalten, wurde jedoch eine Anzahl von Dokumenten gefunden, die eine Auswertung der Leistungsfähigkeit der Suchmaschinen auf dieser Basis möglich macht. Eine statistische Überprüfung ergibt, dass die Unterschiede zwischen den einzelnen Suchmaschinen hinsichtlich des Anteils der prüfbaren Seiten nicht signifikant sind.

Tabelle 1: Anteil der Seiten mit Datumsangaben im gesamten Index

Suchmaschine	Anzahl untersuchte Treffer für die 50 Beispielanfragen*	Anzahl der Seiten mit Datumsangabe	Anteil der Seiten mit Datumsangabe in Prozent
Teoma	933	313	33,55
Google	978	308	31,49
Yahoo	979	296	30,23

* Da je Suchanfrage die ersten 20 Treffer ausgewertet wurden, konnten bei den 50 Anfragen insgesamt maximal 1.000 Treffer erreicht werden. Bei einigen Suchanfragen wurden jedoch weniger als 20 Treffer gefunden, so dass sich die Zahl entsprechend reduziert und je nach Suchmaschine variiert.

Tabelle 2: Anteil der Seiten mit Datumsangaben; nur Dokumente, die von den Suchmaschinen innerhalb der letzten sechs Monate datiert wurden.

Suchmaschine	Anzahl untersuchte Treffer für die 50 Beispielanfragen*	Anzahl der Seiten mit Datumsangabe	Anteil der Seiten mit Datumsangabe in Prozent
Teoma	933	308	33,01
Google	971	279	28,73
Yahoo	972	284	29,22

* Da je Suchanfrage die ersten 20 Treffer ausgewertet wurden, konnten bei den 50 Anfragen insgesamt maximal 1.000 Treffer erreicht werden. Bei einigen Suchanfragen wurden jedoch weniger als 20 Treffer gefunden, so dass sich die Zahl entsprechend reduziert und je nach Suchmaschine variiert.

5 Ergebnisse

5.1 Aktualität der Dokumente

Es wurde gemessen, wie viele der Dokumente aus den Top 20 der Trefferlisten tatsächlich aus den letzten sechs Monaten stammen. Der Anteil dieser Dokumente am Gesamt der untersuchten Dokumente wird im Weiteren als Aktualitätsquote bezeichnet. Diese Quote wurde sowohl für die Suche mit als auch die Suche ohne Datumsbeschränkung errechnet.

Tabelle 3: Aktualitätsquoten der untersuchten Suchmaschinen

Suchmaschine	Aktualitätsquote Standardsuche	Aktualitätsquote bei Suche mit Datumsbeschränkung	Steigerung in Prozent
Teoma	37,06	37,34	0,76
Google	48,70	59,50	22,18
Yahoo	40,54	54,23	33,77

Teoma findet bei der Suche mit Datumsbeschränkung keinen höheren Anteil an aktuellen Dokumenten als bei der Suche ohne Datumsbeschränkung. Auch bietet Teoma den geringsten Anteil an aktuellen Dokumenten. Yahoo liegt bei der uneingeschränkten Suche bei einer Aktualitätsquote von 40,5 Prozent, Google bei 48,7 Prozent. Bei Google stammt also in der uneingeschränkten Suche beinahe jedes zweite Dokument aus dem letzten halben Jahr.

Beschränkt man die Suche auf Dokumente des letzten halben Jahres, so kann Yahoo die Aktualitätsquote auf 54,2 Prozent steigern, Google sogar auf 59,5 Prozent. Dies bedeutet allerdings auch, daß selbst bei der hier am besten bewerteten Suchmaschine Google noch 40 Prozent der gefundenen Dokumente falsch zugeordnet wurden, d.h. nicht innerhalb des eingestellten Zeitraums zu datieren sind.

Betrachtet man die Steigerung der Aktualitätsquote, so zeigt sich, dass Yahoo hier den höchsten Wert vorweisen kann. Während Google mit 59,50 Prozent aktueller Dokumente zwar absolut besser abschneidet, kann Yahoo eine Steigerung von 33,77 Prozent verzeichnen. Google scheint hingegen generell Dokumente, die in kürzeren Abständen aktualisiert werden, zu bevorzugen.

Betrachtet man statt der insgesamt gefundenen Dokumente die Ergebnisse der einzelnen Suchanfragen, zeigt sich bei den einzelnen Suchmaschinen eine unterschiedliche Verteilung (siehe Abbildungen 1-3). Die Aktualitätsquote schwankt bei allen Suchmaschinen zwischen den einzelnen Suchanfragen erheblich. Keine Suchmaschine bewegt sich durchweg bei einer mittleren oder hohen Aktualitätsquote. Google und Teoma gelingt es allerdings häufiger als Yahoo, eine Aktualitätsquote von 100 Prozent zu erreichen. Dafür fällt aber bei beiden Suchmaschinen auch auf, dass sie deutlich öfter als Yahoo eine Quote von weniger als zehn Prozent erreichen. Die Verteilung bei Yahoo ist am ehesten gleichmäßig.

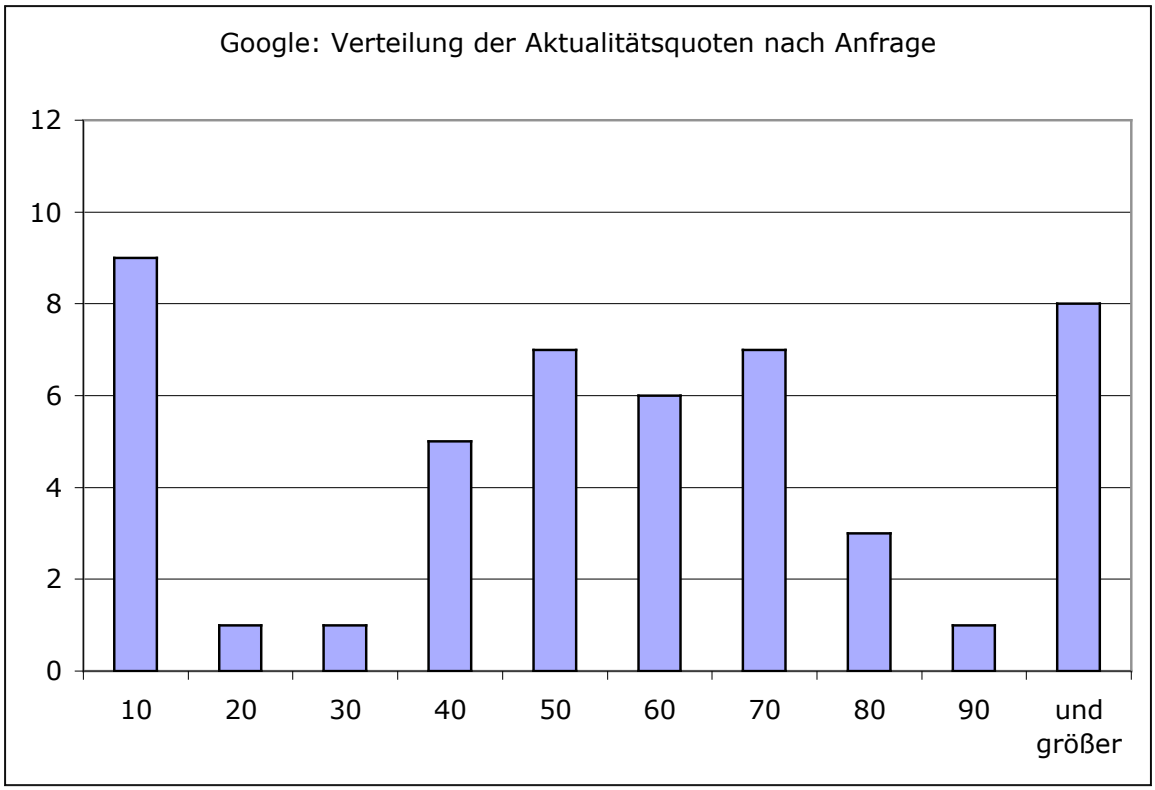

Abbildung 1: Verteilung der Aktualitätsquote nach Suchanfragen bei Google

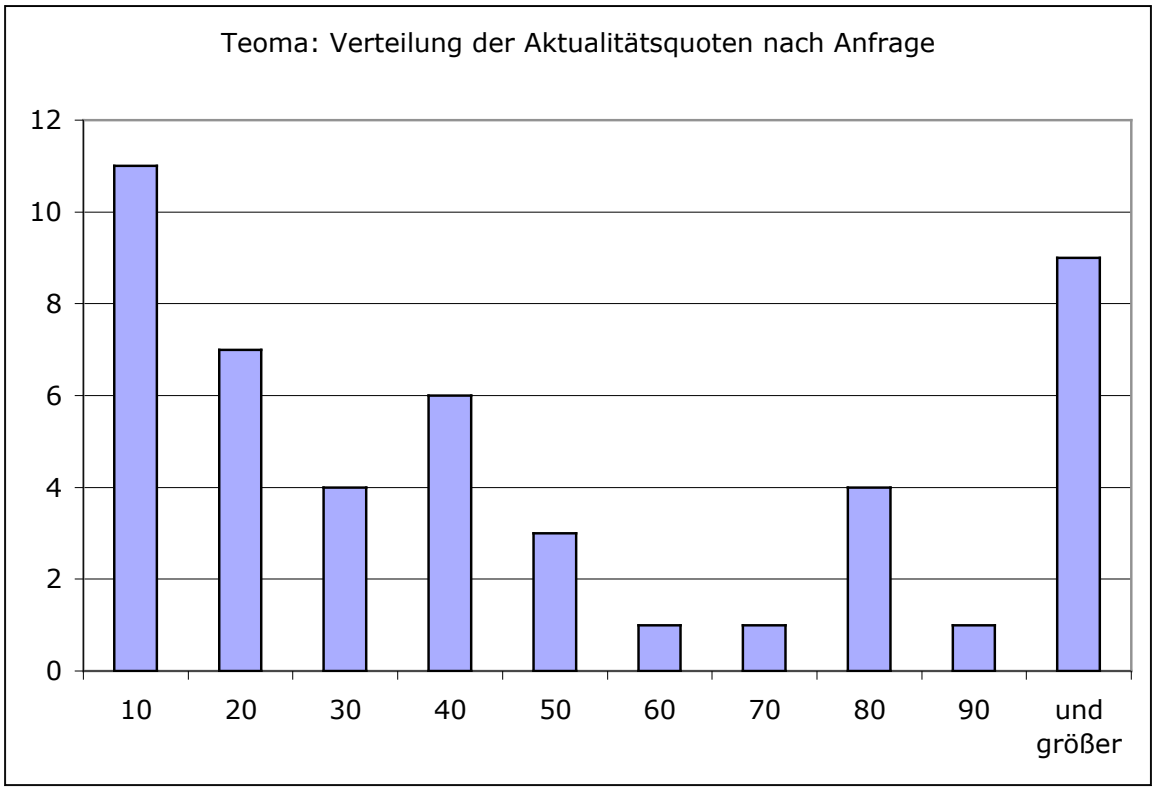

Abbildung 2: Verteilung der Aktualitätsquote nach Suchanfragen bei Teoma

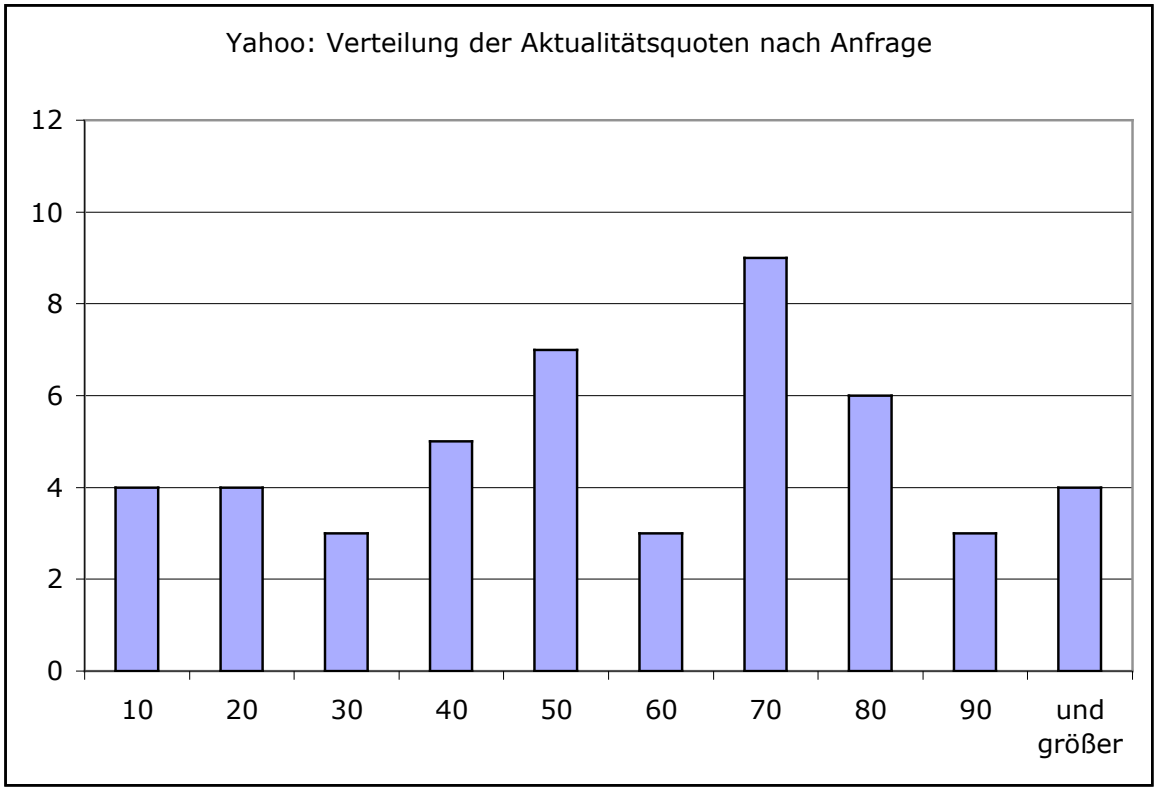

Abbildung 3: Verteilung der Aktualitätsquote nach Suchanfragen bei Yahoo

5.2 Fehlerquote

Für den Suchenden stellt sich nicht nur die Frage, welcher Anteil der gefundenen Dokumente richtig zugeordnet werden konnte, sondern auch die Frage nach den offensichtlich falsch zugeordneten Dokumenten. Bisher wurde als Grundlage für den Sucherfolg der Suchmaschinen nur die Quote der aktuellen Dokumente am Gesamt aller gefundenen Dokumente gewertet. Das Gegenstück zur Aktualitätsquote ist die Fehlerquote – sie misst den Anteil der falsch zugeordneten Dokumente.

Bei Ansicht der Fehlerquoten (Tabelle 4) zeigt sich, dass die Suchmaschine Teoma deutlich mehr Dokumente falsch einschätzt als sie richtig zuordnen kann. Die Fehlerquote liegt bei 62,66 Prozent. Besser schneidet Yahoo ab; hier liegt die Fehlerquote allerdings auch noch bei 45,77 Prozent. Selbst beim Testsieger Google mit der geringsten Fehlerquote werden noch 40,5 Prozent der Dokumente falsch zugeordnet. Die statistische Überprüfung ergibt, dass die Unterschiede signifikant sind.

Tabelle 4: Fehlerquoten bei der Datumsbegrenzung

Suchmaschine	richtig eingeschätzt	falsch eingeschätzt	Fehlerquote in Prozent
Teoma	115	193	62,66
Google	166	113	40,50
Yahoo	154	130	45,77

Die hohen Fehlerquoten aller Suchmaschinen bestätigen die Vermutung, dass die Suchmaschinen das tatsächliche Datum eines Dokuments nur schwer ermitteln können.

Für den Nutzer stellt sich aufgrund der insgesamt unbefriedigenden Ergebnisse aller Suchmaschinen die Frage, ob er die Datumsbeschränkung benutzen soll oder nicht. Tabelle 5 zeigt, in wie vielen Fällen es sich lohnt, die Suche entsprechend einzuschränken oder nicht. Nicht mit in die Auswertung gingen hier diejenigen Suchanfragen ein, bei denen sowohl ohne als auch mit Beschränkung eine Quote von 100 Prozent erreicht wurde.

Yahoo schneidet in dieser Auswertung am besten ab. Allerdings verbessert sich auch bei dieser Suchmaschine das Ergebnis in nur etwas mehr als zwei Dritteln der Anfragen. Interessant ist der bei allen untersuchten Suchmaschinen relativ hohe Anteil von Anfragen, bei denen sich das Ergebnis bei der Datumsbeschränkung verschlechtert sowie der Anteil der Anfragen, bei denen die Datumsbeschränkung nichts verändert.

Tabelle 5: Verbesserung bzw. Verschlechterung der Aktualitätsquote durch die Datumsbeschränkung

Suchmaschine	schlechter	gleich	besser
Teoma	14	17	16
Google	8	12	25
Yahoo	7	10	30

5.3 Sieger je Anfrage

Abbildung 4 zeigt, welche Suchmaschine wie viele Suchanfragen im Vergleich am besten beantworten konnte, unabhängig davon, welche Aktualitätsquote erreicht wurde. Als am besten gilt hier diejenige

Suchmaschine, die in der datumsbeschränkten Suche die beste Aktualitätsquote erreicht. Es wurden jeweils Ränge vergeben; wenn zwei Suchmaschinen die gleiche Aktualitätsquote erreichten, erhielten sie den gleichen Rangplatz und der dritten Suchmaschine wurde der nächst niedrige Rangplatz zugewiesen. Wenn die Aktualitätsquote bei einer Suchmaschine bei Null lag, wurde auf jeden Fall der dritte Platz zugewiesen.

Es zeigt sich, dass Yahoo bei insgesamt 24 Suchanfragen den ersten Platz belegt, Google folgt mit 18 ersten Platzierungen. Zwar konnte in Abschnitt 5.1 festgestellt werden, dass Google insgesamt die höchste Aktualitätsquote erreicht, dies trifft jedoch nicht auf alle Suchanfragen zu. Aus der Verteilung der Sieger nach Suchanfragen lässt sich keine eindeutige Empfehlung aussprechen. Auch der Gewinner Yahoo belegt nur in knapp der Hälfte der Suchanfragen den ersten Platz. Es scheint also stark von der Suchanfrage abzuhängen, welche Suchmaschine die beste Wahl in Bezug auf aktuelle Dokumente ist. Selbst Teoma, also die Suchmaschine, die am schlechtesten abschneidet, liefert in 30 Prozent der Suchanfragen (mit) das beste Ergebnis.

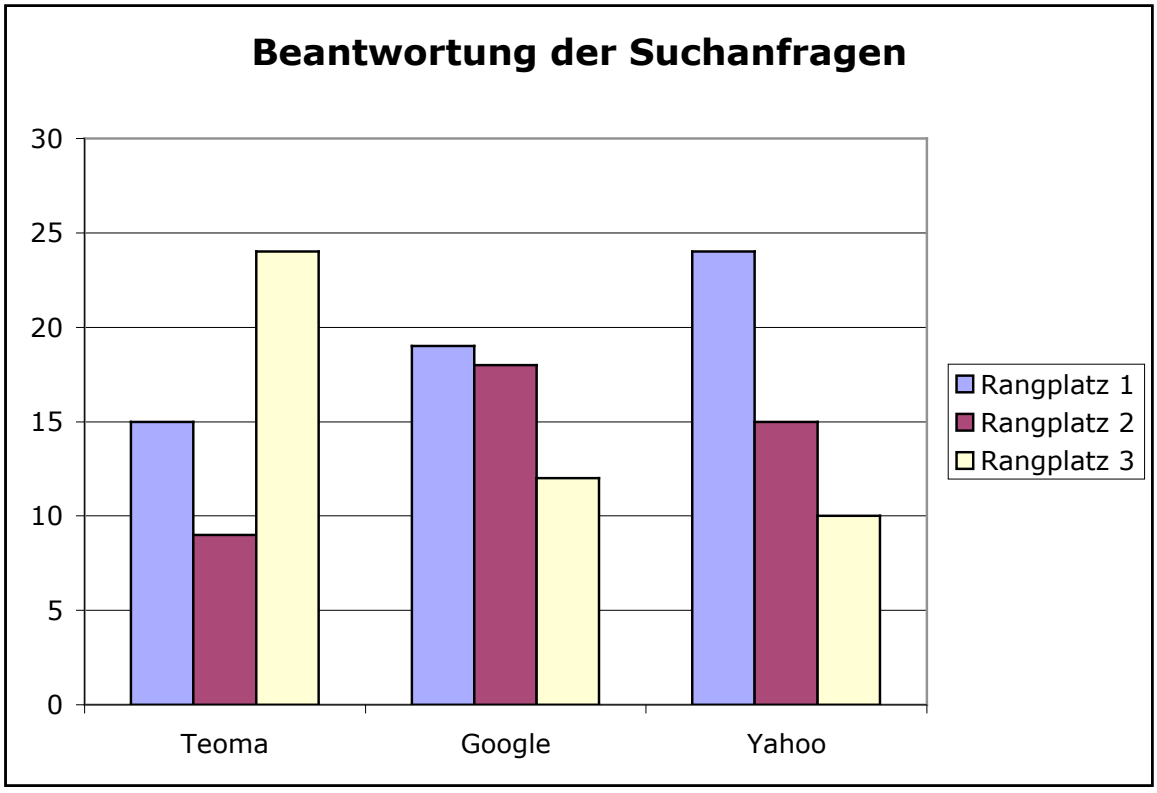

Abbildung 4: Rangplätze in Bezug auf die Datumsbeschränkung der 50 Suchanfragen

6 Fazit

Die Ergebnisse der Studie belegen, dass die Datumsbeschränkung bei den gängigen WWW-Suchmaschinen nur unzureichend funktioniert. Es konnte gezeigt werden, dass von den getesteten Suchmaschinen Google zwar insgesamt die beste Aktualitätsquote erreichte, jedoch bei weitem nicht bei allen Anfragen das beste Ergebnis lieferte.

Für den Nutzer lässt sich keine eindeutige Empfehlung aussprechen. Als sicher kann aber gelten, dass die Datumsbeschränkung bei Teoma nicht lohnenswert ist. Die Aktualitätsquote kann hier nicht signifikant erhöht werden. Zwar kann Teoma bei einigen Suchanfragen die Aktualität der gefundenen Dokumente erhöhen, insgesamt wird jedoch deutlich mehr als die Hälfte der Dokumente falsch eingeschätzt.

Für die Suchmaschine Google bleibt festzustellen, dass die Datumsbeschränkung im erweiterten Suchformular seit längerer Zeit nicht funktioniert. Die einzige Möglichkeit für den Nutzer, die Datumsbeschränkung zu nutzen, besteht in der Benutzung alternativer Suchformulare oder der Eingabe in Kommandosprache.

Die Untersuchung bestätigt die in Abschnitt 2 dargestellte Problematik und zeigt, dass die dargestellten Schwierigkeiten bei weitem noch nicht gelöst sind. In der Forschung ist die Problematik zwar bekannt; die Lösung wird jedoch nicht als dringlich angesehen. So wird das Thema in Werken, die sich mit zukünftigen Fragen des Web-Information-Retrieval auseinandersetzen, schlicht nicht behandelt (so z.B. in Chakrabarti 2003; Henzinger, Motwani, Silverstein 2002).

Offensichtlich ist, dass sich das tatsächliche Datum eines Dokuments, wie es in dieser Untersuchung verstanden wird, nicht allein durch die Ermittlung einer der genannten Datumsangaben ermitteln lässt. Als einzige Möglichkeit, die Zuverlässigkeit zu verbessern, erscheint die Kombination der verschiedenen Angaben und der (wenigstens näherungsweisen) Berechnung des tatsächlichen Datums des Dokuments aus diesen Angaben. Insbesondere die Ergänzung der bisherigen Verfahren um die bisher nicht eingesetzte Möglichkeit der Extrahierung des Datums aus dem Dokumenttext selbst erscheint vielversprechend.

Des Weiteren sollten Suchmaschinen den Veränderungs*grad* von Dokumenten mit in die Datumsermittlung einbeziehen. Ntoulas, Cho und

Olston (2004) fanden heraus, dass die von den meisten Suchmaschinen beachtete Veränderungsfrequenz allerdings kein guter Indikator für den Veränderungsgrad ist. Würde der Veränderungsgrad beachtet, ließen sich kleinere Änderungen am Dokument wie etwa der Austausch von Werbeanzeigen oder Verweisen auf andere Seiten des gleichen Webangebots also solche erkennen und würden nicht mehr wie bisher als Aktualisierung des Dokuments gewertet.

7 Danksagung

Besonders gedankt sei Carmen Wolff, Van Ly, Kristina Eichner und Dr. Chris Wahl für ihre Hilfe bei der Erhebung der Daten.

Literaturangaben

Calishain, T.; Dornfest, R.: Google Hacks: 100 Industrial-Strength Tips & Tools. Sebastopol [u.a.], 2003

Chakrabarti, S.: Mining the web: Discovering knowledge from hypertext data. Amsterdam (u.a.): Morgan Kaufmann, 2003

Henzinger, M., Motwani, R., Silverstein, C.: Challenges in Web Search Engines. SIGIR Forum 36 (2002), http://www.acm.org/sigs/sigir/forum/F2002/henzinger.pdf [18.3.2004]

Lewandowski,, D.: Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen. IWP - Information: Wissenschaft und Praxis 55(2), 97-102 (2004)

Machill, M.; Neuberger, C.; Schweiger, W.; Wirth, W.: Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In: Machill, M.; Welp, C. (Hrsg.) (2003): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. Gütersloh: Verlag Bertelsmann Stiftung, 13-490

Ntoulas, A.; Cho, J.; Olston, C. : What's New on the Web? The Evolution of the Web from a Search Engine Perspective. Proceedings of the Thirteenth WWW Conference, New York, USA (2004). http://oak.cs.ucla.edu/~ntoulas/pubs/ntoulas_new.pdf [25.3.2004]

Price, Gary; Tyburski, Genie: It's Tough to Get a Good Date with a Search Engine. Search Day, 5.6.2002, http://www.searchenginewatch.com/searchday/article.php/2160061 [19.1.2004]

Spink, A. : Web Search : Emerging Patterns. In: Library Trends 52(2), S. 299-306 (2003)

Sullivan, D.: Search Engine Sizes (2.9.2003). http://searchenginewatch.com/reports/article.php/2156481 [8.4.2004]