Mario Fischer

SPAMFILTERBAU: MACH MIR DEN PINGUIN

Das Ausrollen des sog. Pinguin- oder Penguin-Updates am 24.04. dieses Jahres, das vorab in der Szene als „SEO-Filter“ bezeichnet wurde, sorgte für große Aufregung. In Fachforen und -blogs liefen die Kommentarfunktionen heiß. Aus vielen Beiträgen wurde aber auch deutlich, dass selbst vielen Suchmaschinenoptimierern noch nicht ganz klar ist, wie solche „Spamfilter“ entstehen und wie sie letztlich von den Suchmaschinenbetreibern für den Praxiseinsatz trainiert und optimiert werden. Wer das Risiko, mit anvertrauten Seiten aus dem Suchindex zu fallen, vermeiden oder zumindest vermindern möchte, sollte die nachfolgenden Ausführungen zum besseren Verständnis der Denk- und Funktionsweise algorithmischer Updates aufmerksam studieren.

Dass Links nach wie vor großes Gewicht insbesondere bei Google haben, wenn es um die Bewertung der Wichtigkeit und der Relevanz einer einzelnen Webseite geht, ist in Expertenkreisen so gut wie unumstritten. Links und deren Ankertexte waren schon immer das Salz in der Websuppe, mit dem Google versuchte, ein externes Votum für die Beurteilung mit einfließen zu lassen. Googles Problem: Das hat sich mittlerweile bis zum Kirmeslosverkäuferauszubildenden herumgesprochen. Und während immer mehr Agenturen und SEOs fleißig Link um Link zum vermeintlichen Kundenwohl aufbauten, heckten die Indexhüter rund um Matt Cutts in Mountain View mal wieder etwas aus, das Seiten, die „künstlich“ im Ranking nach vorn getrieben wurden, wieder auf unattraktive Plätze zurücksetzt, wo sie nach Googles Meinung hingehören. Dass dies passieren würde, war durchaus absehbar, und Website Boosting hatte im letzten Editorial auch davor gewarnt.

Um den oft falsch angesetzten Zweckoptimis-

Die Hauptgründe für das Setzen von Links

- Gefälligkeit oder Geschäftsbeziehungen
- Linktausch oder Linkkauf
- Online-Marketing-Maßnahmen (Banner, Affiliate etc.)
- Hinweis auf guten und/oder nützlichen Content
- Hinweis auf eine Quelle oder Person
- Beeinflussung des Rankings

Sehr geehrter Herr Fischer,

mein Name ist [redacted] und ich bin zuständige Marketing Managerin von [redacted]Online Marketing. Unsere Kunden haben Interesse an einer strategischen Kooperation mit Ihrer Seite http://[redacted].de signalisiert. Dabei geht es um einen Verweis unserer Kunden innerhalb bestehender Inhalte oder eines für Sie verfasssten, qualitativen Gastartikels. Ihnen entstehen dabei selbstverständlich keine Kosten.

Als Gegenleistung können wir Sie in einem Artikel über Internet-Startups auf den Seiten der [redacted] [große deutsche Tagesezeitung, Anm. d. Red.] verlinken. Die passenden Inhalte würden von unserer Redaktion erstellt werden. Ich denke, dass beide Seiten davon profitieren könnten, da es sich
dabei um eine gut besuchtes Portal mit hoher Sichtbarkeit handelt, von der sich mehr Traffic für Ihre Seite ableiten könnte.

Bei Interesse lassen Sie mir bitte eine kurze Rückmeldung zukommen, im nächsten Schritt würde ich Ihnen gerne ein unverbindliches Angebot unterbreiten.

Mit freundlichen Grüßen aus München,
[redacted]
Online Marketing Managerin

[redacted] [redacted] München
E-Mail: [redacted] marketing.com

Sehr geehrte Frau [redacted],
vielen Dank für Ihre Mail. Habe ich Sie richtig verstanden, dass Sie mir da einen Link von der [redacted] [große deutsche Tagesezeitung, Anm. d. Red.] anbieten? Ich dachte immer, solche „Institutionen“ lassen sich redaktionell nicht reinreden und bleiben unabhängig?
Ist denn das legal oder zumindest moralisch sauber?
Mit Freude erwarte ich Ihr diesbez. Angebot.
Gruß nach München,
Mario Fischer

Hallo Herr Fischer,

vielen Dank für Ihre Nachricht.

Unsere Kunden [redacted] [weltweit agierender Zahlungsanbieter, Anm. d. Red.] und [redacted] (ein großes Reiseportal, Anm. d. Red.] haben Interesse an Ihrer Seite bekundet. Vielleicht haben Sie ja noch weitere Seiten in Ihrem Portfolio?
Die Artikel sollen auf einer Unterseite erscheinen, die mit der Startseite verlinkt ist.
Im Gegenzug würden wir Sie auf einem unserer Portale verlinken - hier ein Beispiel für eine Partnerintegration:
http://[redacted].[redacted].de/startups/2012/03/[redacted]/
Das ist legal und von [redacted] [der Tageszeitung, Anm. d. Red.] auch so erwünscht. Sollten Sie damit einverstanden sein, müssen Sie mir nur noch Ihren gewünschten Link mit Keyword mitteilen und ich leite das an unsere Redaktion weiter.

Mit freundlichen Grüßen,
[redacted]
Online Marketing Managerin

Abb. 1: Die Linktauscher werden immer dreister (mit Anmerkungen ergänzt)

mus, den man häufig in der Branche findet („Das ging bisher auch immer gut.“ „Da werden nur wieder Nebelkerzen von Matt Cutts geworfen.“), soll es an dieser Stelle aber gar nicht gehen, sondern darum, wie solche „Filter“ – genauer eigentlich „Updates im Bewertungsalgorithmus“ – gebaut werden und wie sie wirken. Dabei kämpfen die Suchmaschinen prinzipiell an zwei Fronten. Zum einen geht es darum, eine immer feinere Relevanzmessung vorzunehmen. Mit anderen Worten: Welche Seiten sind die besten, wenn jemand einen Suchbegriff bzw. eine Suchphrase eingibt? Hierzu werden viele Signale herangezogen und die produzierten Ergebnisse werden mittels sog. A/B-Tests (etwa einem Prozent der Suchenden werden anders berechnete Ergebnisse gezeigt und die Verhaltensveränderung gegenüber den normalen Ergebnissen gemessen) und durch die Überwachung von Bounce-Rates (eine schnelle Rückkehr zur Ergebnisliste nach einem Klick auf einen Link, die in der Regel bedeutet, dass der Suchende nicht zufrieden war) ständig überwacht und ggf. weiter optimiert.

Degree of Dependence

Die zweite große Front ist die Klärung von Verwandtschaftsverhältnissen und die Erkennung bewusster Manipulationen von Bewertungssignalen. Es gilt, maschinell den wahren Grund herauszufinden, warum die Webseite A der Domain X einen Link auf Website B der Domain Y setzt. Linkt X auf Y, weil man sich kennt oder es gar die eigene Tochterfirma ist, oder kennen sich X und Y mit hoher Wahrscheinlichkeit überhaupt nicht und der Linkgrund ist tatsächlich, dass Y etwas Nützliches auf seiner Website B hat, auf das man aufmerksam machen möchte.

Linktausch zu erkennen, ist für Suchmaschinen wie Kindergeburtstag

Linktausch ist recht einfach zu erkennen, da beide Parteien hin- und herlinken und somit selbst eine sichtbare Beziehung zwischen sich herstellen. Mittlerweile erkennen Suchmaschinen auch sog. Dreiecksverlinkungen recht zuverlässig, wo eine dritte Website mit einbezogen wird, um die direkte Beziehung zu verbergen.

Solche Linkschaukeleien werden übrigens von Suchmaschinenoptimierern in der Regel tatsächlich oft unter Einbezug einer dritten Partei vorgenommen. Wer hat nicht schon die eine oder andere Mail bekommen, in der die eigene Website in den Himmel gelobt und ein Linktausch vorgeschlagen wurde. Dieses System wird mittlerweile recht offen und offenbar ohne jedes Bedenken in die Breite getreten. Erst vor Kurzem traf in der Redaktion wieder eine automatisiert erstellte Mail ein. Die Besonderheit: Unverhohlen wurde ein redaktioneller Beitrag mit einem Link in der Onlineausgabe einer sehr großen deutschen Tageszeitung angeboten, wenn man aus dem eigenen Blog einen Beitrag (wird geliefert) mit einem Link zu einem weltweit be-

Abb. 2: Menschen helfen, die maschinellen Augen zu schärfen

kannten Zahlungssystem setze. Die Agentur erklärte auf die bewusst naiv formulierte Nachfrage, ob das denn echt legal wäre, redaktionellen Inhalt zu „verkaufen“ sei doch möglicherweise nicht mit den ethischen Prinzipien von Verlagen zu vereinbaren: Jaja, das hätte alles schon seine Ordnung und sei von der Zeitung so gewünscht. Man hätte als Agentur Zugriff auf bestimmte Verzeichnisse des Onlineangebots und könne diese nach eigenem Belieben bestücken. Es ist also zumindest online nicht alles Zeitungsgold, was zu glänzen versucht.

Google, bitte, bitte, schau doch mal her, wie schmutzig meine Hände sind!

Wahrscheinlich ist es nicht die intelligenteste aller Ideen zum Linkaufbau, völlig unkontrolliert maschinelle Linktauschanfragen zu verschicken, vor allem, wenn die Markennamen bekannter Unternehmen dabei auf Vollmast geflaggt werden. Wer die Hintergründe nicht kennt, dem sei kurz erklärt, wie diese Methode funktioniert. Man kauft sich hierzu für einige Hundert Euro Software, gibt dort den gewünschten Suchbegriff ein und die Maschine holt dann von Google z. B. die ersten hundert Treffer. Anschließend werden die gefundenen Domains automatisch auf Adressen und vor allem E-Mail-Adressen (erkennbar im Quellcode durch „mailto:“) durchsucht und diese in Datenbanken gespeichert. Anschließend trägt man unter Verwendung von Platzhaltern einen Mailtext ein und die Maschine verschickt auf Wunsch unbegrenzt Mailanfragen. Die Mails lesen sich oft individuell, denn über die Platzhalter werden auch persönliche Daten von der Website eingefügt. Dem aufmerksamen Leser offenbart sich erst auf den zweiten Blick, dass hier eine Maschine schreibt und nicht etwa ein freundlicher Mensch. Die Problematik bei dieser eigentlich sehr billig durchführbaren „Maßnahme“ liegt darin, dass eben niemand mehr bei den Massen von gesammelten Daten ein Review macht oder erkennen kann, wer da tatsächlich angeschrieben wird. Es soll auch schon vorgekommen sein, dass Mitarbeiter von Google solche Mails erhalten haben, weil auch die ein Privatleben haben, über das sie ggf. im Web „auftauchen“. Solche Großeinladungen zum Linktausch sind sicher ein gefundenes Fressen für die Spamfighter bei Google. Wie immer kann man nur warnen, „Knopfdruck“-Software für dauerhaftes SEO einzusetzen. Wer es schnell und bequem haben will, wird in der Regel nicht lange Freude an auf diese Weise erzeugten Rankings haben. Inwiefern Google hier auch maschinell lernen kann, darauf kommen wir zu einem späteren Zeitpunkt nochmals zurück.

Was sich liebt, das linkt sich

Herkömmliche Methoden der Erkennung von „Verwandtschaft“ im Web funktionieren recht einfach über Auswertung der bekannten Signale, wie z.B.:

» Gleicher IP-D-Block, unter dem die Sites gehosted werden
» Gleicher Admin-C als hinterlegte natürliche Person
» Gleiche Telefon- oder Faxnummer bei den Kontaktdaten, auch gleiche E-Mail-Adressen
» Verwaltung unterschiedlicher Domains in einem Google-Account (insb. Webmaster-Tools und Google Analytics)
» Gleiche Adresse im Impressum

Im Web liest man immer wieder, man solle zur Tarnung im Impressum die Schrift als Bild hinterlegen, damit Google sie nicht auslesen könne. Neben den juristischen Problemen, die man sich damit einhandeln kann, impliziert dies aber auch, die Suchmaschine wäre nicht in der Lage, einfachste OCR-Aufgaben zu erledigen, die jede Software, die Scannern im 50-Euro-Preissegment bereits beigelegt ist, fehlerfrei beherrscht.

Die aufgezählten Signale sind nur ei-

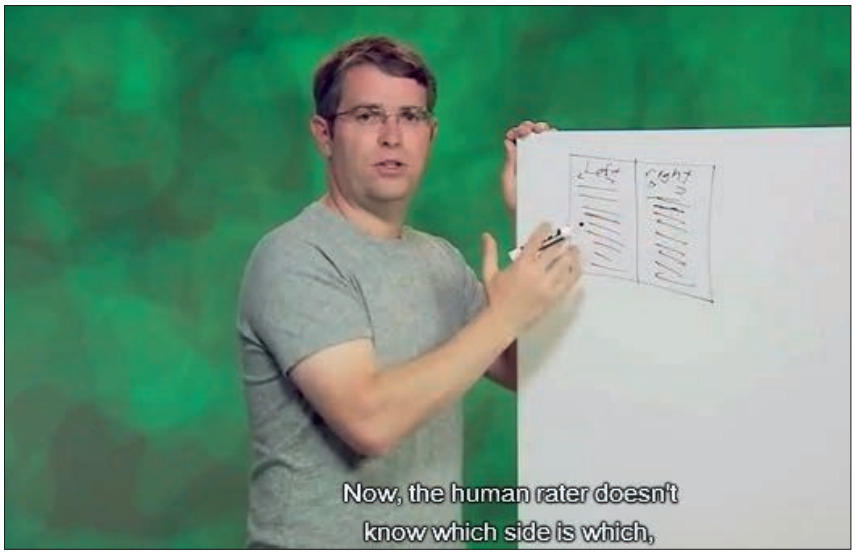

Abb. 2: Auf *http://einfach.st/mchr* erklärt Matt Cutts von Google, wie die Qualitätstester arbeiten.

nige und die augenfälligen. Das Prinzip dahinter ist immer das gleiche: Die Spamfighter finden eine Möglichkeit, ein Signal auszulesen und aus dem Inhalt entsprechende Schlüsse zu ziehen. Kann man mit einer gewissen Wahrscheinlichkeit einen hohen Verwandtschaftsgrad zwischen Websites erkennen, lassen sich die Links zwischen ihnen im Einfluss limitieren. Damit bringt das intensive Verlinken z. B. über eigene Domains fast nichts mehr. Google hat übrigens schon 2004 ein entsprechendes Patent „Determining quality of linked documents" eingereicht, das am 24.08.2010 erteilt wurde (*http://goo.gl/WaOU*).

Mit dieser Methode allein wäre Google allerdings nicht weit gekommen.

Think big – big Data!

Googles Ansatz zur Erkennung von Signalen geht sehr viel weiter und tiefer. Man hat ja schließlich nicht mehr und nicht weniger als (fast) alle im Web offen erreichbaren Dokumente in den eigenen Datenbanken gespeichert. Dies dürften mittlerweile mehrere Hundert Milliarden an der Zahl sein und ihre Verlinkungen untereinander liegen um ein Vielfaches höher. Damit stehen dem Suchgiganten praktisch alle notwendigen Daten für Analysen zur Verfügung. Was jetzt noch fehlen würde, wäre ein Referenzset von guten und nützlichen Webseiten im Gegensatz zu einem Set an Spamseiten, die eigentlich niemand haben möchte. Und in der Tat hat Google auch dieses Prüfset zu Verfügung. Die sog. Qualitätstester (nicht zu verwechseln mit den festangestellten Spamfightern bei Google) prüfen in der Regel auf 400-Euro-Basis aus einer vorgegebenen Liste von Webseiten, welche gut (cool), welche unnütz (uncool) sind und welche möglicherweise sogar Spamtechniken einsetzen, die die Maschine noch nicht erkennen kann.

So entstehen zwei große Pools an Seiten, die von Menschen als gut (Pool A) oder eben als nicht gut (Pool B) klassifiziert wurden. Jetzt machen sich die Maschinen an die Arbeit und versuchen, Unterschiede zwischen A und B zu erkennen. Die Kernfrage lautet dabei: Worin unterscheiden sich Seiten in Pool A von denen in Pool B? Im ersten Moment mag man vermuten, sicher durch die Wortwahl oder die Themen – vielleicht auch durch die Anzahl Links. Aber wer jemals wirklich mit „Big Data" zu tun hatte, weiß, dass es so einfach eben nicht ist. Hier liegt oft auch die Krux: Als Mensch sieht man immer nur weniger als ein Millionstel eines Promilles des gesamten Webs als Ausschnitt. Was auf einer Seite „ganz klar" nach Spammustern aussieht, das kann auf anderen Websites durchaus völlig normal sein. Hierin liegt die Kunst der Spamsignal-Suche: Trennscharfe Signale finden! Da Google verständlicherweise keine Informationen hierüber herausgibt, muss man sich die publizierten Ergebnisse anderer Forscher ansehen, um erkennen zu können, welche zum Teil unerwarteten Signale man auf diese Art und Weise finden kann. Dabei muss man natürlich immer im Hinterkopf behalten, dass Google ungleich mehr Ressourcen für solche Analysen zur Verfügung hat.

2004 fanden Fetterly, Manasse und Najork von Microsoft u. a. heraus, dass z. B. die Länge des Hostnamens ein Indiz für Spam sein kann. Je länger der Hostname, desto wahrscheinlicher handelt es sich um eine Spamdomain. Ebenso, wenn mehr als fünf (Trenn-) Punkte in

Wichtige Google-Updates

» 2009 - Vince-Update
Kern: Reputation von Marken wurde gestärkt und im Suchergebnis bevorzugt

» 2010 - Brand-Update
Kern: Eine Domain kann mehr als 1 bis 2 mal in den Suchergebnissen auftauchen (eigentlich kein Update im traditionellen Sinn; wird oft mit dem Vince-Update verwechselt)

» 2010 - Caffeine-Update
Kern: Veränderung der Infrastruktur für schnellere interne Berechnungsmöglichkeiten und kürzere Reaktionszeiten

» 2010 - Mayday-Update
Kern: Erhöhung der Contentqualität und differenziertere Bewertung sog. Long-Tail-Keywords

» 2011 - Panda-Update *(auch „Farmer"-Update genannt)*
Kern: Berücksichtigung des Userverhaltens und damit Erkennung „wertloser" Webseiten

» 2011 - Freshness-Update
Kern: Hoch aktuelle Inhalte werden kurze Zeit bevorzugt gerankt

» 2012 - Penguin-Update
Kern: Erkennung überoptimierter Seiten, die manipulative Techniken einsetzen

Code 1

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1- transitional.dtd">
<html lang="de" xml:lang="de" xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Das Penguin Update von Google, der "SEO-Filter"</title>
<meta name="description" content="Der 24. April 2012 wird ein denkwürdiges Datum bleiben. An
diesem Tag startete Google mit einem Update im Bewertungsalgorithmus eine Offensive gegen über-
optimierte Seiten"/>
<meta name="keywords" content="Ach ja? Sie befüllen das immer noch???" />
<meta http-equiv="expires" content="0" />
<meta name="author" content="Mario Fischer">
<link rel="shortcut icon" href="http://www.website-boosting.de/favicon.ico" >
<base href="http://www.websiteboosting.com/" />
<link rel="start" href="http://www.websiteboosting.com/" title="Home" />
<link rel="alternate" type="application/rss+xml" title="RSS 2.0"
href="http://www.websiteboosting.com/blog/index.xml" />
<script language="JavaScript" type="text/javascript">
<!--function SymError()
{
  return true;
}
window.onerror = SymError;
//-->
```

Code 2

```
<! "-// // // " "">
< =" ": =" " =" ">
< >
< =" " ="; " />
< ></ >
< =" =""/>
< =" =""/>
< =" "=""/>
< ="" =" ">
< =" " =" // / " >
< =" // /" />
< =" " =" // /" =" " />
< =" " =" / " =". " =" // / /. " />
< =" " =" / ">
<!-- ()
{
  ;
}
=;
//-->
```

der URL vorkamen. 80 der 100 längsten Domainnamen waren Pornografie und weitere elf Spam aus dem Finanzumfeld. Sie fanden ebenso heraus, dass vor allem dann Spam anzutreffen war, wenn viele Domains auf einer IP-Adresse lagen. Bei 3,5 % aller IP-Adressen findet man pro IP mehr als 10.000 Domains – ein starkes Signal. Dabei fanden Sie übrigens auch eine IP-Adresse, die für über 9 Mio. Domains benutzt wurde. Aber auch, wie häufig Content geändert wurde oder wie sog. Near-Duplicate-Content verteilt ist, sind ihrer Ansicht nach trennscharfe Signale (*http://einfach.st/cscol*).

Auch über das sog. Fingerprinting von Dokumenten lässt sich Spam finden. Dabei werden die flexiblen und individuellen Inhalte von HMTL-Seiten entfernt, sodass nur noch der HSS Noise ggf. zusammen mit den reinen Tags übrig bleibt.

Aus dem folgenden Code-Ausschnitt (Code 1) bleibt nach der Entfernung bestimmter Inhalte ja nach Entfernungsregel der „Fingerabdruck“ wie er in Code 2 zu sehen ist übrig:

Wurde eine Webseite mithilfe einer Software oder z. B. eines Skripts erstellt und anschließend nicht manuell individualisiert, dann lässt sich an diesem Fingerprint oft recht zuverlässig ablesen, mit welcher Software diese Seiten erzeugt wurden. Hat man z. B. bei automatisiert erstellten Spamseiten einen solchen gemeinsamen Fingerprint identifiziert, kann man sich damit gezielt auf die Suche nach anderen Seiten manchen, die von demselben System produziert wurden. Weiterführende Überlegungen dazu findet man bei Urvoy et. al. (PDF unter *http://einfach.st/urvoy*).

Webb et al. (*http://einfach.st/webb*) fanden u. a. durch Musteranalyse heraus, dass bei etwa 60 % aller Webseiten mit Spam im Header des Webservers das Muster „x-powerd-by“ übertragen wurde, aber nur bei 29 % aller spamfreien Seiten.

Spam-Algorithmen sind extrem komplex

Natürlich sind dies alles nur einzelne Signale und keinesfalls taugen sie, um bei einzelnem Auftreten eine Seite wegen Spamverdachtes aus dem Index zu verbannen. Erst die Kombination vie-

ler Hundert Signale und das Einstellen von Schwellwerten machen diese Signale zu wertvollen Detektoren im Kampf gegen Webspam. Um es für das notwendige Verständnis der prinzipiellen Mechanik der Wirkungsweise solcher Algorithmen nicht unnötig zu verkomplizieren, ist es förderlich, die in der Praxis extreme Komplexität zu vernachlässigen und einfache Beispiele zu verwenden.

Abbildung 3 zeigt anhand fiktiver Einträge beispielhaft eine mögliche Entscheidungstabelle zu Gewichtung und Verdichtung von Spam-Signalen. In der Spalte Ausprägung werden die maschinell ermittelten Werte des jeweiligen Signals eingefügt. Diese unterschiedlichen Daten werden in der nächsten Spalte auf Werte von z. B. 0 bis 10 normiert, damit sie miteinander vergleichbar werden. Anschließend erfolgt eine Gewichtung des Signals im Vergleich seiner Bedeutung und Trennschärfe (Spam/Nicht-Spam) gegenüber den anderen Signalen. Anschließend werden jedem Signal SpamPoints zugewiesen oder auch Punkte „gutgeschrieben", wenn das Signal positiv zu werten ist (im Beispiel das Domainalter und das Vorhandensein von fünf Links extrem vertrauenswürdiger Domains).

In unserem Beispiel käme die betrachtete Domain auf einen Spam-Wert von 2,88 (6,25 SpamPoints abzüglich der Gutschrift von 3,37 Punkten durch positive Signale). Läge nun der angenommene Wert für ein De-Ranking oder die Verbannung aus dem Index bei einem Wert von 3,0, wäre diese Domain nicht betroffen.

Nimmt man an, dass die Domain statt 11 Jahre wie im ersten Beispiel nur 2 Jahre alt ist, fällt die maschinelle Beurteilung ganz anders aus (siehe Abbildung 4):

Nun läge die Summe auf 3,36 und damit über dem Schwellwert von 3,0. Die Domain würde „bestraft". (In der Tat ist es so, dass ältere Domains bei Google einen höheren Vertrauensstatus genießen.) Ergänzt man das Gedankenspiel

Signal (fiktive Beispiele!)	Ausprägung	Normierung	Gewichtung	SpamPoints	TrustPoints
Anzahl ausg. Links im Footer	5/112	8	0,06	1,48	
Cross-Domain-Verlinkung	3	4	0,02	1,08	
Topic-Relevanz eing. Links	12/385	3	0,08	1,24	
Keyword-Stuffing	12,3	4	0,03	1,12	
Trusted-Domain-Links	5	9	0,09		1,81
Domainalter	11	7	0,08		1,56
Linkverhältnis IP-C/IP-D	31/829	3	0,11	1,33	
...					
...					
Summe: 2,88				*6,25*	*3,37*

Abb. 3: Fiktive Darstellung der Gewichtung von Spamsignalen

Signal (fiktive Beispiele!)	Ausprägung	Normierung	Gewichtung	SpamPoints	TrustPoints
Anzahl ausg. Links im Footer	5/112	8	0,06	1,48	
Cross-Domain-Verlinkung	3	4	0,02	1,08	
Topic-Relevanz eing. Links	12/385	3	0,08	1,24	
Keyword-Stuffing	12,3	4	0,03	1,12	
Trusted-Domain-Links	5	9	0,09		1,81
Domainalter	***2***	***1***	0,08		1,08
Linkverhältnis IP-C/IP-D	31/829	3	0,11	1,33	
...					
...					
Summe: 3,36				*6,25*	*2,89*

Abb. 4: Die Beurteilung einer noch jungen Domain fällt anders aus

Signal (fiktive Beispiele!)	Ausprägung	Normierung	Gewichtung	SpamPoints	TrustPoints
Anzahl ausg. Links im Footer	***0/112***	***0***	0,06	1	
Cross-Domain-Verlinkung	3	4	0,02	1,08	
Topic-Relevanz eing. Links	12/385	3	0,08	1,24	
Keyword-Stuffing	12,3	4	0,03	1,12	
Trusted-Domain-Links	5	9	0,09		1,81
Domainalter	***2***	***1***	0,08		1,08
Linkverhältnis IP-C/IP-D	31/829	3	0,11	1,33	
...					
...					
Summe: 2,76				*5,61*	*2,89*

Abb. 5: Signale können sich gegenseitig überstrahlen bzw. aufheben

> „Ob tatsächlich Spam vorliegt, ist maschinell eindeutig gar nicht so leicht erkennbar!“

nun um eine weitere Änderung (Abbildung 5), nämlich dass keine ausgehenden Footerlinks auf externe Domains vorhanden sind, wird das fehlende Alter wieder ausgeglichen. Die Domain würde wieder ranken, denn der Wert läge nun insgesamt wieder unter 3,0.

Alle Zahlen in diesem Beispiel sind natürlich fiktiv. Aber sie zeigen recht gut die Mechanik, die hinter den Spam-Algorithmen steht. In der Realität kann man davon ausgehen, dass Google hier sicherlich viele Hundert Signale auswertet und verarbeitet. Dabei werden auch die „Schrauben“ transparent, an denen die Spamfighter drehen können. Normierungswerte und Gewichtung können und werden sich sicherlich im Lauf der Zeit ändern. Kann man ein Signal durch neue Techniken genauer bestimmen oder hat zusätzliche Erkenntnisse über die zunehmende Verwendung auf Spamseiten, dann kann man neben der Normierung auch die Gewichtung anpassen. Natürlich lässt sich auch mit dem Schwellwert (im Beispiel 3,0) arbeiten. Bedenkt man nun, dass Google in der Realität auch noch unterschiedliche Arten von Strafen verwendet (auf Domainebene, aber auch auf Einzelkeywords oder einzelnen Verzeichnissen) und diese unterschiedlich ausprägbar sind (Verbannung aus dem Index, De-Ranking um z. B. 30, 60 oder 90 Positionen nach hinten), dann wird deutlich, wie komplex das System ist und wie filigran es steuerbar ist. Schließlich kommt am Ende zu diesen Überlegungen auch der Umstand hinzu, dass Google sicher im Lauf der Zeit auch unterschiedlich tolerant gegenüber bestimmten Spam-Techniken wird und daher auch die Toleranzwerte für die Punktevergabe ändert. Und sicherlich sind auch Signale vorstellbar, die im Extremfall als Ausnahme allein durch ihr singuläres Auftreten andere Signale überdecken, wie z. B. das Cloaking oder die Verwendung sog. Brückenseiten.

Es wird aber auch deutlich, warum zum Teil lange Zeit augenscheinlich eine bestimmte Art von Spamseiten noch immer im Index zu finden ist: Es ist gar nicht so einfach bzw. im Regelfall unmöglich, aufgrund weniger Signale (auch wenn sie für das menschliche Auge beim ersten Blick „leicht“ erkennbar sind) zu entscheiden, ob tatsächlich Spam oder ein Manipulationsversuch des Webseitenbetreibers vorliegt. Ein Beispiel: Eine Website ändert das Design und die frühere Hintergrundfarbe wird neben anderen Dingen von einem Grauwert auf Weiß gesetzt. Vielleicht gibt es noch ein Set von Unterseiten, auf denen teilweise aus optischen Gründen weiße Schrift verwendet wird, die sich vorher von dem grauen Hintergrund ganz gut abhob. Durch die Umstellung auf den nun weißen Hintergrund sind diese Textteile nicht mehr sichtbar. Weiße Schrift auf weißem Grund war früher eine klassische Spamtechnik, um Text vor Besuchern zu verbergen, aber dem Robot der Suchmaschine genügend relevanten Text zu liefern. Nach der Änderung liegt ein Verstoß gegen die Richtlinien von Google vor – unsichtbare Schrift. In diesem speziell konstruierten Fall liegt allerdings nun aber keine manipulative Absicht vor, sich einen Vorteil beim Ranking zu verschaffen.

Dies ist der Grund, warum größere Updates augenscheinlich oft so lange auf sich warten lassen. Die Auswirkungen der Aufnahme neuer Signale oder die Änderung von Gewichtungen müssen sauber getestet werden. Dies kann unter anderem eben auch mithilfe der oben genannten Domainpools vorgenommen werden. Mit ihrer Hilfe kann Google testen, wie viele Spamdomains durch die Änderung „erwischt“, aber vor allem eben auch, wie viele Domains aus dem Pool mit nützlichen Domains zu Unrecht weggefiltert würden! Ziel muss es sein, mit den jeweils neuen Algorithmen möglichst wenige Kollateralschäden zu verursachen. Sicher lässt sich trefflich darüber streiten, was denn nun besser wäre: X Prozent weniger Spam im Index zu haben, dafür aber Y Prozent von den Suchenden als sehr nützlich empfundene Seiten aus den Ergebnissen zu verbannen – und dies eben nur, weil sie zufällig, aus welchen Gründen auch immer, ähnliche Signale senden.

Google ändert seinen Algorithmus praktisch täglich

Auch dies wird häufig vergessen: In der Regel bemerkt man nur größere Änderungen, die als „Update“ auch einen Namen bekommen wie Vince-, Panda- oder Penguin-Update. Tatsächlich aber werden pro Jahr Hunderte Verbesserungen eingespielt. So hat Google offiziell für den Januar dieses Jahres 17 Änderungen publiziert, allerdings nur die „Highlights“ (*http://einfach.st/gjan*). Man darf getrost davon ausgehen, dass es eine ganze Reihe weiterer Änderungen gab. Im Dezember 2011 gab es z. B. 30 solcher Änderungen (*http:// einfach.st/gdez*) und im April dieses Jahres gar 52 (*http://einfach.st/gapr*). Eine Zuordnung einer beobachteten Änderung muss sich also nicht nur auf die großen Updates beziehen, sondern kann praktisch auch wenige Tage vorher oder nachher durch eine der vielen kleineren Änderungen verursacht worden sein. Im April änderte Google z. B. die Art und Weise, wie inhaltlich zusammengehörige Seiten indexiert werden, die über Blätternavigationen aufgeteilt wurden. Diese werden seither als zusammengehörig betrachtet. Hat eine Domain viele solcher aufgeteilten Contentseiten, kann sich das spürbar auf das gesamte Ranking auswirken. Dies ist auch der Grund, warum Analysen und Studien von außen so schwierig sind.

Zum einen ändern sich z. B. die anderen Webseiten, die auch im Ranking eine Rolle spielen, inhaltlich und aufgrund dessen ändert sich die betrachtete Position der eigenen Webseite – ohne dass es dort Änderungen gegeben hätte. Zudem wirken die vielen kleinen Updates sich natürlich in Summe auch an dieser Stelle aus. Von einer Bewegung der eigenen Webseite im Ranking kann also nur mit extremer Unsicherheit auf die einem selbst bekannte vorgenommene Änderung an dieser Website geschlossen werden. Viel wahrscheinlicher könnte mittlerweile sein, dass Änderungen in der Bewertung von Google vor allem auch Auswirkungen auf andere Seiten haben und die eigene nur von deren Positionsveränderungen rauf oder runter verschoben wird. Dies sollte man immer auch kritisch im Hinterkopf behalten, wenn der eine oder andere Vortragende auf Konferenzen einfachste Ursache-Wirkungs-Beziehungen eigener Experimente vorstellt.

Algorithmusopfer oder Strafgefangener?

Stellt man plötzliche Verschlechterungen im Ranking fest, empfiehlt sich, eine fundierte Ursachenanalyse durchzuführen oder durchführen zu lassen. Es macht einen wesentlichen Unterschied, ob die Site „nur“ durch eine Verschärfung des Algorithmus (oder eines Teils dessen) spürbar Positionen verloren hat oder ob Google tatsächlich eine spezifische Domain-, Keyword- oder Verzeichnisstrafe verhängt hat.

Gibt man in den Suchschlitz „site:domain.de“ ein und erhält keinerlei Treffer, dann liegt eine domainweite Strafe vor. Eine Strafe für ein einzelnes Verzeichnis lässt sich ebenfalls auf diesem Weg leicht erkennen, indem man den Verzeichnisnamen an „domain.de/verzeichnis“ anhängt. Um Probleme bei einzelnen Keywords zu diagnostizieren, muss man vorab bereits eine Überwachung geschaltet haben, um die Rankingverläufe und damit drastische und plötzliche Abfälle bei den Positionen sehen zu können. Eine punktuelle Abfrage bringt hier in der Regel wenig Erkenntnisgewinn. Wer professionelle SEO-Tools nutzt, ist hier natürlich besser gewappnet, denn die meisten der Tools überwachen die eigenen definierten oder auch generell passenden Keywords und deren Positionen automatisch. Man kann sich eine Strafe so vorstellen, dass Google aufgrund von Vorkommnissen für ein einzelnes Keyword ein sog. Flag setzt, das bewirkt, dass dieses Keyword eben nicht besser als z. B. Position 30 ranken darf. Liegt eine so geartete Strafe vor, muss man den Grund suchen und beheben. Anschließend stellt man einen Antrag auf Wiederaufnahme (*http://einfach.st/rqantrag*). Ist man jedoch von einem Update wie z. B. Penguin betroffen, muss ein solcher Antrag nicht gestellt werden bzw. ist nutzlos. Zwar muss man natürlich auch hier die (vermutlichen) Gründe für das plötzliche Abfallen im Ranking beseitigen, aber gute Rankings kommen dann von allein wieder. Ein Update verändert grundsätzlich die Art und Weise, wie Google Ergebnisse berechnet. Es handelt sich hier also nicht um eine „Strafe“, sondern schlicht um eine (auch zukünftig!) andere Kalkulation. Es könnte hier z. B. reichen, einige qualitativ hochwertige Links aufzubauen oder auch minderwertige Links (ja, auch das kann ein Grund für ein schlechteres Ranking durch Penguin sein, denn zu einem natürlichen Linkprofil gehören eben nicht nur Top-Links, sondern auch genügend Links, die in der Branche als „Backfill“ bezeichnet werden). Schon ändert sich die Bewertung (siehe Beispiele oben) und die Domain kann zu alter Stärke erblühen. Liegt hingegen eine echte Strafe vor, nützen in der Regel auch die besten Links nichts mehr: Gesetzte Flags verhindern generell ein gutes Ranking.

Wichtige Panda-Updates für den deutschsprachigen Raum

27.04.2102	Panda 3.6
19.04.2012	Panda 3.5
23.03.2012	Panda 3.4
26.02.2012	Panda 3.3
15.01.2012	Panda 3.2
18.11.2011	Panda 3.1
28.09.2011	Panda 2.5 (mit Rollback)
12.08.2011	Panda 2.4

Panda oder Penguin?

Ob Zufall oder wohlkalkuliert: Einige wenige Tage vor dem Rollout von Penguin (24. April) hatte Google auch das Panda-3.5-Update gestartet(19. April). Demensprechend schwierig war und ist es, Änderungen genau auf eines dieser beiden Updates zurückzuführen. Zum besseren Verständnis sei erwähnt, dass Google die Panda-Updates häufiger ausrollt. Die Version 3.5 war schon die zehnte Version, für den deutschsprachigen Raum die fünfte. Die Versionsnummern enthalten übrigens Sprünge, was möglicherweise darauf zurückzuführen ist, dass Google ihnen unterschiedliche Umfänge beimisst. Das erste Panda-Update 1.0, das allerdings nur US und UK betraf, war am 24.02.2011.

Während das Panda-Update aller Wahrscheinlichkeit nach für die Beseitigung von sog. „Thin-Content“ (also Seiten mit spärlichen, schlechten oder wenig Inhalten) zuständig ist, zielt Penguin auf überoptimierte Seiten ab. Viele Experten sind der Meinung, dass Panda vor allem auch das Verhalten von Usern mit einbezieht, die sich bekanntermaßen von unnützen, oft auch noch mit Werbung vollgepackten Seiten abwen-

> „Die Kenntnis, ob „Strafe“ oder „Updateopfer“ ist essenziell für die richtige Vorgehensweise!“

den. Hier genügt es also nicht, am Linkaufbau zu werkeln oder einfach frischen Content nachzuliefern. Die meisten Seiten, die von Panda betroffen waren oder sind, würden auch Besucher mit hoher Übereinstimmung aus dem Index herauswählen. Kurz vor Redaktionsschluss gab es am 25. Mai ein weiteres Penguin-Update 1.1. Hier hat Google nochmals nachgebessert und es heißt, wessen Domain sich jetzt nicht spürbar erholt hat, wird es möglicherweise auch nie mehr tun.

Bei Seiten, die von Penguin betroffen sind, ist die schnelle Erkennung per Augenschein nicht ganz so einfach. Hier sollte die gesamte Site, vor allem ihr Linkprofil, einem professionellen Audit unterzogen werden. Anschließend müssen alle erkannten unnatürlichen Signale entfernt bzw. entschärft werden. Das Erkennen, welches Update einen Absturz im Ranking verursacht hat, ist also essenziell für die Therapie der Site!

Wer übrigens glaubt, dass ihn das Penguin-Update zu Unrecht erwischt hat, kann hier (*http://einfach.st/peng1*) unter Angabe des Suchwortes ein Formular ausfüllen und das Ganze mit einem Kommentar als Feedback an Google senden.

Nachdenkliches zum Schluss

Wenn man es sich genau überlegt, ist es gar nicht so trivial, eine automatisierte Manipulationserkennung zu bauen, die tatsächlich a) zuverlässig arbeitet und b) nicht ständig Sites zu Unrecht in den Rankingabgrund reißt. Die Feinjustierung der jeweiligen Gewichtung und somit des Einflusses der vielen Hundert Drehknöpfe aufeinander ist sicherlich eine sehr diffizile Aufgabe, die viel Fingerspitzengefühl, viel Erfahrung und noch sehr viel mehr Tests auf ihre Wirkung erfordert. Die sog. Engineers bei Google sind um diese Verantwortung nicht zu beneiden, auch wenn man vielleicht ohne Kenntnis dieser Probleme denkt, man könne selbst sehr vieles besser. Viele Fußballfans sind ja auch fest überzeugt davon, dass sie persönlich der weitaus bessere Nationaltrainer wären.

Es ist und bleibt das alte Hase-Igel-Spielchen. Kurzfristig arbeitende Optimierer denken sich neue Manipulationsmöglichkeiten aus und haben damit meist auch eine gewisse Zeit Erfolg. An diesen Stellen zucken viel mit den Achseln und fragen sich, warum Google dagegen nichts unternimmt und sich scheinbar so träge verhält. Das ist aber in Wirklichkeit ein Trugschluss. Mit welchen Techniken die Black-Hat-SEO (die also unerlaubte Methoden verwenden) jeweils aktuell arbeiten, hat die Crew von Google in der Regel schon sehr bald auf dem Schirm. Nur schlägt man dort eben nicht manuell oder unbedacht zu. Man analysiert den Kern des Problems, verschafft sich einen Überblick über die Häufigkeit der Anwendung und entwickelt dann eine algorithmische Abwehrmaßnahme. Solche Maßnahmen werden in kleineren Änderungen beinahe täglich eingespielt, aber eben draußen nicht bemerkt. Wir sehen nur die großen Änderungen in Form vom Updates, die dann auch einen Namen bekommen.

Eines sollte man keinesfalls aus dem Auge verlieren: Google wird immer besser in der Erkennung von Gut und Böse. Selbstverständlich wird es immer Beispiele geben, wo sich in den Ergebnissen Seiten einschummeln, die dort definitiv nichts verloren haben. Das sind aber immer punktuelle Betrachtungen. „Merkt" man sich solch ein schiefes Ergebnis und googelt in ein paar Monaten danach, ist die betreffende Seite oft schon lange wieder verschwunden. Vielleicht die wichtigste Änderung gegenüber der Vergangenheit ist, dass Google nun sehr viel mehr vertrauenswürdige Signale vorliegen hat, also nur die Webseite selbst und die darauf zeigenden Links. Es fließt mehr und mehr ganz massiv das Userverhalten mit ein. Gut – auch das kann man mit entsprechendem (ständig steigendem) Aufwand als Suchmaschinenspammer manipulieren. Aber nun hat Google mit Google+ auch noch ein soziales Netzwerk und damit den sog. „Social Graph" in voller Breite für Analysen zur Verfügung. Damit sind sie in der Lage, echte User von künstlichen Accounts zu unterscheiden, wie ja in der letzten Website Boosting ausführlich erklärt wurde. Und jetzt wird es wirklich eng für die Manipulation von Userverhalten. Wer seine Arbeit als Suchmaschinenoptimierer wirklich ernst nimmt, hat mittlerweile erkannt, dass sein Job nicht mehr nur darin bestehen kann, technische Fehler beim Kunden auszubügeln oder gar künstlichen Linkaufbau zu betreiben. Er muss sich immer mehr wandeln zu einem echten Berater und mit dem Kunden gemeinsam langfristig tragfähige Strategien entwickeln. Hierzu zählt z. B., wertvollen Content zur Verfügung zu stellen (dann kommen Links von allein), eine echte, fachbezogene Reichweite in sozialen Netzwerken aufzubauen (dann erfährt das Fachpublikum jeweils eher von den nützlichen Dingen). Hatten wir das alle nicht schon öfter gehört? Ja, stimmt – aber eben weil das aufwendig ist und Mühe macht, geht man auch heute noch oft den vermeintlich einfacheren Weg der Manipulation. Der wird jedoch in Zukunft mit Sicherheit immer zeit- und kostenaufwendiger.

Damit soll als allerletzte Frage am Ende stehen, wann es günstiger wird, lieber gleich wirklich etwas an den Webseiten zu verändern, statt Geld in manipulative Techniken zu stecken. Denn eines ist wohl sicher: Irgendwann wird Google auch noch so gut Versteckte bemerken bzw. messen können und der plötzliche (!) Absturz ist umso schmerzhafter, aus je größerer Höhe man fällt.¶