Martin Rau

Screaming Frog Log File Analyser

Die Logfile-Analyse ist selbst für viele SEOs ein Buch mit sieben Siegeln, geschweige denn für normale Websitebetreiber. Dabei ist es so einfach, mit einem Tool wie dem Log File Analyser von Screaming Frog diese wichtigen Serverprotokolle zu untersuchen bzw. für SEO-Zwecke zu überwachen. Alles, was man benötigt, ist ein wenig Hintergrundwissen und eine Spur Neugierde.

Alle Erkenntnisse, die man aus einer Logfile-Analyse ziehen kann, basieren auf echten eigenen Server-Daten. Sie stellen somit die eindeutigsten Daten dar, mit denen ein SEO arbeiten kann. Neugierig geworden? Los gehts.

DER AUTOR

Martin Rau ist Technical SEO Manager bei fabfab. Zusätzlich unterstützt er als Freelancer seine Kunden bei der Optimierung ihrer Websites.

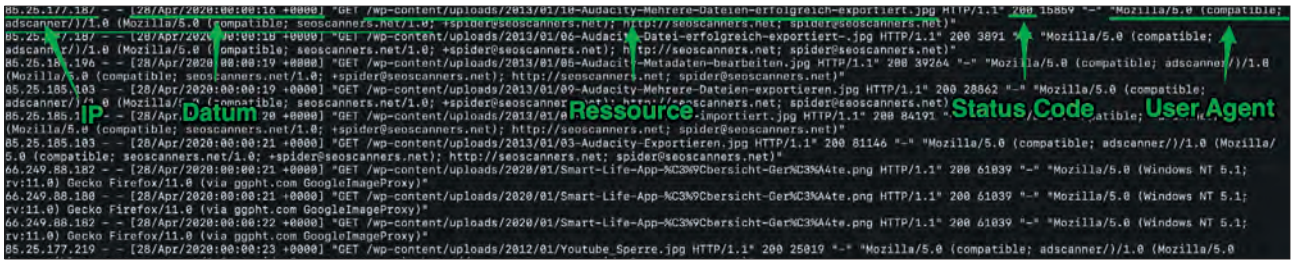

Abb. 1: Auszug aus einem Logfile meines Kunden my-digital-home.de – danke, Michael

Abb. 2: Es lohnt sich, die besonders großen Ausschläge zu analysieren

Mit einer Logfile-Analyse kann man Fragen auf den Grund gehen wie: Welche Seiten werden vom Googlebot besonders häufig gecrawlt? Welche Seiten crawlt der Bot nicht? Verändert sich die Anzahl an Zugriffen des Googlebots auf eine Ressource im zeitlichen Verlauf? Welche Fehler sehen Bots beim Crawlen meiner Webseite? Welche Bots crawlen meine Webseite? Für die Beantwortung solcher Fragen verwendet man am besten ein geeignetes Tool.

Eines dieser Tools ist der „Screaming Frog Log File Analyser“. Es wird von dem britischen Unternehmen Screaming Frog vertrieben. Die Software wird lokal installiert und hat damit Zugriff auf die volle Geschwindigkeit des Rechners. Die kostenpflichtige Lizenz wird für ein Jahr erworben. Nach Ablauf der Zeit muss sie erneut gekauft werden. Unter dem Hauptmenüpunkt Licence und darunter „Enter Licence“ kann man die verbleibende Gültigkeit einsehen. Der Preis für eine Lizenz beläuft sich auf umgerechnet 110 € für ein Jahr. In dieser Zeit kann das Tool ohne Einschränkungen genutzt werden. Für kleinere Webseiten gibt es zudem eine kostenlose Variante. Diese ist allerdings in der maximalen Anzahl an Events begrenzt. Der Log File Analyser läuft auf Windows-, Mac- und Linux-Betriebssystemen.

Stärken des Log File Analysers

Das wichtigste Feature des Log File Analysers ist die IP-Analyse von Events. Ohne diese wüsste man nicht, ob eine Anfrage wirklich vom Googlebot gekommen ist. Womöglich gibt sich ein anderer Crawler als etwas aus, was er nicht ist. Das Tool überprüft die IP jedes einzelnen Events und sortiert diese in „verified“ (überprüft) und „spoofed“ (gefälscht).

Eine weitere große Stärke des Tools ist, dass es mit großen Datenmengen klarkommt. Lädt man 50 Mio. Logfile-Zeilen ein, funktioniert das Programm immer noch einwandfrei. Zum Vergleich: Excel hat bereits Probleme ab ca. 1 Mio. Zeilen.

Das Tool ist sehr stark im Zusammenspiel mit einem Screaming Frog SEO Spider Crawl. Hat man seine Webseite mit dem SEO Spider vollständig gecrawlt, kann man diesen Crawl in den Logfile Analyser importieren. So erhält man einen Abgleich der Daten und kann weitere Erkenntnisse sammeln. Dazu später mehr.

TIPP

Wichtige Fachbegriffe im Tool:

- **IP:** Jeder Computer hat eine IP. Diese wird hier festgehalten. Die IP kann der Log File Analyser überprüfen und so herausfinden, ob ein angegebener User Agent korrekt war – mehr dazu im Abschnitt Analyse Beispiel 4: Gute vs. böse Bots.
- **Datum:** Jahr, Monat, Tag + Uhrzeit
- **Datei:** Welche Ressource wurde angefragt?
- **Status-Code:** Die wichtigsten Status-Codes sind: 200 – alles okay. 301 – die angefragte Ressource ist umgezogen, die neue Adresse der Ressource wurde übermittelt. 404 – die angefragte Ressource wurde nicht gefunden. 5xx – ein Serverfehler ist vorgefallen.
- **User Agent:** Ein User Agent kann bei jeder Anfrage beliebig gewählt werden. Das heißt, jeder kann sich z. B. als Googlebot ausgeben. Damit man in seiner Analyse keine falschen Rückschlüsse zieht, bietet der Log File Analyser ein einzigartiges Feature an – die Überprüfung von User Agents anhand der IPs. Mehr dazu im Bereich: Stärken des Log File Analysers

Was sind Logfiles?

Eine Webseite besteht aus vielen verschiedenen Dateien: HTML-Dokumenten, Bildern, JavaScript, Stylesheets usw. Jede dieser Ressourcen liegt auf einem Server. Ein Nutzer oder anderer Computer (Bot, Crawler) fragt eine einzelne Seite an, zum Beispiel

die Startseite von *https://seo-praxis.de*. Dann wird jede Ressource, die zum Anzeigen der Seite notwendig ist, beim Server angefragt. Das HTML-Dokument wird geladen. Darin enthalten sind Links zu Bildern, Javascripts u. v. m. Jede verlinkte Ressource wird anschließend ebenfalls angefragt. Jede dieser Anfragen wird im Logfile protokolliert mit Informationen zu: IP des Anfragestellers, aktuelles Datum mit Uhrzeit, Ressourcenpfad, Status-Code und der angegebene User Agent (siehe dazu Abbildung 1).

Speicherungszeit von Logfiles

In der Standardkonfiguration von Apache Servern sind Logfiles nur für die letzten sieben Tage gespeichert. Damit kann man keine weitreichenden Kenntnisse erlangen. Wer ernsthaft Logfiles analysieren möchte, sollte sich einen Zeitraum von drei bis sechs Monaten an Logfiles abspeichern. Hier gilt: je mehr, desto besser. Hierfür lohnt es sich, einen automatischen Job einzurichten. Ihr freundlicher System-Admin hilft ihnen hier gerne weiter. Bisher hat das Argument „Meine Analyseergebnisse unterstützen dich, dein System dauerhaft zu optimieren – und ich habe Schokolade dabei" in der Regel sehr gut funktioniert.

Hat man keinen Zugriff auf historische Logfiles, habe ich einen guten Tipp parat. Man öffnet die Google Search Console. In der Leiste klappt man den Bereich „Legacy Tools" aus. Unter „Crawl Stats" sieht man die Googlebot-Aktivitäten auf einer Domain. Diese Graphen schaut man sich regelmäßig jeden Montag an. Jetzt kann man darauf warten, dass ein großer Ausschlag entsteht. Dann besorgt man sich einmalig die Logfiles für die letzten sieben Tage und analysiert gezielt einen speziellen Google Crawl (siehe Abbildung 2).

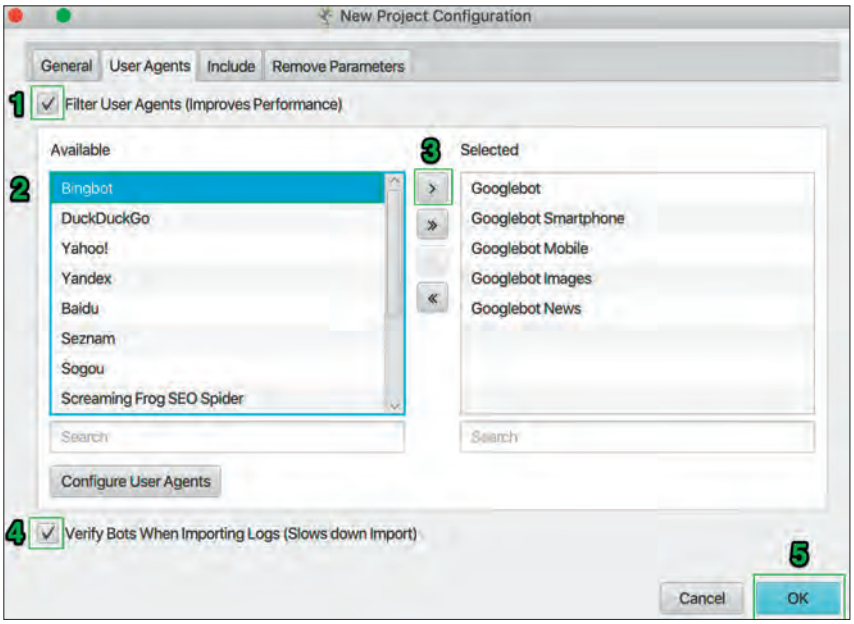

Abb. 3: Beim Anlegen eines neuen Projekts unbedingt „Verify Bots" auswählen

Mit welcher Dateigröße kann das Tool umgehen?

Screaming Frog gibt keine maximale Logfile-Größe an. Die Größe der zu importierenden Dateien wird maßgeblich von der Geschwindigkeit des Systems limitiert. Wird die Anzahl der Events zu groß, wird das Programm sehr schwerfällig. Häufige View-Wechsel benötigen mehr Zeit und das Arbeiten wird erschwert.

Tipp: Es dreht sich alles um die Anzahl der Events. Diese werden im Tool an der unteren Leiste angezeigt. Je mehr Events das Tool gespeichert hat, desto langsamer ist die Performance. Wenn die Performance zu langsam wird, kann man entweder den Zeitraum verkürzen oder bei der Filterung weniger User Agents benutzen. Dadurch verringert man die Anzahl an Events. Notfalls kann man das Projekt aufsplitten.

Wie kann man die ungefähre Größe von Logfiles abschätzen?

- **Blog** mit ca.: 5.000 Nutzern/Tag: gezipptes File: 5 MB
- **Mittelständisches Unternehmen** mit ca. 50.000 Nutzern/Tag: ca. 200 MB
- **Große Unternehmen** mit mehr als 50.000 Nutzern/Tag sollten ein spezielles Filtersystem implementieren, das nur die benötigten Daten abspeichert

Für wen ist das Tool geeignet?

Kurz gesagt, das Tool ist sehr einfach im Umgang und bietet eine hervorragende User Experience. Wirklich jeder kann damit einfach und schnell Logfiles analysieren.

Logfiles in das Tool zu laden und Analysen durchzuführen, ist einfach – der schwierige Part ist das tiefere Verständnis, was einem die Daten letztendlich sagen, und die daraus resultierende Bewertung des Problems.

In der Praxis hat sich herausgestellt, dass es schwierig sein kann, an Logfiles zu gelangen. Entwickler haben in der Regel keine Erfahrung mit dem Thema. Ein kurzer Leitfaden auf Englisch für die Kommunikation mit Entwicklern steht unter seo-praxis.de zur Verfügung. Mit dem Leitfaden im Gepäck kann man sich in das erste Gespräch wagen und gemeinsam eine Lösung erarbeiten.

Gerade für größere Unternehmen ist es oftmals schwierig, Logfiles vorzuhalten. Je mehr Nutzer auf einer Webseite unterwegs sind, desto mehr Events entstehen in den Logfiles. Man bewegt sich sehr schnell im zwei- bis dreistelligen

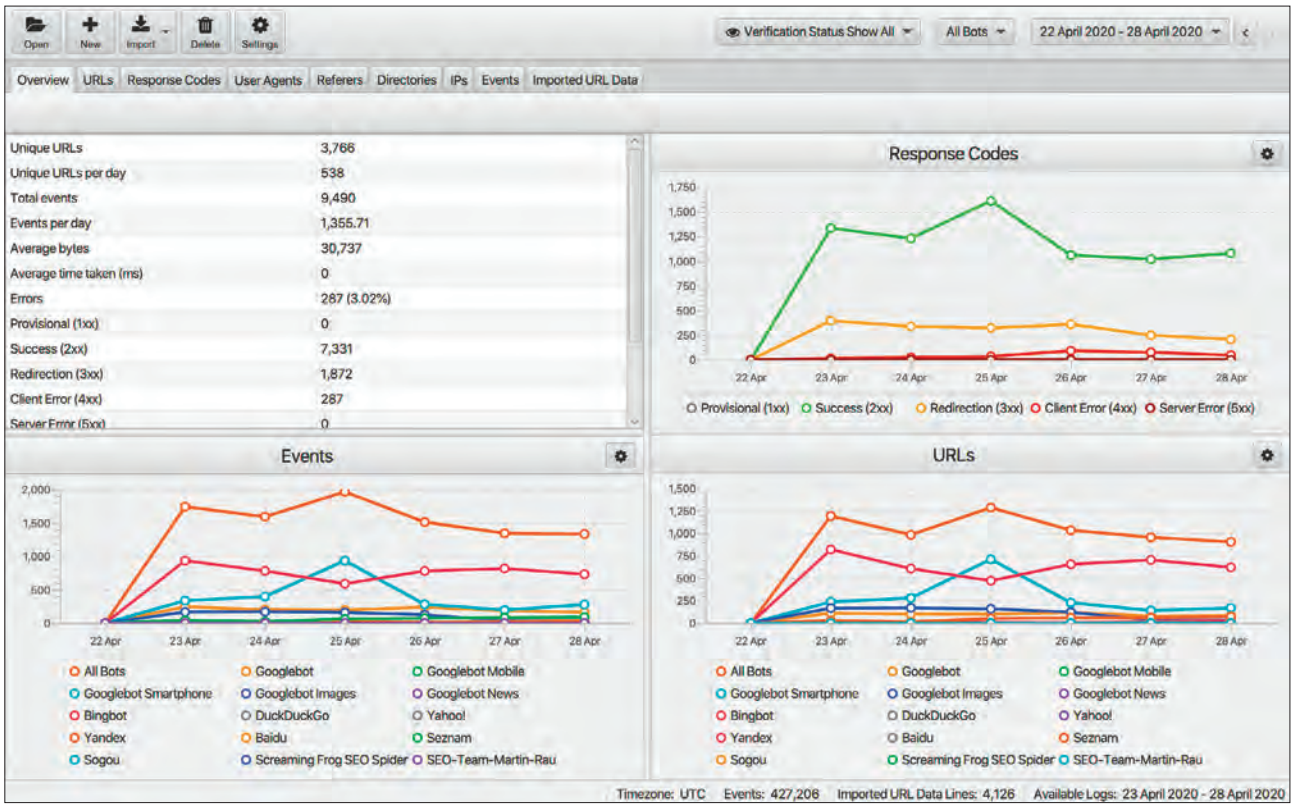

Abb. 5: Im Reiter „URLs“ kann man sich die nicht gecrawlten URLs anzeigen lassen

GB-Bereich pro Tag. Dadurch wird das dauerhafte Abspeichern von Logfiles über einen größeren Zeitraum zu einem echten Kostenfaktor. Zusätzlich werden Systeme in dieser Größenordnung sehr komplex. Das bedeutet, man muss mithilfe von Entwicklern Systeme erschaffen, die die Logfiles an geeigneter Stelle gefiltert abspeichern. Der SEO benötigt zudem Zugriff auf den Ort. Gerade bei großen Konzernen wird deshalb leider auf die Logfile-Analyse verzichtet. Dadurch entgeht ein sehr wichtiges Puzzleteil in der SEO-Analyse.

Um den Einstieg in den Screaming Frog Log File Analyser so einfach wie möglich zu gestalten, habe ich im späteren Verlauf einige Beispiele für die ersten Schritte mitgebracht. Wer mehr Anwendungsbeispiele ausprobieren möchte, den lade ich auf meine Webseite ein. Dort gibt es weitere Anleitungen.

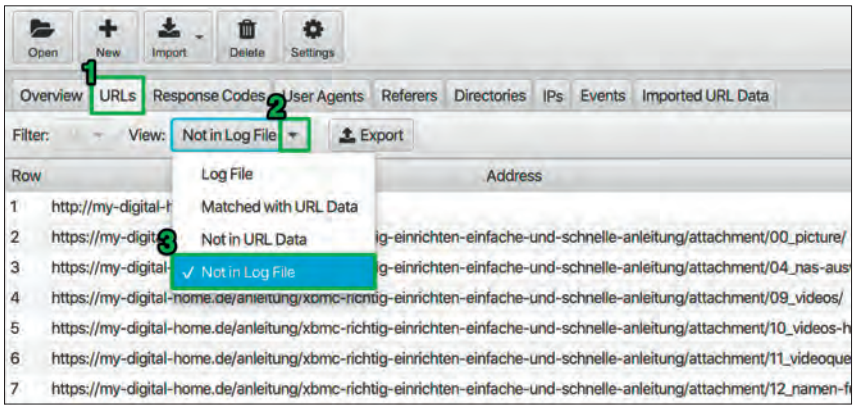

Abb. 5: Im Reiter „URLs“ kann man sich die nicht gecrawlten URLs anzeigen lassen

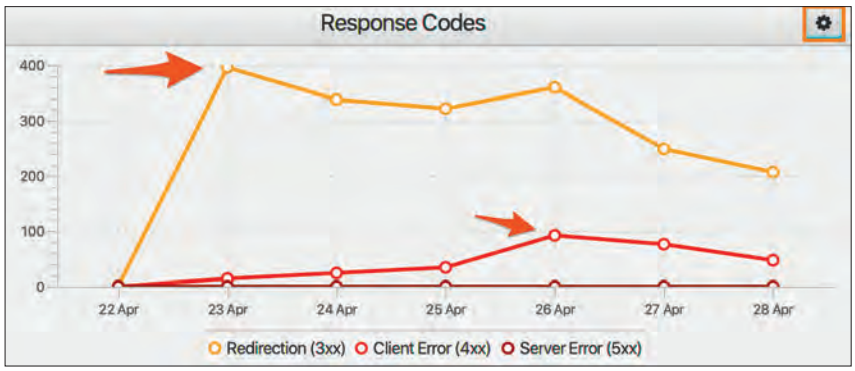

Abb. 6: In der Übersicht der Status-Codes lassen sich schnell Peaks erkennen

Aufbau und Funktionsweise

Vor dem Erstellen eines Projekts sollte man sich Folgendes überlegen:

1. Wie groß sind die vorliegenden Logfiles?
2. Welchen Zeitraum möchte man analysieren?
3. Möchte man alle User Agents überprüfen oder nur spezifische?

Als Erstes kann man ein Testprojekt anlegen. Hier wählt man alle User Agents aus und lässt diese verifizieren. Das ist besonders für die spätere Analyse wichtig (siehe Abbildung 3). Als Nächstes importiert man alle Logfiles. Hierfür können die Logfiles im gezippten Zustand einfach auf die Arbeitsfläche per Drag-and-drop gezogen werden. Alternativ kann man auch den Menübutton „importieren“ benutzen. Der Import kann je nach Größe einige Minuten dauern. Die angegebene Zeit ist nicht realistisch. Keine Sorge.

Wenn die URLs in den Logfiles relativ vorliegen, also ohne das Netzwerkprotokoll oder etwaige Domainangaben, dann beschwert sich das Programm. Im Pop-up gibt man einfach seine Domain ein, z. B.: *https://www.meine-seite.de*.

Das Programm analysiert die Logfiles und verifiziert die IPs der User Agents. Anschließend landet man in der Übersicht, wie in Abbildung 3 zu sehen.

Tipp: Benötigt das Programm nun sehr lange für View-Wechsel, sollte man seine Analyse einschränken. Hierfür gibt es zwei große Hebel. Entweder begrenzt man den zu analysierenden Zeitraum oder man wählt nur wenige, spezifische User Agents aus. Beide Faktoren haben einen Einfluss auf die Gesamtanzahl der Events. Generell gilt: Viele Events machen das Tool langsamer.

Für das Beispiel bekam ich Logfiles meines Kunden zur Verfügung gestellt. Hier wurden alle User Agents ausgewählt, der Filter auf „Verified“ gesetzt und insgesamt Logfiles für sechs Tage importiert.

Man sieht nun vier Bereiche. Oben links dient der allgemeinen Übersicht. Oben rechts befinden sich die Status-Codes als Graph dargestellt. Über den Einstellungsbutton kann man die Übersichten weiter filtern. Unten links gibt es den Events-Bereich. Hier sieht man eine Übersicht aller User Agents. Unten rechts sieht man die Anzahl an URLs, die an dem Tag angefragt wurden.

Am 25. April hat der Googlebot Smartphone einen Peak bei den Events. Gleichzeitig sieht man am 25. April eine erhöhte Anzahl an 4xx-Fehlern. Dies schauen wir uns im Bereich Analyse Beispiel 2: Status-Codes analysieren näher an.

Logfile-Daten mit weiteren Daten anreichern

Der Log File Analyser bietet die Funktion an, URLs aus anderen Quellen zu importieren. Der Clou ist, dass diese mit den URLs aus den Logfiles abgeglichen werden. Somit erhält man eine Übersicht, welche URLs im analysierten Zeitraum noch nicht gecrawlt wurden.

Um weitere URLs hinzuzufügen, wählt man den Bereich „Imported URL Data” aus. Hier können URLs im Format CSV oder Excel importiert werden (CSV

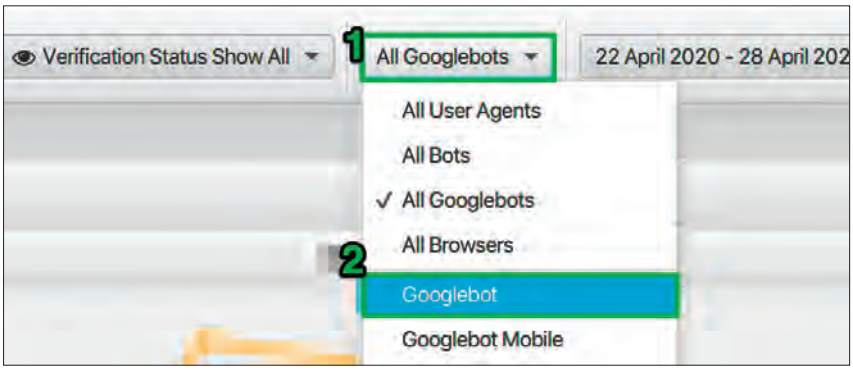

Abb. 7: Mithilfe von Filtern lassen sich einzelne User Agents auswählen

comma separated values). Möchte man seine Sitemaps importieren, müssen diese zunächst aus der XML-Datei in eine Excel-Datei kopiert werden. Man kann ebenfalls seine „Top Pages" aus anderen SEO-Tools exportieren, in eine Excel-Datei einfügen und anschließend importieren. Hat man seine Domain mit dem SEO Spider vollständig gecrawlt, kann man den „Internal"-Tab aus dem SEO Spider exportieren. Die CSV lässt sich dann in den Log File Analyser importieren.

Analyse Beispiel 1: Differenz zwischen vorhandenen und gecrawlten URLs herausfinden

Hat man die Logfiles und den SEO Spider Crawl erfolgreich importiert, kann nun die Analyse starten.

Die Fragestellung lautet: Welche URLs wurden in dem Zeitraum der Logfiles noch nicht gecrawlt?

Dazu wählt man den Bereich „URLs" aus der Hauptnavigation aus. Als Nächstes wählt man unter „View" den Filter „Not in Log File" aus (siehe Abbildung 5).

Nun sieht man alle URLs, die zwar durch den SEO Spider entdeckt oder über andere Datenquellen importiert wurden, aber noch nicht in den Logfiles auftauchen, dementsprechend noch nicht gecrawlt wurden. Tauchen in dieser Liste URLs auf, die relevant für die Seite sind, sollte man unbedingt tiefer in die Analyse gehen. Die einfachste Antwort wäre, dass der Zeitraum an Logfiles zu klein gewählt wurde und dass die Seiten zuletzt vor Monaten gecrawlt wurden und deshalb nicht in den Logfiles auftauchen. Für die wichtigsten Seiten kann man sicherheitshal-

ber eine manuelle Site-Abfrage über die Google Search Console starten, um zu überprüfen, ob die Seiten indexiert und crawlbar sind.

Analyse Beispiel 2: Status-Codes analysieren

Wenn Suchmaschinen eins nicht mögen, dann Seiten im Index zu haben, die einen 404-Error-Code anzeigen. Das größte Augenmerk bei einer Logfile-Analyse sollte man demnach auf 4xx- und 5xx-Fehler legen. Wenn die Suchmaschine eine gehäufte Anzahl an 5xx-Fehlern bemerkt, drosselt sie die Crawl-Geschwindigkeit. Damit sinkt das Crawlbudget. (Mehr zum Thema Crawlbudget im Blogpost von Google: *http://einfach.st/crawlbudget*.)

Je niedriger das Crawlbudget, desto weniger Seiten crawlt Google und desto weniger Seiten werden indexiert. Je weniger Seiten indexiert sind, umso weniger Seiten können ranken und organischen Traffic auf die Seite bringen.

Am 26. April verzeichnete die Domain eine erhöhte Anzahl an 4xx-Fehlern (siehe Abbildung 6). Schauen wir uns das genauer an.

Man wechselt in den Bereich „User Agents". Dort stellt man das Datum auf den 26. April ein. Man wählt „All Bots" und „Verified" aus. In den Spalten sortiert man die Reihe mit 4xx absteigend. Jetzt sieht man, welcher Bot die Fehler ausgelöst hat, in dem Fall war es der Googlebot.

Für die weitere Analyse filtert man mit obigem Filter nun auf den besagten Googlebot (siehe Abbildung 7).

Als Nächstes wechselt man in den Bereich „Response Codes". Oben wählt man den Filter „Client Error 4xx" und wählt das Häkchen bei „Last Response" aus (Last Response zeigt nur den zuletzt gemessenen Wert für die URL an).

Nun erhält man eine Liste an URLs, die der Bot im genannten Zeitraum angefragt hat und die einen 4xx-Status-Code ausgegeben haben. Selektiert man eine URL, kann man im unteren Bereich den Tab „Events" auswählen. Nun sieht man Events zu dieser URL und ob sich der Status-Code im Laufe der Zeit geändert hat.

Als Nächstes kann man den Ursachen hierfür im System auf den Grund gehen. Im Falle meines Kunden sind die 404-Fehler aus gelöschten JavaScripts entstanden. Hier wird man den Server noch besser konfigurieren müssen. Für gelöschte Ressourcen sollte der Status-Code 410 ausgegeben werden.

Dieses Vorgehen lässt sich auf alle 3xx-, 4xx- und 5xx-Status-Codes anwenden.

Analyse Beispiel 3: Welche URLs sind für die Suchmaschine am wichtigsten?

Die URLs, die besonders häufig von einer Suchmaschine gecrawlt werden, sind für sie am wichtigsten. Soweit die These. Die Häufigkeit wird durch die Art des Contents ebenfalls beeinträchtigt. Ist es informatorisch orientiert – Evergreen Content – dann wird die Suchmaschine nicht so häufig vorbeikommen wie wenn es sich um brandaktuellen Newscontent handelt.

Als Erstes startet man wie immer mit der Auswahl seiner Zeitspanne. Als Nächstes wählt man die „Verified"-Bots aus, die man betrachten möchte. Man wählt den Tab „URLs" aus. Dort setzt man den Filter auf „HTML", View „Logfile".

Man erhält eine Liste an URLs. Scrollt man nach rechts, entdeckt man den Bereich „Num Events". Weiter rechts sind die Anzahl an Events für jeden Bot einzeln aufgelistet. Diese Liste sortiert man sich am besten absteigend nach „Num Events".

Jetzt sieht man, welche Seiten in dem Zeitraum am meisten Events haben. Doch Vorsicht! Angenommen, man hat den Filter nicht explizit auf „HTML" gesetzt, sondern auf „All", dann wird die Liste ein anderes Ergebnis haben.

Tipp: Wählt man im Filter „Images" aus, kann man über die Spalte „Average Bytes" besonders große Bilder ausfindig machen.

Analyse Beispiel 4: Gute vs. böse Bots

Möchte man wissen, wie stark man von fremden Bots gecrawlt wird, fragt man entweder sein IT-Security Team – oder man analysiert Logfiles. Hierzu wählt man den Bereich „User Agents" aus und stellt einen möglichst großen Zeitraum ein. Anschließend wählt man „All Bots" und „Verified" aus.

Nun addiert man alle Events. Angenommen, die Summe beträgt 100.

Nun filtert man die „Spoofed" User Agents und addiert wieder die Events. Angenommen, die Summe beträgt 25. 25 / 100 = 0,25. Das bedeutet, der Anteil an „bösen Bots" beträgt für diesen Zeitraum ca. 25 %. Mit diesem Wert in der Tasche kann man sich überlegen, ob es sinnvoll ist, Gegenmaßnahmen einzuleiten.

Fazit

Zusammenfassend kann man sagen, dass der Log File Analyser unschlagbare Funktionen zu einem sehr attraktiven Preis mitbringt. Hat man sich an die Imports gewöhnt, macht es Spaß, die Logfiles in übersichtlicher Form zu analysieren und Erkenntnisse zu sammeln, die einem sonst verborgen geblieben wären. Eine automatisierte Abspeicherung der Logfiles sollte stets angestrebt werden. Über die Zeit legt man sich einen unschätzbaren Datenschatz an. Die interne IT kann einem hier auf jeden Fall weiterhelfen. Logfiles analysieren zu können, sollte im Handwerkskoffer eines jeden SEOs sein. Wer mehr wissen möchte, dem lege ich die Dokumentation (EN) ans Herz: *https://www.screamingfrog.co.uk/log-file-analyser/user-guide/*. ¶