Engineer

Crawler

🕷 ein Programm, das Webseiten zur Indexierung abruft

Was es ist

Es gibt kein zentrales Verzeichnis aller Webseiten. Eine Suchmaschine muss sie deshalb selbst finden, und das dafür ausgesandte Programm ist ein Crawler. Er beginnt mit bereits bekannten Adressen, ruft jede über HTTP ab, liest die Links in der Antwort und nimmt neue Adressen in seine Liste auf. Googles Leitfaden zur Funktionsweise der Suche nennt weitere gebräuchliche Namen: „auch als Crawler, Robot, Bot oder Spider bekannt“. RFC 9309, der Standard zum Ausschluss von Robots, beschreibt Crawler als automatisierte Clients, die von Suchmaschinen betrieben werden, „um Links zur Indexierung rekursiv zu durchlaufen“.

Links sind der wichtigste Zugangsweg. Google sagt, Googlebot finde neue Adressen „hauptsächlich über Links, die in zuvor gecrawlte Seiten eingebettet sind“. Der Leitfaden nennt den vertrauten Fall einer Kategorieseite, die auf einen neuen Beitrag verlinkt. Eine Sitemap ist der zweite Weg: eine Liste, die der Websitebetreiber direkt übergibt. Eine Seite ohne beides gibt einem Crawler keine Möglichkeit, von ihrer Existenz zu erfahren.

Der Abruf ist die erste der drei Stufen, die Google beschreibt. Crawling lädt Text, Bilder und Videos einer Seite herunter. Indexierung analysiert das Heruntergeladene und speichert es. Die Auslieferung beantwortet eine Suchanfrage aus diesem Index. Beim Crawlen rendert Google die Seite auch und führt ihr JavaScript in einer aktuellen Chrome-Version aus, weil es ohne diesen Schritt „diese Inhalte möglicherweise nicht sehen würde“. Keine Stufe ist zugesichert: Google „garantiert nicht, dass Ihre Seite gecrawlt, indexiert oder ausgeliefert wird“, und „nimmt keine Zahlungen an, um eine Website häufiger zu crawlen“.

Ein Crawler soll außerdem sagen, wer er ist. RFC 9309 verlangt für jeden Crawler einen eigenen Namen, ein Produkttoken, das nur aus Buchstaben, Unterstrichen und Bindestrichen besteht. Laut RFC sollte dieses Token in der vom Crawler gesendeten Identifikationszeichenfolge stehen, bei HTTP also im User-Agent-Header. Die Zeichenfolge sollte den Zweck des Crawlers beschreiben. Der Beispielheader des RFC enthält `ExampleBot/0.1` und einen Link auf eine Seite über den Crawler. Die dazu passende Zeile in robots.txt lautet `user-agent: ExampleBot` . Der Name ist eine Behauptung, kein Berechtigungsnachweis, denn jedes Programm kann ihn senden.

Nicht jede automatisierte Anfrage von Google ist ein Crawl. Die Übersicht teilt seine Clients in drei Arten ein: gewöhnliche Crawler wie Googlebot, die „bei automatischen Crawls immer die robots.txt-Regeln beachten“; Spezialcrawler, die aufgrund einer Vereinbarung mit der Website arbeiten und diese Regeln beachten können oder auch nicht; sowie vom Nutzer ausgelöste Abrufprogramme, die auf dessen Wunsch meist eine einzelne Anfrage stellen und robots.txt ignorieren.

Das Konzept für eine Entscheidung nutzen

Nutzen Sie „Was ist ein Crawler?“ für eine konkrete Entscheidung. Wir können helfen, das Konzept mit dem System in Verbindung zu bringen, das Sie entwickeln oder einkaufen. Eine Website oder App, die die KI-Assistenten Ihrer Kunden finden, lesen und bedienen können — bei jeder Veröffentlichung geprüft statt einmal versprochen.

Unsere Leistungen entdecken Kontakt aufnehmen →

Die Annahmen hinter dem Begriff prüfen

Nutzen Sie die Erklärung und die verlinkten Arbeiten, um Fragen für Ihr Projekt zu formulieren. Wir können die Annahmen gemeinsam mit Ihnen prüfen, bevor ein Werkzeug ausgewählt wird.
Hugo-Website mehrsprachig in 3 Sprachen aufgebaut — 108 Vorlagendateien für drei Sprachen und vier Formate

Hugo-Website mehrsprachig in 3 Sprachen aufgebaut Kontakt aufnehmen →

Warum es wichtig ist

Ein Crawler entscheidet selbst, wann er wiederkommt. Sein Verhalten bestimmt einen Teil der Kosten dafür, gefunden zu werden. Google sagt, Googlebot solle auf den meisten Websites „im Durchschnitt nicht öfter als einmal alle paar Sekunden auf Ihre Website zugreifen“. Seine Crawler reduzieren ihre Aktivität, wenn eine Website Schwierigkeiten hat. Im Suchleitfaden bedeuten HTTP-500-Fehler: „langsamer werden“. Googles Leitfaden zum Crawl-Budget beschreibt den gesamten Mechanismus. Er richtet sich aber an Websites mit ungefähr einer Million wöchentlich veränderten Seiten oder zehntausend täglich veränderten Seiten. Laut Google braucht eine Website ihn nicht, wenn ihre Seiten am Tag der Veröffentlichung gecrawlt werden.

Ein Crawler liest außerdem nur einen begrenzten Teil jeder Datei. Googlebot liest beim Crawlen für die Suche die ersten 2 MB jeder Datei eines unterstützten Typs und die ersten 64 MB einer PDF-Datei, gemessen vor der Komprimierung. Den Rest ignoriert er.

Der Schwachpunkt ist der Name. Google warnt, dass der von Googlebot gesendete User-Agent-Header „oft von anderen Crawlern gefälscht wird“. Eine Logzeile mit dem Namen Googlebot beweist deshalb allein nichts. Google dokumentiert zwei Prüfungen. Für eine einzelne Adresse führen Sie eine umgekehrte DNS-Abfrage aus und prüfen, ob der Hostname auf `googlebot.com`, `google.com` oder `googleusercontent.com` endet. Lösen Sie diesen Namen anschließend vorwärts auf und prüfen Sie, ob dieselbe Adresse zurückkommt. Googles eigenes Beispiel:

```
host 66.249.66.1
1.66.249.66.in-addr.arpa domain name pointer crawl-66-249-66-1.googlebot.com.
host crawl-66-249-66-1.googlebot.com
crawl-66-249-66-1.googlebot.com has address 66.249.66.1
```

Für viele Adressen zugleich vergleichen Sie jede mit den von Google als JSON-Dateien in CIDR-Notation veröffentlichten IP-Bereichen. Es gibt eine Datei pro Clientart. Beide Prüfungen stützen sich auf das Netzwerk, aus dem eine Anfrage kam. Der Absender kann es nicht einfach in einen Header hineinschreiben.

Wie es in der Praxis aussieht

Der Server dieser Website führt ein Zugriffslog. Die Datenschutzseite erklärt, wie darin Crawler unterschieden werden. Jede Zeile hält den User Agent fest, „womit wir Menschen von Suchmaschinen-Crawlern unterscheiden“. Danach folgt ein zweiter Schritt: „Ein Crawler, der einen Namen nennt, wird anhand der von seinem Betreiber veröffentlichten Netzwerke geprüft, mit der unten beschriebenen vergröberten Adresse.“ Das ist Googles zweite Methode, und sie passt zu den gespeicherten Logdaten. Der letzte Teil jeder IP-Adresse wird durch Nullen ersetzt, bevor die Zeile auf der Festplatte landet. Die vollständige Adresse, die eine umgekehrte DNS-Abfrage benötigt, wird daher nie gespeichert.

Die Zusammenfassungen behandeln Crawler als eigenes Publikum: „dasselbe für alles Automatisierte“ neben der für menschliche Besuche genannten Aufbewahrung von dreißig Tagen, eine tägliche Zählung von „so vielen gelesenen

Seiten, so vielen davon durch Crawler, so vielen Aufrufen der Kontaktseite“ und eine monatliche Rangliste, „welche Crawler und automatisierten Scanner vorbeikamen“. Ein Crawler, der sich ehrlich benennt, wird unter seinem eigenen Namen gezählt. Ein bekannter Name aus einem Netzwerk, das sein Betreiber nicht veröffentlicht, besteht die Prüfung nicht.

Die Seite zur Vorbereitung auf KI-Agenten beginnt beim selben Leser: „Ein Agent beginnt dort, wo ein Crawler beginnt“, bei robots.txt, Sitemap und Suchindizes. Ihre Antwort lautet, jeden dieser Wege bewusst zu gestalten, bis hin zu einer IndexNow-Mitteilung bei jeder Bereitstellung. Teilnehmende Suchmaschinen „erfahren so am Veröffentlichungstag von einer neuen Seite statt erst bei ihrem nächsten Besuch“.

Datumsangaben sind eine Möglichkeit, mit einem Crawler zu sprechen. Sie können aber auch nichts Nützliches aussagen. Der Bericht über eine Sitemap, die 172 ihrer 176 Seiten denselben Zeitstempel gab macht die Bedeutung deutlich: „Ein Änderungsdatum verspricht einem Crawler, dass sich der Inhalt einer Seite geändert hat.“ Eine Website, die dieses Versprechen für alle Seiten zugleich gibt, „sagt entweder die Wahrheit über eine vollständige Überarbeitung oder sagt niemandem etwas Nützliches“.

Wo es schiefgeht

Anfragen mit dem Namen Googlebot kommen nicht von Google. Das Log zeigt viel Traffic unter Googlebots Namen. Der Betreiber lässt ihn entweder als willkommenen Gast durch oder sperrt den Namen ganz. Weil andere Crawler diesen Namen häufig kopieren, ist keine der beiden Reaktionen sicher: Die erste vertraut einem Hochstapler, die zweite sperrt zusammen mit ihm den echten Crawler aus. Prüfen Sie das Quellnetzwerk, bei einer Adresse mit umgekehrter und anschließender vorwärts gerichteter DNS-Abfrage, bei vielen anhand der von Google veröffentlichten Bereiche. Entscheiden Sie danach, nicht nach dem Namen.

Eine veröffentlichte Seite taucht nie in einem Crawl auf. Nichts verlinkt auf sie, oder sie ist erst nach einer Anmeldung erreichbar. Ohne zentrales Verzeichnis erfährt ein Crawler durch Links und Sitemaps von einer Adresse. Googles Leitfaden nennt Seiten hinter einer Anmeldung ausdrücklich als für seinen Crawler unerreichbar. Verlinken Sie die Seite von einer bereits gecrawlten Stelle, führen Sie sie in der Sitemap auf und halten Sie alles, was für die Suche bestimmt ist, außerhalb des Anmeldebereichs.

Nach einer schwierigen Phase des Servers sinkt die Crawl-Rate. Täglich werden weniger Seiten abgerufen, und Änderungen werden später bemerkt. Langsame Antworten, 5xx-Fehler und 429-Antworten senken alle Googles Crawl-Kapazitätsgrenze. Sie steigt erst wieder, wenn Antworten stabil und schnell werden. Alle Google-Crawler auf dem Host teilen sich den verbleibenden Spielraum. Beheben Sie zuerst Fehler und Latenz. Bei einer Website mit einigen Tausend Seiten ist Googles Crawl-Budget-Leitfaden nicht der richtige Ausgangspunkt.

Eine unverlinkte Seite erscheint in den Anfragen eines Crawlers. Ihre Adresse wurde nie veröffentlicht, vielleicht handelt es sich um einen privaten Entwurf oder eine Testkopie. Trotzdem erscheint sie unter dem Namen eines Crawlers im Zugriffslog. Google nennt es „fast unmöglich, eine Website allein dadurch geheim zu halten, dass keine Links darauf veröffentlicht werden“. Ein Klick von dieser Seite auf eine andere Website kann ihre Adresse in einem später veröffentlichten Referrer-Log hinterlassen. Schützen Sie alles Private mit einem Passwort. Google nennt das als Weg, sowohl Crawler als auch Menschen fernzuhalten.

Weiterführende Lektüre

- RFC 9309: Robots Exclusion Protocol — die Sicht des Standards auf Crawler: Produkttokens, User-Agent-Zeichenfolge und die Regeln, um deren Einhaltung Crawler gebeten werden.
- Ausführlicher Leitfaden zur Funktionsweise der Google Suche — Entdeckung, Crawling, Rendering und Indexierung, wie Google sie beschreibt.
- Übersicht über Googles Crawler und Abrufprogramme (User Agents) — die drei Google-Clientarten, Dateigrößenbegrenzung, Protokolle und Caching.
- Anfragen von Google-Crawlern und Abrufprogrammen prüfen — DNS-Abfragen und veröffentlichte IP-Bereiche mit ausgearbeiteten Beispielen.
- Das Crawl-Budget optimieren — wie Google seine Crawl-Rate festlegt und welche Websites sie abstimmen müssen.
- Googlebot — Smartphone- und Desktopcrawler, Besuchshäufigkeit, Dateigrößengrenze und warum eine unverlinkte Seite nicht geheim ist.

Glossar

Für Ihre tatsächlichen Rahmenbedingungen auswählen

Beschreiben Sie den Arbeitsablauf, die vorhandenen Schnittstellen und die Nutzer, die die Veränderung benötigen. Wir können eine prüfbare Lieferung definieren und die Übergabe so gestalten, dass Ihr Team die gelieferten Bestandteile warten kann.

Preise ansehen Kontakt aufnehmen →

Die Begriffe praktisch nutzbar machen

Fragen Sie, wie „Was ist ein Crawler?“ auf Ihre Organisation zutrifft. Ein konkretes Beispiel aus Ihrem bestehenden System reicht aus, um das Gespräch zu beginnen.

Unsere Leistungen entdecken

welcome@engineer.company +45 50 10 61 84 Kopenhagen, Dänemark

Nehmen Sie diese Seite mit

Kontaktkarte

Mit Sorgfalt und Liebe gestaltet von D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Copyright © 2025 – heute · Engineer ApS · Alle Rechte vorbehalten.

Gedruckt von https://engineer.company/de/glossary/crawler/