Institut für
Kommunikationswissenschaft
Westfälische Wilhelms-Universität Münster

Abschlussbericht:
Test der Effektivität von Optimierungsverfahren zur externen Manipulation von Suchmaschinen

Im Rahmen des Projekts „Transparenz im Netz“
im Auftrag der Bertelsmann Stiftung

Christoph Neuberger/Stefan Karzauninkat

Stand: 3. November 2003
Beobachtung, Auswertung und Bericht: Christoph Neuberger
Konzept und technische Realisierung: Stefan Karzauninkat
Mitarbeit (Codierung und Datenverwaltung): Birgit Eiglmeier, Yvonne Lünzmann, Kerstin Popp

1 Einführung

Suchmaschinen sind „Black boxes“, über deren Arbeitsweise nicht nur den Nutzern, sondern auch Experten relativ wenig bekannt ist. Welche Eigenschaften einer Webseite bestimmen ihre Beachtung und Platzierung? Diese Frage lässt sich beantworten, wenn man die Dokumenteigenschaften systematisch variiert und die Seiten unter gleichen Randbedingungen (Domains, Zeitraum, Sitestruktur, Anmeldung) Suchmaschinen im WWW „anbietet“. Zugleich lässt sich durch einen solchen Test prüfen, wie effektiv die Optimierungsverfahren sind, die zur externen Manipulation von Suchmaschinen eingesetzt werden. In einem achtmonatigen Test von neun deutschen Suchmaschinen wurden sieben verschiedene Optimierungsverfahren – teils separat, teils kombiniert – eingesetzt. Die drei Parameter zur Erfolgsmessung waren die Aufnahme einer Seite in den Suchmaschinen-Index (vgl. Abschnitt 3.1), das Ranking, also die Höhe der Platzierung in der Ergebnisliste (vgl. Abschnitt 3.2), sowie die Crawlingtiefe, also die Erfassung der Zeichenzahl und damit wie viel vom Inhalt einer Webseite im Index der Suchmaschine landet (vgl. Abschnitt 3.3). Da es eine vergleichbare Studie bisher nicht gab, musste zunächst ein Testdesign entwickelt werden (vgl. Abschnitt 2).

2 Methode und Analysetool

Domains und Sitestruktur: Bei diesem Effektivitätstest wurde mit insgesamt fünf Domains gearbeitet. Dies ermöglichte es, die Stabilität der Ergebnisse über verschiedene Domains hinweg zu prüfen und die Härte des Wettbewerbs bei verschiedenen populären Suchwörtern zu messen. Diese Domains waren:

- „www.britney-spears-fanpage.de“, optimiert für: „Britney Spears“
- „www.hotel-lupe.de“, optimiert für: „Hotel“
- „www.mallorca-freak.de“, optimiert für: „Mallorca“
- „www.spielespiel.de“, optimiert für: Spiele
- „www.versicherungsvergleicher-online.de“, optimiert für: „Versicherungen“

Tabelle 1 gibt einen Überblick über die Verzeichnisstruktur dieser Websites. Dort wurden verschiedene Spam-Methoden alleine und in verschiedenen Kombinationen realisiert (vgl. als aktuellen Überblick zu den angewandten Verfahren des Spamming: Karzauninkat 2003). Die Seiten waren alle von den jeweiligen (nicht optimierten) Startseiten („index. html“) aus verlinkt, um den Robots den Zugang zu allen Unterseiten zu ermöglichen. Die Seiten waren – bis auf jene im Ordner /l1 bis l4 – nicht querverlinkt, damit sie sich nicht gegenseitig beeinflussten. Auch die Startseiten der Linkfarmen waren mit der Homepage verlinkt.

„Keyword Stuffing“:

Für die Bestimmung geeigneter Keywords diente der „MetaGer-Assoziator“ („http://metager.de/asso.html“). Mit diesem Tool wurden zum Ausgangswort verwandte, deutschsprachige Wörter erzeugt. Ausgangsbegriff war der jeweils zu opti-

mierende und eine Variante davon. Schon bei einer leichten Variation erzeugte der Assoziator ganz unterschiedliche Begriffslisten. Aus der MetaGer-Liste wurden Eigennamen (z.B. Hotelnamen) entfernt, um möglichen rechtlichen Einwänden zuvorzukommen. Missbräuchliche Verwendung vom Markennamen auf Webseiten könnte für betroffene Firmen ein Klagegrund sein. Die Basiswörter für die Listen waren: 1. Versicherung, Versicherungen, 2. Britney Spears, Britney, 3. Spiele, Online Spiele, 4. Mallorca, Verreisen, 5. Hotel, Urlaub. Diese verwandten Wörter wurden zugemischt, um eine optimalen Keyworddichte zu erreichen. Jede Seite enthält Metainformationen mit Keywords, den Titel mit Keyword und mehrere Hervorhebungen als Überschrift (h1, h2, h3). Das Stichwort selbst wurde in einer Dichte von etwa 5% dem Text beigemischt. Der Text hat eine Länge von ca. 3000–4000 Zeichen. Die Seiten wurden nach den Empfehlungen und dem Vorbild professioneller Optimierer gebaut.[1]

Tabelle 1: Verzeichnisstruktur der Websites mit statischen Spam-Methoden und ihrer Kombination, die in den Verzeichnissen realisiert wurden

Spam Methode	Keyword Stuffing	Unsichtbarer Text	Weiterleitungs-seiten	Text in Kommentaren und Alt-Tags	Links von einer Seite mit hohem Page Rank
Kombiniert mit:					
Ohne Kombination	/1	/2	/3	/4	/6
Keyword Stuffing	–				
Unsichtbarer Text	/u1	–			
Weiterlei-tungsseiten	/w1	/w2	–		
Text in Kommentaren und Alt-Tags	/k1	/k2	/k3	–	
Interne Links	/l1	/l2	/l3	/l4	

Dynamische Methoden: Neben den statischen Spam-Methoden gibt es dynamische Verfahren. Für den Test wurden Cloaking und Linkfarmen kombiniert. Die im Test eingesetzte Cloaking-Methode beruht auf einem selbst entwickelten Script, das auf dem Prinzip eines Freeware-Scriptes basierte und stark erweitert wurde. Hierbei wurde eine eingekaufte IP-Tabelle von Fantomaster[2] verwendet, die die IP-Nummern der aktiven Suchmaschinenrobots enthält und regelmäßig aktualisiert wird. Jeder Suchmaschinenrobot nutzt (wie jeder, der im Internet kommuniziert) eine eindeutige IP-Adresse, die sich im Falle der Robots selten oder nie ändert. Entsprechend kann ein Nutzer, der als Robot identifiziert wurde, auch von anderen als solcher er-

1 Vgl. zum Beispiel das „Suchmaschinen-Tutorial“ von Klaus Schallhorn unter http://www.kso.co.uk/de/tutorial/

2 Marktführer für das „Cloaking“ von Websites.

kannt werden. Die Linkfarmen können wegen des Cloaking nicht von normalen Surfern gesehen werden. Ein spezielles Script erkennt anhand der IP-Adresse den Nutzer und liefert an diesen speziell vorbereitete Seiten aus. Besucher, deren IP Nummer nicht auf der Liste steht, bekommen ganz andere Seiten zu sehen. Die Einstiegsseiten für die Linkfarmen waren für vier Domains die Verzeichnisse: /hotel, /mallorca, /spiele und /versicherung.

Die Linkfarmen bestanden aus jeweils 200 bis 400 Seiten mit jeweils einem Haupt- und vier Nebenschlüsselwörtern. Alle Seiten innerhalb der Linkfarm einer Domain waren untereinander verlinkt. Zusätzlich war nach dem Zufallsprinzip jeweils jede zehnte Seite der Nachbardomains mit verlinkt. Die Seiten wurden bei jedem Aufruf dynamisch generiert, d.h. dass die Seiten nicht als normale statische Datei auf dem Server existierten, sondern ein php-Script erzeugte im Augenblick des Seitenaufrufes eine Datei, die für den Robot wie eine normale statische Seite aussah. Der normale Nutzer sah lediglich eine Seite mit einem animierten Bild. Die Seiten der Cloakingpages bestanden aus kopierten Textfragmenten aus dem „Projekt Gutenberg“[3] und Textfragmenten aus „Zeit“-Artikeln, also recht willkürlich gemischten deutschsprachigen Artikeln. In diese Texte waren Variablen-Felder eingebaut, die Stichwörter zumischten. Die bei den anderen Spam-Seiten gewählte Methode der Aneinanderreihung von Keywords wurde hier nicht für den normalen Text gewählt, da diese Keywords als komplette Liste noch einmal zu allen anderen Seiten der Farm verlinkten. Auf diese Weise wurde die Keyworddichte etwas gesenkt. Eine zu hohe Zahl von Keywords ohne Fülltext hätte zu einer frühzeitigen Identifizierung als Spam Seiten führen können. Zusätzlich wurde am Ende eine Liste mit den internen Links und viermal zehn zufällig ausgesuchten Links der anderen Domains angefügt.

Testverfahren: Geprüft wurde die Aufnahme in den Index und die Platzierung auf den Rängen 1 bis 20. In einem Analysetool wurde dafür die Domain und der jeweilige Begriff eingegeben. Dabei konnte je nach Auslastung der Suchmaschinen die ganze Liste auf einmal abgearbeitet werden. Wegen der hohen Beanspruchung des Servers und der langen Antwortzeiten der Suchmaschinen kam es vor, dass der Server einen „Timeout“ lieferte. In diesem Falle wurde die Liste gekürzt. Als Ergebnis erschien eine Trefferliste, in der gefundene URLs blau hervorgehoben wurden. Zusätzlich zu den gespamten Wörtern gab es Kontrollwörter, mit deren Hilfe sich feststellen ließ, ob die Seiten überhaupt im Index stehen. Die Kontrollwörter wurden aus einmaligen Zeichenketten gebildet und lauteten: britneyspearswurst, hotelwurst, mallorcawurst, spielewurst, versicherungsvergleicherwurst.

Alle Seiten wurden bei den Volltextsuchmaschinen am 02.01.2003 erstmals angemeldet.[4] Ihr Auftauchen im Index der Suchmaschinen wurde bis 28.08.2003 wöchentlich am Donnerstag geprüft. Die Trefferlisten wurden gespeichert, die Ergebnisse wurden in einer SPSS-Datenmaske erfasst.

---

3 http://www.gutenberg2000.de

4 Auch bei der durch das Page-Rank-Verfahren bekannt gewordenen Suchmaschine Google kann man direkt anmelden: http://www.google.de/intl/de/addurl.html Bei allen untersuchten Volltextsuchmaschinen außer Metager existiert eine Add-URL-Funktion.

Analysetool: Für die Beschleunigung und eine höhere Genauigkeit der Auswertung wurde ein Analysetool eingesetzt:

- Ausgangsmodul war ein Robot, der als simulierter Web-Client Metasuchmaschinen-ähnlich auf die Seiten der Suchmaschinenbetreiber zugreift.
- Es musste ein Abfragemechanismus entwickelt werden, der die gewünschten Stichworte in den Suchmaschinen übergab und gleichzeitig für den späteren Abgleich die zu vergleichende URL zwischenspeichert.
- Die Rückgabeseiten der Suchmaschinenbetreiber wurden daraufhin analysiert, ob es einen Treffer gab oder nicht. Wenn „ja", wurden die ermittelten URLs mit der zuvor angegebenen URL verglichen, um zu bestimmen, ob die gewünschte URL ebenfalls unter den Treffern war.
- Die Suchmaschinen stellen die URL in ihren Trefferseiten sehr unterschiedlich dar, teilweise in einen zusätzlichen Code integriert. Das machte die Extraktion der exakten URL aufwendiger als erwartet.
- Bei der Abfrage jeder einzelnen Suchmaschine wurden die Treffer-URLs einzeln ausgegeben und – falls identisch – mit der gesuchten URL optisch hervorgehoben.
- Die Ergebnisse wurden auf einer Seite ausgegeben und in einem Archiv protokolliert. Eine Verzeichnisseite mit allen erfolgten Abfragen wurde ebenfalls angelegt.
- Die Analysetechnik zur Erkennung der Seitenbestandteile erforderte einen hohen Aufwand bei der Pflege der Interpretationsregeln, da es immer wieder zu nicht vorhersehbaren Design- und Strukturänderungen auf Seiten der Suchmaschinenbetreiber kam.
- Eine aufwendige Erweiterung bestand darin, dass nicht nur eine einzelne Query, sondern eine Liste von Queries auf einmal mit dem Tool abgefragt werden konnte. Damit wurde eine deutliche Verringerung des Zeitaufwandes bei der Analyse erreicht.

Crawlingtiefe: In einem zweiten Test wurde die Tiefe des Suchmaschinen-Crawlings ermittelt, also wie tief „nach unten" ein Dokument erfasst wird. Dafür wurde eine Seite mit 1000 kB Textmenge produziert („index.html"). Eine weitere Seite („index2.html") hatte die identische Textmenge (1 MB), aber zu Beginn etwa 50 kB JavaScript. So konnte ermittelt werden, ob JavaScript erfasst wird oder nicht. Jede Seite enthielt im Abstand von 1 kB (ca. 1000 Zeichen) insgesamt tausend einzigartige Kontrollwörter, nach denen gesucht werden konnte. Sie wurden nach dem folgenden Muster gebildet:

- „wurst1brottwettbewerb50" meint: „Treffer nach 50 kB"
- „wurst2brottwettbewerb172" meint: „Treffer nach 172 kB"
- „wurst3brottwettbewerb577" meint: „Treffer nach 577 kB"

So konnte auf 1 kB genau ermittelt werden, welche Textmenge vom Crawler pro Seite erfasst wurde. Der verwendete Fülltext entstammte dem „Projekt Gutenberg“, das von Copyrights frei ist. Die verwendete Domain war „http://www.wurstbrotwettbewerb.de“. Auch hier konnte das Analysetool eingesetzt werden. War die Suche nach „Wurstbrotwettbewerb1000“ erfolgreich, war die gesamte Seite bis zum Ende indexiert. Andernfalls musste die Zahl am Ende schrittweise niedriger bzw. bei einem Treffer wieder höher eingegeben werden. So ließ sich schrittweise bis auf 1 kB genau die Crawlingtiefe bestimmen. Um festzustellen, ob darüber hinaus der Inhalt des JavaScript indexiert wird, wurde zusätzlich das einmalige Wort „Hamsterbackenlizenz“ im JavaScript eingebaut.

Untersuchungszeitraum und Suchmaschinen: Einbezogen waren in den Optimierungstest die Suchmaschinen AltaVista, AOL, Fireball, Google, Lycos, MSN, T-Online, Web.de und Yahoo, als die gleichen Suchmaschinen wie im Leistungsvergleich bei Machill, Neuberger, Schweiger, Wirth (2003). Lediglich MetaGer wurde hier nicht betrachtet, weil eine Metasuchmaschine von außen nicht direkt manipulierbar ist, sondern nur über den Umweg über andere Suchmaschinen. Alle Seiten wurden bei den Suchmaschinen am Donnerstag, den 02.01.2003, erstmals angemeldet. Ihr Auftauchen im Index der Suchmaschinen wurde ab 09.01.2003 wöchentlich am Donnerstag überprüft. Dabei wurden die Trefferlisten zu Dokumentationszwecken gespeichert, die Ergebnisse in einer SPSS-Datei erfasst. Laufzeit des Projekts waren acht Monate, was 34 Prüfterminen entspricht (bis 28.08.2003).

Vorauszuschicken ist, dass die Ergebnisse interne Mechanismen der Suchmaschinen widerspiegeln, die für Außenstehende kaum einsehbar sind. Deshalb lässt sich nicht bis ins Detail erklären, wie Ergebnisse zustande gekommen sind. Angesichts der geringen Fallzahlen werden hier in der Regel absolute Werte ausgewiesen. Um den zeitlichen Verlauf besser überblicken zu können, wurde der Untersuchungszeitraum für einige Auswertungen in drei etwa gleich lange Abschnitte aufgeteilt.[5]

5 Der erste Abschnitt beginnt mit der dritten Abfrage, bei der erstmals indizierte Seiten angezeigt wurden. Die drei Zeiträume umfassen die Abfragen 3-13 (n=183), 14-24 (n=290) und 25-34 (n=346).

3 Ergebnisse

3.1 Aufnahme in den Index

Im Untersuchungszeitraum wurden bei den 34 wöchentlichen Abfragen insgesamt 906 Treffer für optimierte Seiten ermittelt. Zieht man hier Mehrfachanzeigen ab, also jene Seiten, die bei einer Abfrage mehr als einmal als Treffer angezeigt wurden, verbleiben 819 Treffer.[6] Ebenfalls ausgeschlossen wurden in der Ergebnisdarstellung Seiten, die zu Linkfarmen gehörten und dem Zweck dienten, die in den Verzeichnissen /hotel, /mallorca, /spiele und /versicherung abgelegten Seiten zu optimieren (also nicht selbst Gegenstand der Optimierungsaktion waren).[7]

Erfolg der Optimierungsverfahren nach Suchmaschinen: Alle neun Suchmaschinen haben auf die manipulierten Seiten reagiert. Wie anfällig waren die einzelnen Suchmaschinen für Manipulationsversuche? Betrachtet man hier die absolute Zahl der manipulierten Seiten, die im Untersuchungszeitraum angezeigt wurde, so ergibt sich, dass Fireball am schlechtesten gegen Einflussnahmen von außen gewappnet war. Bei Fireball ließen sich 230 Treffer registrieren, was 28% aller von den neun Suchmaschinen angezeigten Seiten entspricht.[8]

Fireball besaß nicht nur einzelne Schwachstellen, sondern ließ sich als einzige der untersuchten Suchmaschinen durch jedes der getesteten Verfahren überlisten. Fireball reagierte als erste Suchmaschine auf die optimierten Seiten, nämlich schon bei der dritten Abfrage (AltaVista folgte als nächster Anbieter erst bei der siebten Abfrage) (vgl. Tabelle 2). Einen regelrechten Einbruch erlebte Fireball nach der 29. Untersuchungswoche (24.07.2003): Nachdem zuvor die Zahl der Seiten konstant bei vier gelegen hatte, machte sie nun einen Sprung auf 27 und erreichte nach 31 und 33 Wochen ein Maximum von 29 Seiten. Mehr als die Hälfte aller angezeigten Seiten

---

6 Von diesen 87 Mehrfachanzeigen entfielen auf Web.de alleine 56. Die restlichen fanden sich bei MSN (14), Fireball (15), Google (1) und AltaVista (1). Diese Dubletten wurden in der Zwischenauswertung (bis 22.05.2003) berücksichtigt, weshalb die hier dargestellten Ergebnisse leicht abweichen (besonders in Tabelle 2).

7 Solche Linkfarm-Seiten (insgesamt: 186) tauchten nach 15 Untersuchungswochen bei MSN auf, nach 23 Wochen bei AltaVista und nach 30 Wochen bei Fireball. Fireball zeigte zwar nur in fünf Wochen Linkfarm-Seiten an, erfasste aber die mit Abstand meisten Seiten (Maximum nach 34 Wochen mit 22 Seiten). Dagegen registrierte AltaVista maximal zwei Linkfarm-Seiten an; MSN kam über sieben Seiten nicht hinaus. Die durch die Linkfarm optimierten Seiten entdeckten Fireball (4 Treffer) und MSN (10), nicht aber AltaVista.

8 Um Missverständnissen vorzubeugen: Wir unterscheiden zwischen der Aufnahme in den Suchmaschinen-Index (ohne Platzierung) und dem Ranking auf den Plätzen 1-20. Da bei diesem Experiment Suchwörter ausgewählt wurden, die sehr häufig verwendet werden, erzielt man damit leicht Tausende von Treffern. Eine genaue Lokalisierung des Treffer-Platzranges auch jenseits von Rang 20 hätte einen zu großen Zeitaufwand bedeutet. Außerdem gibt es empirische Hinweise darauf, dass für die User nur die Plätze 1-20 von Interesse sind. Den Rest beachten sie kaum (vgl. Machill, Neuberger, Schweiger, Wirth 2003: 94f., 255; Greenspan 2002).

stammte ab der 29. Abfrage bis zum Ende des Untersuchungszeitraums jeweils von Fireball.

AltaVista (131 Treffer im Untersuchungszeitraum) und Google (104) waren – wenn auch deutlich weniger ausgeprägt als Fireball – ebenfalls empfänglich für optimierte Seiten. Nur zwei Suchmaschinen zeigten sich weitgehend immun, nämlich Yahoo und AOL. Sie hatten nicht nur insgesamt sehr wenige Treffer vorzuweisen (Yahoo: 11, AOL: 2), sondern reagierten auch deutlich später auf die optimierten Seiten als die anderen Anbieter (Yahoo: 24 Wochen, AOL: 27 Wochen). Im Fall von Yahoo ist dies vermutlich auf die redaktionelle Prüfung angemeldeter Seiten zurückzuführen. Bei AOL überrascht das Ergebnis deshalb, weil AOL-Treffer bei dem Leistungsvergleich von Machill, Neuberger, Schweiger, Wirth (2003: 106) weitgehend mit jenen von Google übereinstimmten, Google aber zumindest für kurze Zeit viele optimierte Seiten indizierte. Eine mögliche Erklärung ist in internen Umbaumaßnahmen bei AOL und möglicherweise unsynchronisierten Daten zu sehen. Möglicherweise nutzt AOL nur bestimmte Bereiche des Google Serverparks. Im Gegensatz dazu hat sich Web.de – der Anbieter kooperiert ebenfalls mit Google – fast im Gleichschritt mit Google verhalten und weitgehend gleiche Seiten angezeigt.

Google und Web.de reagierten erst sehr spät, dann jedoch gleich relativ stark: Beide Anbieter zeigten erst nach zehn Wochen jeweils acht Treffer an. Dies waren zu diesem Zeitpunkt deutlich mehr Treffer als bei den anderen Suchmaschinen, die schon längere Zeit einzelne Seiten erfasst hatten. So schlagartig, wie die Seiten bei Google und Web.de auftauchten, verschwanden sie auch nach 19 Wochen (Web.de) bzw. 20 Wochen (Google) aus dem Index.

Die Sorge, dass die Domains als gespamt aufgefallen waren und deshalb völlig gesperrt wurden, erwies sich als unbegründet: Nach 24 Wochen erschien dann wieder jeweils eine Seite bei Google und Web.de. Bis zum Ende des Untersuchungszeitraums wurde das Niveau des „Zwischenhochs“ von März bis Mai jedoch nicht wieder erreicht. Hier gab es also wie bei Fireball starke Schwankungen, wobei abrupte Ab- und Zunahmen den Zeitpunkt von Index-Aktualisierungen markieren. Allerdings lassen sich aus den Daten keine Aktualisierungszyklen herauslesen.[9]

Ein ähnliches Paar wie Google und Web.de bildeten Lycos und T-Online. Sie meldeten erstmals in der achten Woche Treffer und zeigten ebenfalls fast die gleichen Seiten an. Sie kamen im gesamten Zeitraum der Untersuchung bei keiner Abfrage über fünf angezeigte Seiten hinaus.

9 Starke Veränderungen lassen darauf schließen, dass eine Aktualisierung stattgefunden hat. „Zyklus“ meint, dass die Zeiträume, in denen sie stattfinden, nicht nachvollziehbar sind. Dafür gab es zu selten deutliche Veränderungen.

Tabelle 2: Zahl der indizierten Seiten pro Woche und Suchmaschine, bereinigt um Mehrfachanzeigen und Linkfarm-Seiten (absolut, leere Felder = 0)

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
AltaVista							4	4	4	4
AOL										
Fireball			2	2	3	3	3	3	3	3
Google										8
Lycos								4	4	5
MSN									1	
T-Online								4	4	5
Web.de										8
Yahoo										

	11	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
AltaVista	4	4	4	4	4	4	4	4	4	4
AOL										
Fireball	3	3	3	3	3	3	3	3	3	3
Google	8	7	8	8	8	7	8	8	12	
Lycos	4	4	4	4	4	4	4	2	2	2
MSN	5	3	6	6	5	4	4	6	6	6
T-Online	4	4	4	4	4	4	2	2	1	2
Web.de	8	6	6	8	8	7	8	6		
Yahoo										

	21	22	23	24	25	26	27	28	29	30
AltaVista	4	4	6	6	6	6	6	6	5	6
AOL							1	1		
Fireball	3	4	4	4	4	4	4	4	27	23
Google				1	1	2	4	2	2	2
Lycos	2	3	2	3	1	1	3	4	4	4
MSN	5	5	5	3	2	2	3	3	3	3
T-Online	2	1	4	2	3	3	2	4	4	4
Web.de				1	1	2	2	2	2	2
Yahoo				1	1	2	2	2	1	

	31	32	33	34	Gesamt
AltaVista	5	5	5	5	131
AOL					2
Fireball	29	20	29	19	230
Google	2	2	2	2	104
Lycos	4		4	4	86
MSN	2	1	1	2	92
T-Online	4	1	1		79
Web.de	1	2	2	2	84
Yahoo	1	1			11

Insgesamt waren die Reaktionszeiten der Suchmaschinen – trotz mehrfacher Anmeldung der Seiten – unerwartet schlecht, was auf eine geringe Aktualisierungsfrequenz hindeutet. Nur Fireball wies schon nach der dritten Woche Treffer aus, die anderen Anbieter folgten erst nach sieben bis zehn Wochen oder – wie AOL und Yahoo – noch später. Die Anmeldung erfolgte nicht aggressiv durch ein automatisches Script, sondern manuell im Abstand von mehreren Tagen. Zwar behaupten Suchmaschinenbetreiber, dass ein aggressives Anmelden nicht notwendig sei und sogar bestraft werde, es könnte aber trotzdem ein Erfassen der noch nicht registrierten Seiten durch die Robots (man spricht von „spidern" oder „crawlen") initiieren. Das Erfassen der Seiten durch die Robots hätte durch eine intensivere Verlinkung der Seiten von anderen Domains aus beschleunigt werden können, hätte aber das Ergebnis möglicherweise verzerrt, da wegen der Menge der Seiten nicht alle Seiten von außen hätten verlinkt werden können. Zudem haben die Dateien der Linkfarm keine statischen Dateinamen und sind daher für eine Verlinkung nicht geeignet.

Erfolg der Optimierungsverfahren nach Suchmaschinen: Welche Optimierungsverfahren sind am erfolgreichsten (vgl. Tabelle 3)? Darauf lässt sich eine klare Antwort geben: Mehr als ein Drittel aller indizierten Seiten (37%) waren solche, auf die ein Link von einer anderen Seite mit einem hohen „Page Rank" führte. Dies war auch das einzige Verfahren, bei dem alle untersuchten Suchmaschinen reagierten. Und es zeigte schnell Wirkung: Von der fünften bis zur neunten Abfrage waren jeweils über 60% der angezeigten Treffer solche mit „Page Rank". Das Verfahren wirkte am besten – betrachtet man die absoluten Werte – bei Lycos, T-Online, MSN und Fireball. Damit wird die Wirksamkeit der derzeit angeblich effektivsten Methode bestätigt. Allerdings überrascht hier die geringe Sensibilität der Suchmaschine Google, die damit ihre Marktführerschaft errungen haben soll.

Sieht man von den nicht optimierten Homepages der Websites ab (14%), waren – aber mit deutlicher geringerer Wirkung – Keywords (10%), unsichtbarer Text (6%) sowie Keywords mit internen Links (6%) erfolgreiche Spam-Verfahren. Fast ohne Effekt waren Weiterleitungsseiten, und zwar alleine oder in Kombination mit anderen Verfahren, wobei komplizierte, hier nicht angewandte Formen der Weiterleitung (mittels JavaScript verschlüsselte Parameter) mehr Wirkung zeigen könnten. Ebenfalls enttäuschend fiel das Ergebnis für die dynamischen Seiten (durch Cloaking und Linkfarm) aus (2%). Nur zwei Suchmaschinen sprachen darauf an, nämlich Fireball und MSN.

Welche Suchmaschine lässt sich durch welche Verfahren am stärksten beeinflussen? Bei Google versprechen wider Erwarten eher primitive Formen der externen Manipulation Erfolg: Keyword Stuffing, alleine (25%) und in Kombination mit Text in Kommentaren und Alt-Tags (17%), schnitten am besten ab, erst dann folgte die Verlinkung von Seiten mit hohem „Page Rank" (14%).

Die „Page Rank"-Technologie spielt dagegen eine entscheidende Rolle bei anderen Suchmaschinen wie T-Online (86%), Lycos (81%) und MSN (58%). Während Fireball, Google und (im Gefolge) Web.de eine breite Palette von Manipulationsverfahren zuließen, gibt es bei anderen Suchmaschinen nur wenige Lücken in der Spam-Kon-

trolle. Die beiden einzigen Treffer bei AOL gab es für Seiten, die zu anderen Seiten mit hohem „Page Rank" verlinkt waren. Yahoo erwies sich als anfällig für interne Links in Kombination mit Text in Kommentaren und Alt-Tags (36%), Keyword Stuffing (27%) und „Page Rank" (27%), wobei es hier – wie erwähnt – absolut nur um sehr wenige Treffer ging. MSN ließ – neben „Page Rank"-Seiten (58%) und nicht-optimierten Homepages (10%) – intern verlinkte Seiten mit Keyword Stuffing passieren (21%). AltaVista bevorzugte nicht-gespamte Homepages (55%), was für die Qualität der Suchmaschine spricht; vier der fünf Einstiegsseiten konnte AltaVista erfassen.

Erfolg der Optimierungsverfahren nach Domains: Am erfolgreichsten waren die Domains, die auf die Suchwörter „Britney Spears" (235 Treffer) und „Spiele" (176) optimiert waren. Weniger Treffer erzielten „Hotel" (156), „Mallorca" (155) und „Versicherungen" (97). Die Ergebnisse zeigen, dass trotz strukturell identisch angelegter Domains und Verlinkungsstrukturen die Domains vollkommen unterschiedlich tief und ausführlich durch Robots erfasst werden. Das Versprechen der Suchmaschinen, automatisch allen Links zu folgen und alle Dokumente auf einer Domain zu erfassen, wird nicht eingelöst.[10]

Im Erhebungszeitraum gab es markante Verschiebungen: Es fiel zunehmend leichter, „Britney Spears"-Seiten in die Indizes einzuschleusen. Dies lässt sich an der absoluten Zahl der erfassten Seiten zeigen (1. Abschnitt: 34, 2. Abschnitt: 66, 3. Abschnitt: 135), aber auch am Anteil, den die Domain an allen angezeigten Treffern hatte (1. Abschnitt: 19%, 2. Abschnitt: 23%, 3. Abschnitt: 39%). Im Gegenzug wurde es tendenziell schwerer, auf „Versicherungen" optimierte Seiten erfassen zu lassen (1. Abschnitt: 37 Treffer=20%, 2. Abschnitt: 46=16%, 3. Abschnitt: 14=4%).

Auch zwischen den Suchmaschinen gab es deutliche Unterschiede beim Erfassen der Domains: Lycos (47%) und T-Online (44%) präferierten „Britney Spears"-Seiten, Fireball reagierte am stärksten auf „Spiele"-Seiten (32%), MSN (34%) auf „Mallorca"-Seiten. Google (27%) und Web.de (27%) zogen auf „Hotel" optimierte Pages vor. Eine Abgrenzung zwischen einer Nicht-Indexierung einer Seite aus Gründen interner Steuerung der Crawler oder aufgrund erkannter Spam-Methoden ist nicht möglich. Wahrscheinlich ist aufgrund der dennoch aufgenommenen Spam-Seiten, dass die Robots zufallsgesteuert manche Domains mehr, andere weniger gut erfassen, zumindest in den ersten acht Monaten nach der Initialisierung.

10 Google erklärt das folgendermaßen: „Die Anmeldung ist jedoch weder notwendig, noch garantiert sie eine Aufnahme in den Index" (vgl. http://www.google.de/intl/de/webmasters/1.html). Und weiter: „Für die Anmeldung einer Site ist nur die Angabe der Top-Level-Domäne erforderlich, da die Suchroboter automatisch den internen Links zu den übrigen Seiten Ihrer Site folgen." Ähnliches ist auf der Seite bei Fireball zu lesen, über die URL angemeldet werden können (http://rubriken.fireball.de/Suchen/urlmelden.csp), und auf den FAQ bei AltaVista (http://addurl.altavista.com/help/search/faq_web).

Tabelle 3: Sensibilität der Suchmaschinen für Optimierungsverfahren, bereinigt um Mehrfachanzeigen und Linkfarm-Seiten (absolut)

Optimierungsverfahren	*Gesamt*	*Alta-Vista*	*AOL*	*Fireball*	*Lycos*	*T-Online*	*Google*	*Web.de*	*MSN*	*Yahoo*
Keyword Stuffing	83	5		12	10	7	26	20		3
unsichtbarer Text	49	17		3	6	4	10	9		
Weiterleitungsseiten	10			10						
Text in Kommentaren und Alt-Tags	23			12			6	4	1	
Links von einer Seite mit hohem Page Rank	304	29	2	52	70	68	14	13	53	3
Unsichtbarer Text & Keyword Stuffing	25			10			10	4	1	
Weiterleitungsseiten & Keyword Stuffing	6			6						
Weiterleitungsseiten & unsichtbarer Text	5			5						
Text in Kommentaren und Alt-Tags & Keyword Stuffing	44			9			18	17		
Text in Kommentaren und Alt-Tags & unsichtbarer Text	22			18			1	3		
Text in Kommentaren und Alt-Tags & Weiterleitungsseiten	20			20						
interne Links & Keyword Stuffing	45	8		10			5	3	19	
interne Links & unsichtbarer Text	7			6			1			
interne Links & Weiterleitungsseiten	9			9						
interne Links & Text in Kommentaren und Alt-Tags	38			17			9	8		4
Cloaking & Linkfarm	13			4					9	
Startseite (nicht optimiert)	116	72		27			4	3	9	1
Gesamt	819	131	2	230	86	79	104	84	92	11

3.2 Ranking auf den Plätzen 1-20

Innerhalb von acht Monaten gelang es bei zwei Suchmaschinen, jeweils die zwei gleichen Seiten unter die ersten zwanzig Plätze in der Trefferliste zu bringen. Es handelte sich dabei um die Suchmaschinen Lycos und T-Online sowie um zwei Seiten der „Britney Spears"-Domain, deren Seiten ja – wie erwähnt – am häufigsten angenommen wurden und für die Lycos und T-Online eine Vorliebe hatten.

Optimiert waren die beiden hoch eingestuften Seiten für Keyword Stuffing alleine (B2) sowie in Kombination mit unsichtbarem Text (B7). „Britney Spears"-Seiten ließen sich im Erhebungszeitraum zunehmend leichter in die Indizes einschleusen. Entsprechend erschienen auch erst kurz vor dem Ende der Untersuchung, nämlich nach 30 von insgesamt 34 Wochen, die ersten Seiten unter den ersten zwanzig Treffern in den Ranglisten.

Der zeitliche Verlauf war bei Lycos und T-Online weitgehend parallel: Zunächst erschien die Seite B2, in der darauffolgenden Woche stattdessen B7. Während bei T-Online kein weiterer Treffer mehr zu verzeichnen war, blieb bei Lycos die B7-Seite bis zum Erhebungsende in der „Top Twenty"-Liste und erreichte am 28.08. als höchste Platzierung Rang 13.

3.3 Crawlingtiefe

Die Domain „www.wurstbrottwettbewerb.de" für die Analyse der Crawlingtiefe tauchte im Untersuchungszeitraum bei sechs Suchmaschinen im Index auf (AltaVista, MSN, Lycos, T-Online, Google, Web.de). Auch bei diesem Teil des Tests zeigten sich Übereinstimmungen zwischen Google und Web.de sowie zwischen Lycos und T-Online. Sie erfassten paarweise im Gleichschritt die Testseite und behielten sie ungefähr gleich lange im Index (Lycos/T-Online: 12. Abfrage bis 34./31. Abfrage, Google/Web.de: 15. Abfrage bis 19./18. Abfrage). Außerdem erreichten sie genau den gleichen Wert bei der Crawlingtiefe, der sich dann auch im weiteren Verlauf des Tests nicht mehr änderte (Lycos/T-Online: 784 kB, Google/Web.de: 120 kB). AltaVista erzielte die geringste Tiefe mit 76 kB. Mit Ausnahme von MSN verharrten alle Suchmaschinen auf dem Ausgangsniveau der Seitenerfassung. Nur bei MSN gab es zweimal Fortschritte zu verzeichnen, und zwar von der 9. zur 10. Abfrage (114 zu 411 kB) und von der 33. zur 34. Abfrage (411 zu 436 kB).

Aus diesem Ergebnis lässt sich die Erkenntnis ableiten: Weder erfassen Crawler die Seiten in ihrer Gesamtheit, zumindest nicht ab einer bestimmten Größe, noch sind sie (sieht man von der einen Ausnahme ab) in der Lage, die beim ersten Besuch angefangene Erfassung bei weiteren Besuchen fortzusetzen.

4 Fazit

Die hier vorgestellte Pionierstudie über die Effektivität von Optimierungsverfahren liefert einen Einblick in die Arbeitsweise von Suchmaschinen und ihre Fähigkeit, externe Manipulationsversuche abzuwehren. Zwischen den neun untersuchten Suchmaschinen zeigten sich dabei klare Unterschiede: Während AOL und Yahoo nahezu immun gegen die Einflussnahme waren, ließen sich bei Fireball auf vielfältige Weise, schnell und in großer Zahl optimierte Seiten in den Index einschleusen. Hier scheint die Spam-Protection versagt zu haben.

Am deutlich effektivsten erwies sich unter den getesteten Optimierungsverfahren die Verlinkung mit Seiten, die über einen hohen „Page Rank" verfügen, also von Suchmaschinen bereits hoch gelistet werden. Dieses Ergebnis war nicht überraschend – eher schon, dass für den Marktführer Google nicht dieses, sondern ein traditionelles Verfahren, das Keyword Stuffing, den größten Erfolg brachte.

In einigen Fällen ließen sich starke Zu- und Abnahmen bei der Zahl erfasster Seiten feststellen (Fireball, Google, Web.de), die Hinweise auf Aktualisierungszeitpunkte geben. Regelmäßige Zyklen ließen sich jedoch nicht ermitteln.

Mehrfach ließen sich für Google und Web.de sowie für Lycos und T-Online parallele Verläufe und übereinstimmende Ergebnisse registrieren. Sie dürften aus (zeitweisen) Kooperationen zwischen diesen Anbietern resultieren.

Zwar gelang es, dass viele der optimierten Seiten von den Suchmaschinen erfasst wurden. Als sehr viel schwerer erwies es sich, die für die Nutzung relevanten Rangplätze 1 bis 20 zu erklimmen. Dies gelang nur für zwei Seiten bei zwei Suchmaschinen, und dies auch erst kurz vor dem Ende des achtmonatigen Untersuchungszeitraumes.

Ein letzter Test galt der Vollständigkeit, mit der Webseiten von Suchmaschinen-Crawlern erfasst werden. Hier zeigte sich, dass Crawler die Seiten sehr unterschiedlich tief analysierten und sie (mit einer Ausnahme) auch nicht fähig waren, im Laufe der Zeit tiefer in Dokumente einzudringen.

Abschließend sind noch einige Anmerkungen zur Methodik der Studie zu machen: Der gewählte Untersuchungszeitraum von acht Monaten scheint ausreichend lang gewesen zu sein, um die Reaktionen der Suchmaschinen zu testen. Allerdings fällt es schwer, die Verallgemeinerbarkeit der Ergebnisse abzuschätzen. Zeitlich gesehen, dürfte sich das Konkurrenzumfeld für die gewählten Domains und die Effektivität der Optimierungsverfahren rasch ändern. Außerdem ist es erforderlich, die zum Untersuchungszeitpunkt tatsächlich angewandten Verfahren zu imitieren.

Mit der vorgelegten Studie ist es zweifellos gelungen, ein praktikables Design für den Vergleich unterschiedlicher Optimierungsverfahren zu entwickeln. Dennoch ist rückblickend im Vergleich mit der aktuellen Entwicklung festzustellen, dass für ein effektives Spamming die Zahl der Domains und die Zahl der Seiten am unteren Ende dessen rangierten, die seit etwa sechs Monaten von professionellen Google-Manipulatoren eingesetzt werden. Üblich geworden sind Hunderttausende von Seiten auf Dutzenden von Domains und vor allem eine höhere Verlinkung der manipulierten

Seiten von bereits länger existierenden Seiten mit einem hohen „Page Rank“ (vgl. Karzauninkat 2003). Im Rahmen der Möglichkeiten der Studie war ein derart massiver Einsatz von Ressourcen und das tägliche aufwendige Nachbessern der Verlinkungsstrukturen nicht leistbar.

Literatur:

Greenspan, Robyn (2002): Search Engine Usage Ranks High. In: CyberAtlas. 14.11.2002. http://cyberatlas.internet.com/markets/advertising/article/0,1323,5941_1500821,00.html (03.11.2003)

Karzauninkat, Stefan (2003): Google zugemüllt. Spam überschwemmt die Suchergebnisse. In: c’t. Nr. 20 v. 22.09.2003, S. 88-91.

Machill, Marcel/Neuberger, Christoph/Schweiger, Wolfgang/Wirth, Werner (2003): Wegweiser im Netz. Qualität und Nutzung von Suchmaschinen. In: Marcel Machill / Carsten Welp (Hrsg.): Wegweiser im Netz. Qualität und Nutzung von Suchmaschinen. Gütersloh: Verlag Bertelsmann Stiftung, S. 13-489.