Aktualität als erfolgskritischer Faktor bei Suchmaschinen

Dirk Lewandowski, Düsseldorf

Aktualitätsfaktoren spielen bei Suchmaschinen auf unterschiedlichen Ebenen eine Rolle: Einerseits geht es um die Aktualität der Datenbestände (Index-Aktualität), andererseits um Aktualitätsfaktoren in der Recherche, und schließlich um Aktualitätsfaktoren im Ranking. Zu allen drei Bereichen werden die wesentlichen Arbeiten diskutiert. Der Artikel gibt damit einen Überblick über die Vielschichtigkeit des Aktualitätsthemas und zeigt, dass Aktualität für Suchmaschinen ein erfolgskritischer Faktor ist und in Zukunft noch in verstärktem Maß werden dürfte.

Up-to-dateness as a ranking factor in Web search engines

Up-to-dateness is an important factor for Web search engines in different aspects. Firstly, the web page index should be up to date, secondly, the use of date-restricted queries should be supported, and last but not least up-to-dateness is an important ranking factor. This article discusses the research findings in all three areas. The importance of the mentioned factors is stressed and it is shown that they are key for ranking web documents. Their importance will even increase with the age of the WWW.

1 Einleitung

Suchmaschinen unterliegen durch das sich stetig verändernde Web unterschiedlichen Aktualitätsproblematiken. An erster Stelle muss natürlich der Datenbestand aktuell gehalten werden; hier geht es neben dem Auffinden aktueller Dokumente auch um die Vermeidung von Verweisen auf nicht mehr vorhandene Dokumente, also um die Vermeidung der leidigen 404-Fehlermeldung. Es ist allerdings illusorisch, anzunehmen, dass Suchmaschinen in der Lage sein können, alle Dokumente in ihrem Datenbestand (der oft mehrere Milliarden Dokumente umfasst) in gleicher Frequenz zu aktualisieren. Webseiten werden in sehr unterschiedlichen Abständen aktualisiert, teils geschieht dies minütlich. Die Suchmaschinen müssen also auswählen, welche Seiten häufig und welche seltener besucht werden. Insbesondere bei Nachrichtenangeboten ist eine häufige Aktualisierung notwendig.

Aktualitätsfaktoren spielen aber auch bei der Recherche eine Rolle: Die Einschränkung auf aktuelle Dokumente ist eine der typischen (und dazu eine einfache) Möglichkeit, die Trefferzahl zu beschränken. Suchmaschinen haben jedoch Probleme bei der Bestimmung des tatsächlichen Erstellungs- bzw. Aktualisierungsdatums einer Datei. Deshalb ist die Funktionstüchtigkeit dieser Recherchemöglichkeit eingeschränkt.

Da die Bestimmung des exakten Datums eines Dokuments schwierig (wenn nicht gar unmöglich) ist, arbeiten die Suchmaschinen mit Indikatoren für eine relative Aktualitäts- bzw. Aktualisierungsbestimmung. Diese wird vor allem für das Ranking benötigt: Hier ist es sinnvoll, aktuelle Dokumente zu bevorzugen oder diesen zumindest einen „Bonus" mit auf den Weg zu geben, da aufgrund der von den Suchmaschinen zur Qualitätsbewertung eingesetzten linktopologischen Verfahren tendenziell ältere Dokumente bevorzugt werden. Der vorliegende Artikel wird auf die genannten Problematiken ausführlich eingehen. Nach einer Beschreibung der Dynamik des Web werden die zu den einzelnen Problembereichen vorliegenden Studien ausführlich beschrieben und die sich aus ihnen ergebenden Konsequenzen dargelegt.

2 Die Dynamik des Web

Das Web verändert sich stetig und in hoher Geschwindigkeit. Neue Websites und Dokumente kommen hinzu, andere werden verändert oder gelöscht. Suchmaschinen müssen versuchen, beim Aufbau und der Pflege ihrer Indizes mit dieser Dynamik mitzuhalten, wobei sie stets nur ein Abbild der Vergangenheit bieten können. Falsch ist dabei jedoch die Annahme, dass jede Suchmaschine ein konsistentes Abbild des Web zu einem bestimmten Zeitpunkt bietet. Da das Auffinden und Erfassen der Dokumente durch die Crawler der Suchmaschinen ein kontinuierlicher Prozess ist, verändert sich auch der Index fortlaufend und ist zu keinem Zeitpunkt der gleiche [Risvik und Michelsen, 2002].[1] Auch eine kurzfristige Reproduzierbarkeit der Ergebnisse ist aufgrund verschiedener Faktoren nicht gegeben [Bar-Ilan, 2004], weshalb sich die Suchmaschinen zwar nur eingeschränkt als verlässliche Instrumente für die Erforschung des Web eignen [Wouters, et al., 2004]. Dafür können sie dem Nutzer aber einen in möglichst großen Teilen aktuellen Datenbestand bieten.

In welch hohem Maß und mit welcher Geschwindigkeit sich das Web verändert, zeigt eine Studie von Ntoulas, Cho und Olsten aus dem Jahr 2004 [Ntoulas, et al., 2004]. Auf das gesamte Web hochgerechnet ergibt sich nach dieser Untersuchung, dass pro Woche 320 Millionen neue Seiten entstehen (wobei auch auf eine andere URL verschobene Seiten zu diesen gerechnet werden). 20 Prozent der heute vorhandenen Seiten werden in einem Jahr nicht mehr vorhanden sein, inhaltlich werden innerhalb eines Jahres 50 Prozent des Webs neu sein. Noch schneller ändert sich allerdings die Linkstruktur: innerhalb eines Jahres werden 80 Prozent aller Links neu oder verändert sein.

Diese Daten zeigen, dass es für Suchmaschinen essentiell ist, ihren Index möglichst aktuell zu halten. Ohne oder mit nur partiellen Aktualisierungen würde der Datenbestand sehr schnell veralten und damit wertlos werden.

3 Index-Aktualität

Bei Fragen der Index-Aktualität muss zwischen den verschiedenen Datenbeständen der Suchmaschinen unterschieden werden. Neben dem Hauptindex (also dem Index der Webseiten) führen viele Suchmaschinen gesonderte Indizes für Nachrichten, Produkte und Bilder. Jeder dieser Datenbestände unterliegt besonderen Anforderungen an die Aktualisierung; so wird der Bilderbestand bei den meisten Suchmaschinen nicht besonders häufig aktualisiert, während der Nachrichtenbestand eine sehr häufige Aktualisierung aufweist. Untersuchungen zur Index-Aktualität müssen also differenziert für die einzelnen Datenbestände durchgeführt werden; ihre Ergebnisse können keine Gültigkeit für alle Angebote einer Suchmaschine erheben.

1 Beim Aufbau der Suchmaschinen-Indizes ist zwischen batch und incremental crawlers zu unterscheiden (Risvik u. Michelsen 2002). Ein Batch-Crawler baut den Index vollständig auf und beginnt dann von neuem mit dem Aufbau eines komplett neuen Index. Der Incremental Crawler ist dagegen in der Lage, dem bestehenden Index kontinuierlich neue Dokumente (bzw. neue Versionen von bekannten Dokumenten) hinzuzufügen. Moderne Suchmaschinen arbeiten mit Incremental Crawlers.

Im Folgenden werden Untersuchungen zur Index-Aktualität der Web-Indizes der wichtigsten Suchmaschinen sowie zu den Nachrichtenbeständen der deutschen Suchmaschinen diskutiert.

3.1 Aktualität der Web-Bestände

[Notess, 2003] untersucht die Aktualität der Datenbestände von acht verschiedenen Suchmaschinen (MSN, HotBot, Google, AlltheWeb, AltaVista, Gigablast, Teoma und Wisenut) auf der Basis von sechs Suchanfragen. Für jede Anfrage werden alle in der Trefferliste angeführten URLs, die den folgenden zwei Kriterien entsprechen, ausgewertet. Erstens müssen die betreffenden Seiten täglich aktualisiert werden und zweitens muss das Datum des letzten Updates auf der Seite selbst angezeigt werden. Für jede der so ausgewählten Seiten geht deren Alter in die Untersuchung ein. In der Ergebnisdarstellung wird pro Suchmaschine sowohl das Alter der neuesten als auch der ältesten gefundenen Seite sowie ein grober Mittelwert aller Seiten aufgeführt. In der letzten Aktualisierung dieses eine Zeitlang kontinuierlich durchgeführten Tests (Mai 2003) haben die großen Suchmaschinen MSN, HotBot, Google, AlltheWeb und AltaVista jeweils Seiten indexiert, die aktuell oder nur einen Tag alt waren. Die Datenbanken der kleineren Suchmaschinen Gigablast, Teoma und Wisenut sind dagegen weit weniger aktuell. Die durchschnittliche Aktualität der einzelnen Suchmaschinen reicht von vier Wochen bis zu sieben Monaten, wobei der Durchschnitt der großen Suchmaschinen bei einem Monat liegt (Ausnahme: AltaVista mit einem Durchschnitt von etwa drei Monaten).

Die Studie hat leider einige gravierende Probleme, wobei das schwerwiegendste wohl die unvollständige Darlegung der Methodik ist. Weder wird beschrieben, wie die Suchanfragen ausgewählt wurden noch wie die groben Durchschnittswerte ermittelt wurden. Weiterhin wird nicht erläutert, wie das Indexierungsdatum ermittelt wurde. Nicht alle untersuchten Suchmaschinen bieten eine Cache-Funktion mit Anzeige des Indexierungsdatums. Da insgesamt nur sechs Suchanfragen verwendet wurden, wurden je Suchmaschine nur zwischen zehn und 46 Treffer ausgewertet. Diese Menge ist aber schlicht zu klein, um mehr als explorative Ergebnisse zu erbringen.

Da die Untersuchung zwischen 2001 und 2003 mehrmals durchgeführt wurde, kann aus ihr zumindest ein Trend erkannt werden: Die Suchmaschinen haben sich dahingehend verbessert, dass mehr aktuelle Seiten zeitnah in den Index aufgenommen werden; die Durchschnittswerte konnten damit aber nicht verbessert werden. Alle Suchmaschinen haben veraltete Seiten in ihrem Bestand.

Die nur bedingt aussagekräftigen Ergebnisse von Notess sowie die Tatsache, dass diese mittlerweile (auch durch die seit 2003 wesentlich veränderte Suchmaschinenlandschaft) veraltet sind, war Anlass, eine umfassende Studie zur Aktualität der wichtigsten internationalen Suchmaschinen (Google, Yahoo, MSN) durchzuführen [Lewandowski, et al., 2006].

Mit der Untersuchung sollten die folgenden Fragen beantwortet werden:

- Wie häufig aktualisieren die Suchmaschinen ihre Datenbestände? Gibt es klare Intervalle?
- Welche Unterschiede gibt es in der Aktualität der Datenbestände der unterschiedlichen Suchmaschinen?
- Verfolgen die Suchmaschinen unterschiedliche „Update-Strategien"?

Zur Beantwortung dieser Fragen wurden über einen Zeitraum von sechs Wochen (Februar/März 2005) 38 deutschsprachige Webseiten, die täglich aktualisiert werden, untersucht. Neben großen Nachrichtenseiten wurden regionale Newsportale, wissenschaftsorientierte Seiten und Special-Interest-Seiten ausgewählt.

Die Seiten wurden täglich auf ihre Aktualität in den verschiedenen Suchmaschinen-Indizes hin untersucht. Dazu wurde die Cache-Funktion der Suchmaschinen verwendet, die den Stand der jeweiligen Seite zum Zeitpunkt der Indexierung anzeigt. Bei Google und MSN wird oberhalb dieser Cache-Version auch das Datum der Indexierung angezeigt. Da dies aber bei Yahoo nicht der Fall ist, konnten nur Seiten verwendet werden, die auch innerhalb ihres Inhaltsteils das Aktualisierungsdatum (oder generell das aktuelle Datum) anzeigen.

Die Untersuchung zeigt, dass sich die Suchmaschinen deutlich unterscheiden. Google hat bei weitem am meisten Seiten tagesaktuell indexiert; von insgesamt 1.558 Datenpunkten je Suchmaschine sind bei Google 1.291 (82,86 Prozent) tagesaktuell oder höchstens einen Tag alt.[2] Die beiden anderen Suchmaschinen liegen deutlich dahinter mit 748 (48,01 Prozent) für MSN und 652 (41,85 Prozent) bei Yahoo.

Allerdings darf der Anteil der täglich aktualisierten Seiten nicht als alleiniges Qualitätsmerkmal der Indexaktualität angesehen werden. Betrachtet man die Mittelwerte, so kommt Google auf eine durchschnittliche Aktualisierungsfrequenz von 3,1 Tagen, MSN auf 3,5 Tage und Yahoo auf 9,8 Tage. Damit schwindet der Unterschied zwischen Google und MSN deutlich.

Der Unterschied liegt darin, dass Google zwar am meisten top-aktuelle Seiten aufweisen kann, jedoch auch einige deutliche Ausreißer zeigt, die lange Zeit gar nicht aktualisiert werden. Abbildung 1 zeigt die Verteilungen der unterschiedlichen Aktualitätswerte bei den verschiedenen Suchmaschinen. Das Schaubild für Google zeigt eine steil abfallende Kurve; viele Seiten werden täglich aktualisiert, während einige in Intervallen von zwei bis etwa zwölf Tagen aktualisiert werden. Einige Seiten werden noch seltener besucht; die älteste gefundene Seite war 54 Tage alt. Bei MSN ergibt sich ein ganz anderes Bild: die Kurve fällt zwar auch relativ schnell ab. Es gibt deutlich weniger Seiten, die täglich aktualisiert werden, viele werden im Intervall von drei Tagen bis zwei Wochen aktualisiert. Bemerkenswert ist, dass die älteste gefundene Seite nur 17 Tage alt ist. MSN scheint also in der Lage zu sein, seinen Index schneller komplett zu aktualisieren als die untersuchten Konkurrenten. Dies kann aber auf der Basis der täglich aktualisierten Seiten nur vermutet werden; das Aktualisierungsverhalten bezüglich weniger häufig aktualisierter Seiten ist nicht bekannt.

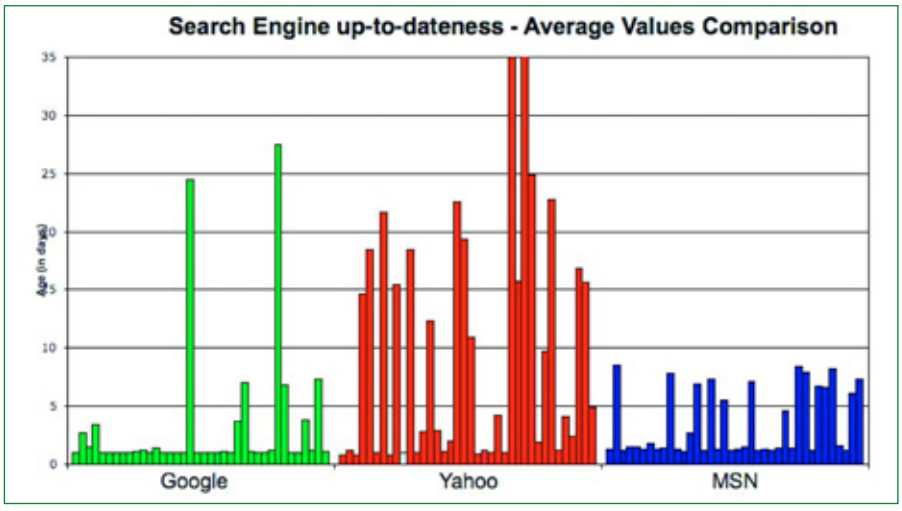

Abbildung 1: Verteilung der unterschiedlichen Aktualitätswerte bei den verschiedenen Suchmaschinen [Lewandowski et al., 2006]

2 Zwischen tagesaktuell und einem Tag alt konnte in der Untersuchung nicht unterschieden werden. Eine solche Unterscheidung hätte u.U. eine einzelne Suchmaschine benachteiligt, wenn diese ihren Index zwar täglich aktualisiert hätte, dies aber jeweils nach dem täglichen Untersuchungszeitpunkt (am frühen Abend) getan hätte.

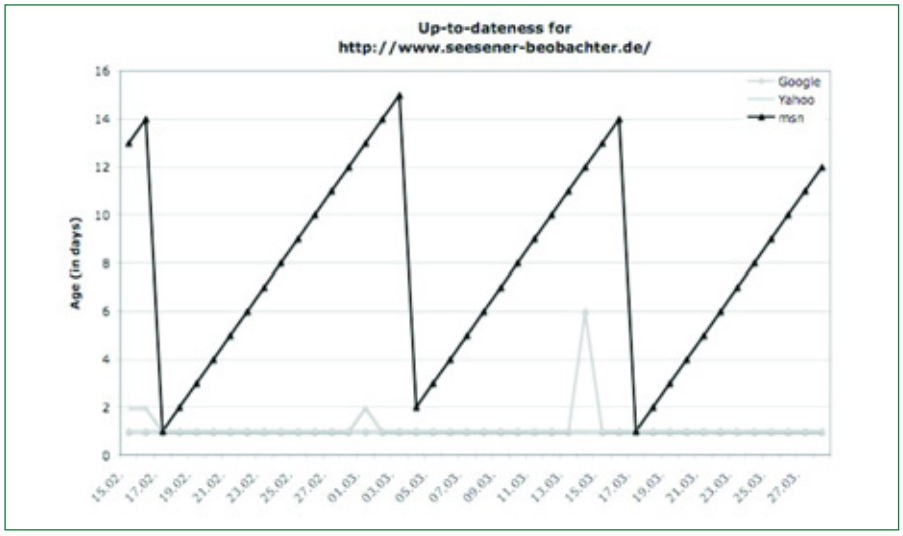

Abbildung 2: Aktualisierungsfrequenz bei MSN am Beispiel der Website des Seesenheimer Beobachters [Lewandowski et al., 2006]

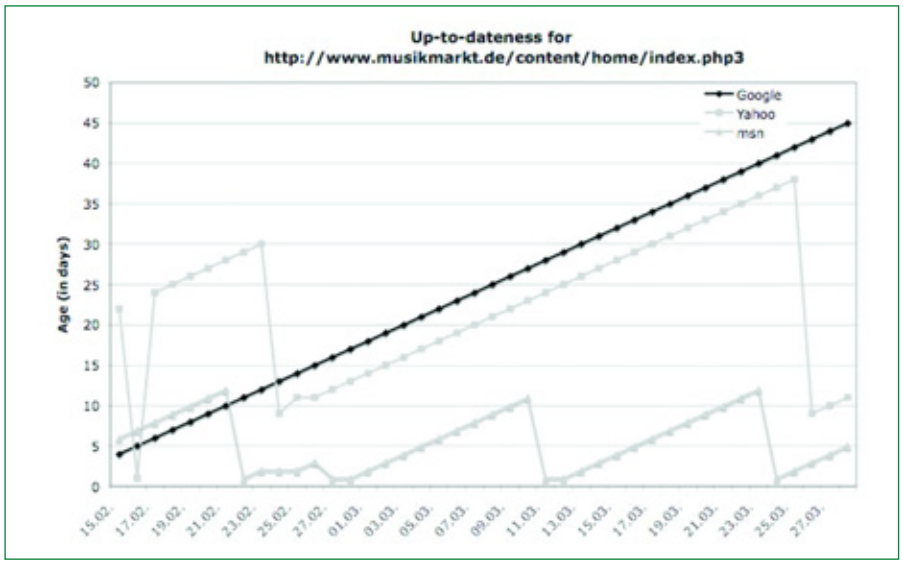

Abbildung 3: „Vergessene" Seite bei Google [Lewandowski et al., 2006]

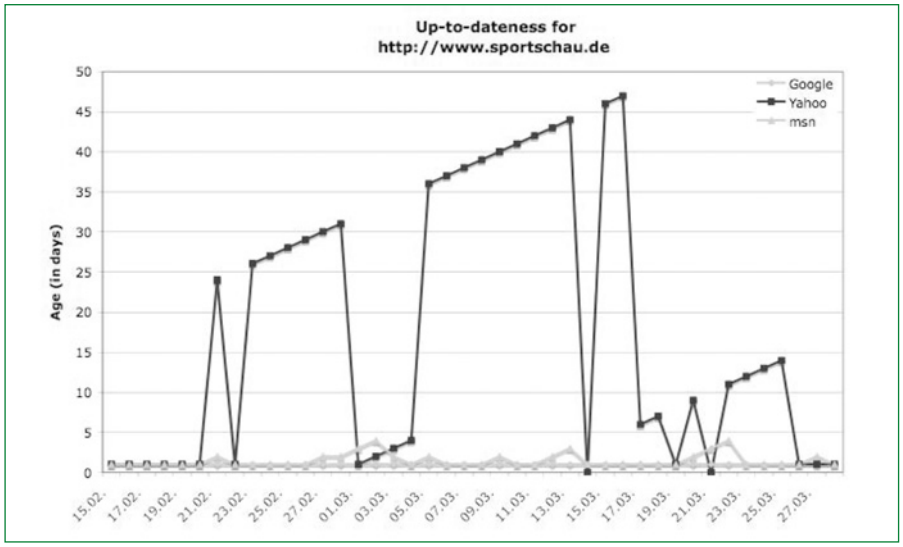

Abbildung 4: „Pattern Break" bei Yahoo [Lewandowski et al., 2006]

Als dritte Suchmaschine bietet Yahoo ein schlechtes Bild: Noch weniger Seiten als bei MSN werden täglich aktualisiert, viele Seiten weit seltener. Die Kurve hat einen langen Verlauf bis zur ältesten Seite, die bei Yahoo 62 Tage alt ist.

Aus der Untersuchung geht also Google einerseits als klarer Sieger hervor, wenn man bewertet, welche Suchmaschine am meisten Seiten tagesaktuell hält. Wenn man allerdings den Aktualisierungszyklus des gesamten Index (der tagesaktuellen Seiten) betrachtet, so schneidet MSN am besten ab. Yahoo hat offensichtlich massive Probleme, seinen Index aktuell zu halten. Insgesamt ist festzustellen, dass es keiner der untersuchten Suchmaschinen gelingt, alle tagesaktuellen Seiten zeitnah zu erfassen. Bei der Recherche kann also nicht davon ausgegangen werden, dass der Datenbestand auf dem neuesten Stand ist. Bei zeitkritischen Anfragen ist es vielmehr sinnvoll, relevante Websites direkt anzusteuern und eventuell die dortige Suchfunktion zu verwenden.

Interessant ist auch ein Blick auf die Indexierungsmuster der einzelnen Suchmaschinen, welche anhand einiger typischer Beispiele erläutert werden sollen.

Abbildung 2 zeigt hervorgehoben für MSN ein Indexierungsmuster, wie es eigentlich von allen Suchmaschinen zu erwartet gewesen wäre. Die Suchmaschine hat offensichtlich ein Intervall gefunden, in dem die Seite aktualisiert wird (bzw. im Datenbestand aktualisiert werden soll). Der Robot kehrt in regelmäßigen Abständen zu der Seite zurück und indexiert diese. Ein solch klares Indexierungsverhalten konnte einzig für MSN festgestellt werden.

Während Google bei den meisten Seiten gut abschneidet, finden sich bei dieser Suchmaschine auch Seiten, die in der Aktualisierung offensichtlich „vergessen" werden (Abb. 3). Typisch für Yahoo (allerdings teils auch bei Google zu finden) sind „Pattern Breaks", also Aktualisierungen, die schon einen oder wenige Tage später nicht mehr zu sehen sind (Abb. 4). Stattdessen wird wieder auf eine alte Version derselben Seite zurückgegriffen. Dass es sich bei diesem Verhalten nicht um einen temporären Fehler bei Yahoo handelt, ließ sich einerseits im Pretest der Studie und andererseits in einer weiteren Überprüfung im November 2005 festgestellen.

Wie bereits erwähnt, haben die Ergebnisse nur Aussagekraft für die täglich aktualisierten Seiten. Hinsichtlich „beständigerer" Seiten sind weitere Forschungen notwendig; es ist davon auszugehen, dass die Suchmaschinen noch weit ältere Seiten in ihren Indizes haben. Interessant wäre zu wissen, ob MSN hier eine Ausnahme bildet und in der Lage ist, tatsächlich alle Seiten im Index in einem Intervall von unter 20 Tagen zu besuchen.

3.2 Aktualität der News-Bestände

Die aus der Problematik der Index-Aktualität erwachsenden Schwierigkeiten wurden von den Suchmaschinen (wenigstens zum Teil) pragmatisch gelöst, indem gesonderte Datenbestände mit Nachrichten aufgebaut und mit entsprechenden Suchoberflächen versehen wurden. Besonders deutlich wurde die Schwäche der konventionellen Suchmaschinen, mit aktuellen Er-

eignissen mitzuhalten, nach dem 11. September 2001. Als hilflose Versuche, den Suchmaschinen-Nutzern aktuelle Nachrichten zu bieten, sind manuelle Verweise von den Suchseiten auf die Seiten von Nachrichtensendern und Zeitungen zu werten (vgl. [Wiggins, 2001]). Zu dieser Zeit bestanden gesonderte Nachrichten-Suchmaschinen noch nicht wie heute als integraler Bestandteil aller größeren Suchmaschinen. Diesen wiederum war die schnelle Integration der aktuellen Meldungen in ihre Indizes nicht möglich. In der Folge wurden von allen wichtigen Suchmaschinen eigene Nachrichtenbestände aufgebaut [Machill, et al., 2005]. Die Verwendung eines gesonderten Index ermöglicht es dabei, auf diesen andere Regeln anzuwenden als auf den regulären Web-Index: So werden die Nachrichten-Sites wesentlich öfter von den Suchmaschinen besucht als andere Sites im Index; die Erschließung kann wesentlich genauer erfolgen, weil der in der Regel einheitliche Aufbau aller Unterseiten eines Webangebots ausgenutzt werden kann. Probleme mit Spam bestehen nicht, weil die zu durchsuchenden Sites manuell ausgewählt werden und daher nur vertrauenswürdige Quellen Verwendung finden.
Die Problematik der Datumsermittlung bei aktuellen Nachrichten ist weniger schwierig als bei regulären Web-Dokumenten, hat jedoch selbst einige Besonderheiten, die von den Suchmaschinen berücksichtigt werden müssen. Die zu unterscheidenden Datumsangaben sind hier das redaktionelle Datum, das technische Datum, das Datum in den Metaangaben sowie das Änderungsdatum [Machill, et al., 2005]: Nachrichten werden auf Websites oft nicht in einer endgültigen Form veröffentlicht, sondern nach dem ersten Erscheinen entsprechend der aktuellen Ereignisse weiter überarbeitet.

In einem „Reaktionstest" wird in der Studie von [Machill, et al., 2005] untersucht, wie schnell die News-Suchmaschinen aktuell erschienene Nachrichten indexieren. Dazu wurden neun deutsche Suchmaschinen eine Stunde, drei Stunden und fünf Stunden, nachdem eine Nachricht über den Ticker der Nachrichtenagentur Reuters lief, abgefragt. Allerdings wurde nur eine Anfrage („Hubschrauber Absturz im Irak") verwendet, was die Ergebnisse auf explorative Aussagen beschränkt.
Die besten Reaktionszeiten mit nur einer Stunde erreichten Google, Yahoo und AltaVista. Alle anderen Suchmaschinen (mit Ausnahme von T-Online) konnten die Nachricht nach drei Stunden anzeigen. Yahoo, AltaVista und Google führten bei der Trefferzahl sowie der Relevanz der Treffer.

Neben der Limitierung durch die Verwendung nur einer Suchanfrage ist bei der Interpretation der Ergebnisse zu beachten, dass der Zeitpunkt der Untersuchung schon etwa zwei Jahre zurückliegt und sich gerade bei den Nachrichtensuchmaschinen seitdem deutliche Veränderungen ergeben haben. Weitere Forschung ist auch hier nötig.

4 Aktualitätsbeschränkung in der Recherche

Außer der Problematik der Index-Aktualität ist die Aktualitätsfrage auch für die erfolgreiche Durchführung bestimmter Recherchen kritisch. Zwar sind die meisten Nutzer nicht willens oder in der Lage, Suchanfragen zu formulieren, die in ihrer Komplexität über die typischen Zwei- bis Dreiwortanfragen hinausgehen [Machill, et al., 2003;Spink und Jansen, 2004]. Gerade aber in Hinblick auf die Aussage, dass „die meisten Nutzer nicht willens [sind], bei der Formulierung ihres Suchziels allzu viel kognitive und zeitliche Energie aufzuwenden" [Machill, et al., 2003], S. 169, dürfte die Datumsbeschränkung – weil intuitiv verständlich – eine attraktive Möglichkeit der Anfragenbeschränkung bieten.
In einer weiteren Untersuchung [Lewandowski, 2004] konnte allerdings gezeigt werden, dass die Suchmaschinen nicht in der Lage sind, das tatsächliche Erstellungs- bzw. Aktualisierungsdatum einer Seite zuverlässig zu bestimmen.
Dazu wurden 50 Suchanfragen an die Suchmaschinen Google, Teoma und Yahoo gestellt; einmal ohne Datumsbeschränkung, das andere Mal mit einer Beschränkung auf Dokumente, die in den letzten sechs Monaten aktualisiert wurden.
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.
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. 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.
Bei dieser Untersuchung ist zu beachten, dass MSN als eigenständige Suchmaschine zum Untersuchungszeitpunkt noch nicht bestand und deshalb nicht berücksichtigt wurde. Aufgrund der veränderten Suchmaschinenlandschaft wäre eine Folgeuntersuchung erforderlich.
Das insgesamt schlechte Abschneiden aller Suchmaschinen ist damit zu erklären, dass die Suchmaschinen Dokumente zu leicht als aktualisiert betrachten. Sobald eine Aktualisierung innerhalb des Dokuments auftritt (und sei dies nur die Veränderung eines automatisch gesetzten tagesaktuellen Datums oder der Uhrzeit), wird das gesamte Dokument als aktualisiert angesehen. Dass die alleinige Prüfung der Veränderung nicht tauglich für die Datumsermittlung ist, zeigt die Studie von [Ntoulas, et al., 2004]. Sie unterscheiden zwei Kennzahlen, um festzustellen, wie stark sich Seiten verändert haben: Veränderungsfrequenz *(frequency of change)* und Veränderungsgrad

Tabelle 1: Aktualitätsquoten der untersuchten Suchmaschinen [Lewandowski, 2004], S. 310

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 (s. Tabelle 1). 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 schon in der uneingeschränkten Suche beinahe jedes zweite Dokument aus dem letzten halben Jahr.

(degree of change). Dabei stellen sie fest, dass die von den meisten Suchmaschinen beachtete Veränderungsfrequenz kein guter Indikator für den Veränderungsgrad ist. Allerdings stellen die Autoren eine signifikante Übereinstimmung zwischen dem in der Vergangenheit gemessenen und dem für die Zukunft zu erwartenden Veränderungsgrad fest. Diese Korrelation variiert aber signifikant zwischen unterschiedlichen Seiten.
Erfolgversprechend erscheint eine (wenigstens näherungsweise) Ermittlung des tatsächlichen Aktualisierungsdatums durch die Kombination mehrerer Indikatoren (Serverdatum, Datum des ersten Auffindens, Metadaten, inhaltliche Auswertung) [Lewandowski, 2005], S. 180ff. Welche dieser Faktoren aktuell von den gängigen Suchmaschinen eingesetzt werden, kann aus den Ergebnissen allerdings nicht geschlossen werden.

5 Aktualitätsfaktoren im Ranking

Wenn nun die Suchmaschinen bisher offensichtlich nicht in der Lage sind, das Dokumentendatum korrekt zu bestimmen, wie können dann Aktualitätsfaktoren überhaupt im Ranking eingesetzt werden? Und warum sind solche Faktoren überhaupt notwendig?
Da die heute eingesetzten Rankingverfahren, die zu einem wesentlichen Teil auf der Auswertung der Linktopologie im Umfeld der Dokumente basieren, potenziell ältere Dokumente, die bereits gut verlinkt sind, bevorzugen [Lewandowski, 2005], S. 117-138, muss für neue Dokumente ein Ausgleichsfaktor verwendet werden, damit diese im Ranking überhaupt eine Chance haben. Manche Suchmaschinen scheinen neue Dokumente generell zu bevorzugen. Zum besseren Verständnis soll hier nochmals der Weg einer neuen Seite von der Erstellung bis zu einem potenziell hohen Ranking in den Suchmaschinen beschrieben werden.
Abbildung 5 zeigt diesen Weg schematisch anhand der beiden gegenübergestellten Rankingansätze der bevorzugten Verwendung von textstatistischen Verfahren und der bevorzugten Verwendung linktopologischer Verfahren. Die alleinige Verwendung textstatistischer Verfahren ist nicht (mehr) üblich. Selbstverständlich sind Rankingpositionen niemals statisch, allerdings ergeben sich mit der Zeit je nach Ansatz früher oder später relativ stabile Rankingpositionen.
Im Fall der Bevorzugung der textstatistischen Verfahren erreicht das Dokument schon zum Zeitpunkt seiner Erfassung durch die Suchmaschine einen hohen

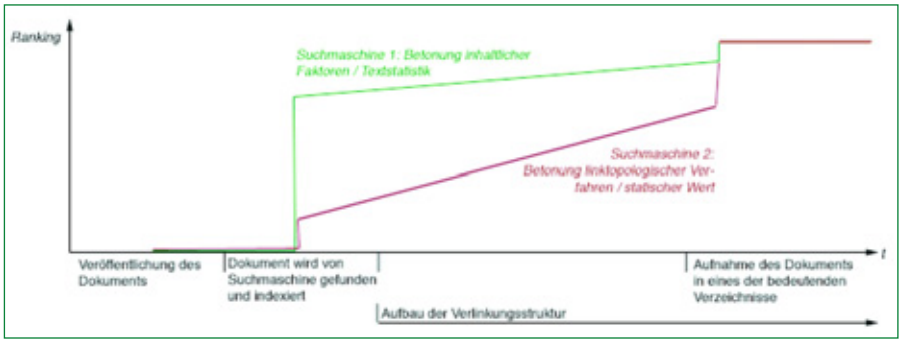

Abbildung 5: Weg eines neuen Dokuments von der Veröffentlichung bis zu einer stabilen Rankingposition [Lewandowski 2005], S. 183

Rang, weil das Ranking hauptsächlich auf inhaltlichen Aspekten basiert, die ja schon bei der Veröffentlichung des Dokuments feststehen. Im weiteren Zeitverlauf wird das Dokument höher bewertet, wenn es Links auf sich ziehen kann. Sein Ranking verbessert sich mit der zunehmenden Anzahl von Links stetig. Die im Schaubild dargestellte Aufnahme in ein Web-Verzeichnis (und die damit verbundene starke Erhöhung der Linkpopularität des Dokuments) bedeutet nochmals einen Popularitätsschub und damit ein weiter verbessertes Ranking.
Der Weg des gleichen Dokuments verläuft in einer Suchmaschine, die sich stärker auf die Linktopologie ausrichtet, anders. Hier wird das Dokument zum gleichen Zeitpunkt in den Index aufgenommen, allerdings erreicht es zuerst nur eine relativ niedrige Rankingposition. Erst mit der zunehmenden Verlinkung des Dokuments steigt seine Rangposition. Im Schaubild ist dies als kontinuierlicher Prozess dargestellt, unter realen Bedingungen wird sich diese Steigerung allerdings kaum linear darstellen. Mit der Aufnahme des Dokuments in ein Verzeichnis schließlich erreicht das Dokument seine stabile Rankingposition; im Schaubild erreicht das Dokument nun in beiden Suchmaschinen die gleiche Position.

Es dauert im Fall der linktopologisch orientierten Suchmaschine wesentlich länger, bis eine hohe Position erreicht werden kann. Bedenkt man dazu noch die Tendenzen zum *preferential attachment*[3], so ergeben sich bei einem solchen Ranking gravierende Nachteile für neue Dokumente. Deshalb ist ein Ausgleichsfaktor nötig, der den neuen Dokumenten gleiche oder zumindest ähnliche Chancen im Ranking einräumt.
Weiterhin ist zu bedenken, dass die Aktualität bei der Masse der im Web vorhandenen Dokumente eine große Rolle für die Relevanz haben kann. Wenn statische Dokumente mit Dokumenten unterschiedlicher Aktualisierungsfrequenzen um die Rangpositionen konkurrieren, ist zu entscheiden, welche zu bevorzugen sind. In vielen Fällen mag eindeutig das aktuelle Dokument zu bevorzugen sein (wie etwa bei Nachrichten), in anderen Fällen (wie wissenschaftlichen Papers) dürfte die Autorität vor der Aktualität wichtig sein. Um die Bedeutung für die einzelnen Dokumente zu bestimmen, können wiederum Aktualitätsfaktoren eingesetzt werden, nun jedoch beispielsweise bei der Verlinkung des entsprechenden Dokuments. Bei einem wissenschaftlichen Beitrag dürfte die Zahl der darauf zeigenden Links nach einer kurzen Phase des schnellen Zuwachses eher kontinuierlich steigen, während bei Nachrichten von einer schnellen Anhäufung von Links auszugehen ist, während später kaum noch neue hinzukommen oder die Gesamtzahl der verweisenden Links sogar abnimmt.
[Acharya, et al., 2005] beschreiben in ihrer Patentanmeldung unterschiedliche Aktualitätsfaktoren, die im Ranking verwendet werden können. Dabei wird davon ausgegangen, dass jedem Dokument ein statischer Aktualitätswert (ähnlich dem statischen Wert der Linkpopularität) zugewiesen wird. Die Möglichkeiten der Verwendung der Aktualitätsfaktoren teilen sich in die folgenden Gruppen:

Datum der Dokumenterstellung *(Document Inception Date)*. Acharya et al. nennen die Möglichkeiten des ersten Auffindens des Dokuments durch die Suchmaschine in Form einer Anmeldung, in Form des Auffindens im Crawling oder in Form des Auffindens eines Links, der auf das Dokument verweist.
Inhaltliche Aktualisierungen bzw. Veränderungen *(Content Updates/Changes)*. Hier sollen Dokumente, die häufig aktualisiert werden, anders bewertet werden als solche, die nicht oder nur selten aktualisiert werden. Dazu wird (ähnlich wie in der Untersuchung von [Ntoulas, et al., 2004]) zwischen der Update-Frequenz *(update frequency)* und dem Update-Grad *(update amount)* berücksichtigt.

Analyse der Abfragen *(query analysis)*. Hier wird das Nutzerverhalten ausgewertet, um solche Dokumente zu bevorzugen, die von den Nutzern entweder häufig angeklickt werden oder, was im Kontext hier von größerer Bedeutung ist, in einer gewissen Zeitspanne (beispielsweise innerhalb des letzten Monats) wesentlich häufiger angeklickt wurden als in einem vergleichbaren vorangegangenen Zeitraum. So kann ge-

3 Links werden bevorzugt auf Dokumente gesetzt, die bereits gut durch Suchmaschinen gefunden werden bzw. die eine hohe Wahrscheinlichkeit haben, von einem Nutzer angesehen zu werden.

messen werden, welche Dokumente an Popularität gewinnen bzw. verlieren.

Veränderungen in der Verlinkung *(Link-Based Criteria)*. Sowohl das Auftauchen von neuen Links als auch das Verschwinden bestehender Links kann ausgewertet werden, um festzustellen, welche Dokumente wohl aktuellere Inhalte haben und welche veraltet sind. Im letzteren Fall ist anzunehmen, dass die Zahl der Links mit der Zeit abnimmt, während die Zahl der Links bei aktuellen Dokumenten in der Anfangsphase erst einmal zunimmt

Ankertext *(Anchor Text)*. Ergeben sich in den von der Suchmaschine erfassten Ankertexten, die auf ein Dokument oder eine Domain verweisen, wesentliche Änderungen, so kann davon ausgegangen werden, dass sich die Inhalte des Zieldokuments bzw. der Zieldomain verändert haben. Beispielsweise kann die Domain verkauft worden und die Inhalte entsprechend ersetzt worden sein. Werden von der Suchmaschine statische Werte der Linkpopularität eingesetzt, ergibt sich oft das Problem, dass Domains bevorzugt gerankt werden, deren Inhalte mit denen zum Zeitpunkt der Linksetzungen nichts mehr gemein haben. Acharya et al. schlagen vor, den Zeitpunkt der Änderung der Inhalte zu ermitteln und entsprechend alle Links, die vor diesem Zeitpunkt gesetzt wurden, bei der Berechnung der Linkpopularität auszuschließen.

Traffic *(traffic)*. Wird der Traffic, der auf ein Dokument gelenkt wird, beobachtet, so kann ermittelt werden, ob dieses Dokument mit der Zeit weniger populär wird. Manche Dokumente werden in unterschiedlichen Jahreszeiten unterschiedlich häufig nachgefragt. Werden diese Gesetzmäßigkeiten im Traffic erfasst, können die Dokumente entsprechend gerankt werden.

Nutzerverhalten *(User Behavior)*. Das Nutzerverhalten kann ausgewertet werden, indem die durchschnittliche Verweildauer eines Nutzers bei einem Dokument gemessen wird. Nimmt die Verweildauer im Lauf der Zeit deutlich ab, so kann darauf geschlossen werden, dass das Dokument nun nicht mehr aktuell ist und deshalb auch nicht mehr bevorzugt gerankt werden sollte.

Informationen über die Domain *(Domain-Related Information)*. Informationen über die Domain, auf der ein Dokument liegt, können berücksichtigt werden, um die Verlässlichkeit der Dokumente zu bestimmen. So können häufige Veränderungen des Domaininhabers oder des Host als Indikator dafür dienen, dass die entsprechende Domain nur vorübergehend genutzt wird, etwa um ein Angebot aufzubauen, das künstlich Verlinkungsstrukturen generiert, um anderen Dokumenten zu einer bevorzugten Position im Ranking zu verhelfen.

Ranking im Lauf der Zeit *(Ranking History)*. Die Daten, wie ein Dokument für bestimmte Suchanfragen im Lauf der Zeit gerankt wird, können ausgewertet werden. Dabei kann eine plötzliche signifikante Verbesserung des Ranking darauf hindeuten, dass das Ranking manipuliert wurde. Allerdings kann es sich auch schlicht um ein heißes Thema handeln, durch das das Dokument entsprechend besser verlinkt oder genutzt wird. Acharya et al. schlagen einen Abgleich beispielsweise mit seriösen News-Quellen vor: Sie nehmen als wahrscheinlich an, dass echte heiße Themen auch in den Nachrichten erwähnt werden. Weiterhin soll eine Beschränkung im Maß der Steigerungsmöglichkeit im Ranking eingeführt werden, um massive Verbesserungen im Ranking, die in der natürlichen Entwicklung nur selten vorkommen, zu verhindern.

Durch die Nutzer generierte Daten *(User Maintained/Generated Data)*. Durch die Auswertung der Bookmarks, des Browser-Caches oder der Cookies eines Nutzers sollen Trends festgestellt werden. Faktoren dabei können unter anderem sein, wie oft ein Dokument sich in den Bookmarks von Nutzern findet, wie oft dieses aus den Bookmarks aufgerufen wird, wie oft ein Dokument aus den Bookmarks gelöscht wird.

Einzelne Wörter, Wortpaare, Phrasen im Ankertext *(Unique Words, Bigrams, Phrases in Anchor Text)*. Ankertexte werden oft in Massen einheitlich generiert, um das Ranking des Zieldokuments für die in den Ankertexten vorkommenden Begriffe zu verbessern. Häufen sich plötzlich gleiche Ankertexte oder es können verdächtige Texte herausgefunden werden, so kann das Zieldokument entsprechend schlechter bewertet werden.

Verlinkungsstruktur *(Linkage of Independent Peers)*. Wenn plötzlich viele Dokumente auf ein Dokument verweisen (also ein künstlicher Web-Graph erzeugt wird), so kann daraus geschlossen werden, dass es sich um einen Spamming-Versuch handelt.

Themen *(Document topics)*. Wenn die Dokumente (zumindest groben) Themen zugeordnet werden, so lässt sich bei einer Veränderung des Themas feststellen, dass eine Neubewertung des Dokuments vorgenommen werden sollte.
In der umfangreichen Darstellung von Acharya et al. zeigt sich also, dass Aktualitätsfaktoren in vielfacher Weise eingesetzt werden können, um das Ranking der Ergebnisse zu verbessern. In vielfacher Weise werden hier Aktualitätsfaktoren als Maße zur Bestimmung der Qualität der Dokumente vorgeschlagen: Die zugrunde liegende Annahme ist hier, dass aus der Dynamik des Web (in allen ihren Formen) Aussagen über die Relevanz eines Dokuments zu einem bestimmten Zeitpunkt getroffen werden können. Zumindest einige der beschriebenen Kriterien dürften bei den gängigen Web-Suchmaschinen zum Einsatz kommen.

6 Fazit und Ausblick

Aktualitätsfaktoren spielen bei Suchmaschinen in unterschiedlicher Weise eine bedeutende Rolle. In der Betrachtung fällt zunächst die Index-Aktualität ins Auge: Suchmaschinen müssen ihre Datenbestände möglichst aktuell halten, um mit der Dynamik des Web

wenigstens annähernd Schritt halten zu können. Um bei den riesigen Datenbeständen ressourcenorientiert zu arbeiten, müssen sie erkennen, welche Dokumente häufig und welche selten aktualisiert werden. Bei der Bestimmung des tatsächlichen Erstellungs- bzw. Aktualisierungsdatums der Dokumente ergeben sich aber massive Probleme, wie sich auch bei der Recherche nach aktuellen Dokumenten zeigt.

Schließlich sind Aktualitätsfaktoren im Ranking zu berücksichtigen. Die Aktualität eines Dokumentes kann ein entscheidendes Qualitätsmerkmal sein: In vielen Fällen ist davon auszugehen, dass der Nutzer für seine Anfrage aktuelle Dokumente bevorzugt. Insofern sind hier Aktualitätsfaktoren als Ausgleichsfaktoren zu den Bewertungen der linktopologischen Verfahren zu sehen, die tendenziell ältere Dokumente bevorzugen. Letztlich kann in der Verwendung von Aktualitätsfaktoren auch die Erweiterung der verwendeten Qualitätsfaktoren (wie Linkpopularität, Klickpopularität) gesehen werden. Insgesamt dürften Aktualitätsfaktoren (auch in Hinblick auf das weiter wachsende Web und dem damit einhergehenden höheren Anteil an älteren Dokumenten) an Bedeutung in der Relevanzbestimmung gewinnen.

Literatur

Acharya, Anurag; *Cutts*, Matt; *Dean*, Jeffrey; *Haahr*, Paul; *Henzinger*, Monika; *Hoelzle*, Urs; *Lawrence*, Steve; *Pfleger*, Karl; *Sercinoglu*, Olcan und *Tong*, Simon (2005): Information retrieval based on historical data US-Patent 20050071741, March 31

Bar-Ilan, Judit (2004): Search Engine Ability to Cope With the Changing Web. In: Mark Levene und Alexandra Poulovassilis: Web Dynamics: Adapting to Change in Content, Size, Topology and Use, S. 195-215. Heidelberg: Springer Verlag

Lewandowski, Dirk (2004): Datumsbeschränkung bei WWW-Suchanfragen: Eine Untersuchung der Möglichkeiten der zeitlichen Einschränkung von Suchanfragen in den Suchmaschinen Google, Teoma und Yahoo. In: Bernard Bekavac, Josef Herget, Marc Rittberger (Eds.): Information zwischen Kultur und Marktwirtschaft, Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004), Chur, 6.-8. Oktober 2004. Schriften zur Informationswissenschaft 42 Hochschulverband für Informationswissenschaft S. 301-316

Lewandowski, Dirk (2005): Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt am Main: DGI.

Lewandowski, Dirk; *Wahlig*, Henry und *Meyer-Bautor*, Gunnar (2006): The Freshness of Web search engine databases. In: Journal of Information Science (Band 32), Nr. 2, S. 133-150.

Machill, Marcel; *Lewandowski*, Dirk und *Karzauninkat*, Stefan (2005): Journalistische Aktualität im Internet. Ein Experiment mit den „News-Suchfunktionen" von Suchmaschinen. In: Marcel Machill und Norbert Schneider: Suchmaschinen: Herausforderungen für die Medienpolitik, S. 105-164. Berlin: Vistas.

Machill, Marcel; *Neuberger*, Christoph; *Schweiger*, Wolfgang und *Wirth*, Werner (2003): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In: Marcel Machill und Carsten Welp: Wegweiser im Netz. Gütersloh: Bertelsmann Stiftung.

Notess, Greg R. (2003): Search Engine Statistics: Freshness Showdown, 4.1.2005, www.searchengineshowdown.com/stats/freshness.shtml

Ntoulas, Alexandros; *Cho*, Junghoo und *Olston*, Christopher (2004): What's New on the Web? The Evolution of the Web from a Search Engine Perspective, Thirteenth WWW Conference, New York, USA.

Risvik, Knut Magne und *Michelsen*, Rolf (2002): Search engines and Web dynamics. In: Computer Networks (Band 39), Nr. 3, S. 289-302.

Spink, Amanda und *Jansen*, Bernard J. (2004): Web Search: Public Searching of the Web. Dordrecht: Kluwer. (Information Science and Knowledge Management; Bd.6).

Wiggins, R.W. (2001): The Effects of September 11 on the Leading Search Engine. In: First Monday (Band 7), Nr. 10, www.firstmonday.org/issues/issue6_10/wiggins/

Wouters, Paul; *Hellsten*, Iina und *Leydesdorff*, Loet (2004): Internet time and the reliability of search engines. In: First Monday (Band 9), Nr. 10, http://firstmonday.org/issues/issue9_10/wouters/index.html

Suchmaschine, Zeitfaktor, Qualität, Bewertung

Der Autor

Dr. Dirk Lewandowski

ist freier Berater zum Themenbereich Suchmaschinen und Suchtechnologie sowie Lehrbeauftragter an der Heinrich-Heine-Universität Düsseldorf. Seine Forschungsinteressen sind Web Information Retrieval, Qualitätsfaktoren von Suchmaschinen sowie das Rechercheverhalten der Suchmaschinen-Nutzer. Weitere Informationen sind auf seiner Website zu finden.

Heinrich-Heine-Universität Düsseldorf
Institut für Sprache und Information
Abt. Informationswissenschaft
Gebäude 23.21
Universitätsstraße 1
40225 Düsseldorf
E-Mail:
dirk.lewandowski@uni-duesseldorf.de
www.durchdenken.de/lewandowski

Informationen

Ausschreibung des VFI-Förderungspreises 2006

Erstmals schreibt der österreichische Verein zur Förderung der Informationswissenschaft (VFI) für den gesamten deutschsprachigen Raum einen Förderungspreis für herausragende akademische Abschlussarbeiten auf Teilgebieten der Informationswissenschaft aus. Es können bis zu drei Preise vergeben werden. Das Preisgeld beträgt zwischen 300 und 500 Euro; die Preisträger werden in der Fachpresse bekannt gemacht. Bis zum 15. September 2006 können Arbeiten eingereicht werden, die 2005 oder 2006 approbiert worden sind. Neben der wissenschaftlichen Qualität spielen vor allem Originalität/Neuartigkeit des Themas, Praxisrelevanz, Relevanz für die theoretische Weiterentwicklung des gewählten Teilgebiets, Qualität und Originalität hinsichtlich Methodik und Themenbehandlung, Qualität der Präsentation und des Stils und Brauchbarkeit als Lehrtext oder Übersichtsarbeit eine Rolle.

Die Arbeiten sind, gemeinsam mit einer Approbationsbestätigung der betreffenden Hochschule, in elektronischer Form einzusenden. Die Vergabe eines Preises ist an die Vorlage einer als Zeitschriftenaufsatz publizierbaren Kurzversion gebunden. Details zu den Regelungen für den VFI-Förderungspreis sind auf der Webseite http://www.ub.tuwien.ac.at/vfi/VFI_Preis.html zu finden.

Josef *Pauser*