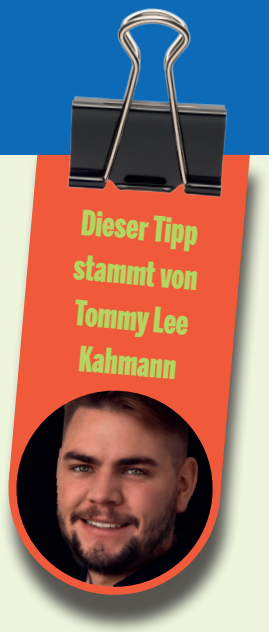

WAS VERBIRGT SICH HINTER DEM „CRAWL-DELAY“?

Crawling gehört in der Suchmaschinenoptimierung immer dazu. Ohne einen vernünftigen Crawler könnten wohl die wenigsten SEOs ordentlich arbeiten. Wenn man dann aber den Screaming Frog startet und der Crawl einfach nicht enden will und mehr als zwei Drittel der URLs mit dem Status-Code-5xx-Fehler antworten, fragt man sich schnell, woran dies liegt. In meinem bestimmten Beispiel hat es sich um einen Crawl-Delay in der robots.txt gehandelt. Was dies für den Check bedeutet, erläutere ich in dem folgenden Artikel genauer.

Wie wichtig ist eine robots.txt?

Aber fangen wir erst einmal ganz von vorne an. Was ist überhaupt eine robots.txt? Die robots.txt ist eine Datei im Hauptverzeichnis einer Website, die den Bots einer Suchmaschine und den verschiedenen Crawlern Anweisungen gibt, welche Seiten oder Bereiche der Website nicht gecrawlt werden sollen. Sie ermöglicht Website-Betreibern zu steuern, welche Inhalte von Suchmaschinen aufgenommen oder ausgeschlossen werden sollen. Dabei ist die robots.txt aus mehreren Gründen wichtig:

1. Kontrolle über das Crawling: Mit der robots.txt können Website-Betreiber steuern, auf welche Bereiche ihrer Website Suchmaschinen zugreifen dürfen. Dies kann helfen, sicherzustellen, dass wichtige Seiten gecrawlt und gefunden werden, während irrelevante oder sensible Seiten ausgeschlossen werden.
2. Ressourceneffizienz: Durch das Ausschließen bestimmter Bereiche oder Dateitypen kann der Website-Betreiber sicherstellen, dass die Bots nicht unnötige Ressourcen verwenden.
3. Einblicke in Crawler-Aktivitäten: Über das „Crawl-Delay“-Direktiv in der robots.txt können Website-Betreiber die Crawling-Rate bestimmter Suchmaschinen steuern, um Serverüberlastungen zu vermeiden.

Und beim dritten Punkt sind wir schon bei meinem Problem. Insgesamt lief der Crawl für knapp eine Million URLs über eine Woche durchgehend und ich konnte mir nicht erklären, wieso zwei Drittel meiner URLs einen 5xx-Statuscode zurückgespielt haben. Dass es sich um einen Crawl-Delay handeln könnte, habe ich zu dem Zeitpunkt aber noch nicht vermutet.

Was macht der Crawl-Delay?

Nach den ganzen Basis-Checks, die man als SEO macht, habe ich immer die Sitemap und danach die robots.txt angeschaut und habe etwas gefunden, was ich vorher noch nie gesehen hatte. Unter der Zeile „User-agent:*“, die für eine robots.txt eigentlich der Normalfall ist, stand der Crawl-Delay (siehe Abb. 1).

Da ich bis zu diesem Zeitpunkt noch gar nicht mit dem Thema „Crawl-Delay“ zu tun hatte, musste ich mich erst einmal damit befassen und wusste bis zu diesem Zeitpunkt immer noch nicht, dass ich damit auf das Problem meiner 5xx-Fehler gestoßen bin.

Der Crawl-Delay in der robots.txt gibt nämlich an, wie viele Sekunden ein Web-Crawler zwischen einzelnen Anfragen warten sollte, wenn er eine Website durchsucht. Genutzt wird dies, um die Serverlast zu reduzieren, indem man das Crawling-Tempo von Bots steuert. Verhindern soll dies, dass Bots die Website zu schnell durchsuchen. In diesem Fall hat es also bedeutet, dass der Crawler zwei Sekunden zwischen den Anfragen warten sollte, bevor er eine weitere Seite der Website abruft. Da die Geschwindigkeit des Crawlers allerdings nicht eingestellt war und dieser weniger als zwei Sekunden mit den Abfragen gewartet hat, kam immer wieder der Statuscode 5xx vom Server zurück,

```
User-agent: *
Crawl-delay: 2

Sitemap: https://www.          /sitemapindex.xml
```

Abb.1: Die robots.txt des besagten Kunden

wenn zwischen diesen zwei Sekunden Abfragen gestartet wurden.

Hier ist allerdings noch zu ergänzen, dass eine effiziente Crawler-Steuerung maßgeblich von anderen Dingen, besonders bei großen Domains, abhängt. Der Crawl-Delay ist nur eine von vielen Möglichkeiten, um die Crawlbarkeit einer Seite zu optimieren. Andere Faktoren wie beispielsweise eine saubere Seitenstruktur, Seiten mit „noindex-Tags" und Canonicals tragen ebenso einen Anteil zur Crawlbarkeit bei.

Vorteile des Crawl-Delays

Website-Betreiber müssen daher abwägen, ob sie einen Crawl-Delay in der robots.txt haben wollen oder nicht. Doch welche Vorteile bietet dieser überhaupt?

1. Reduzierung der Serverlast: Durch das Limitieren der Crawler-Anfragen kann der Server entlastet werden, wodurch die Performance für echte Nutzer optimiert wird.
2. Vermeidung von Überlastung: Ein zu häufiges Crawling kann den Server überlasten und sogar zu Ausfallzeiten führen. Mit einem Crawl-Delay kann das Risiko solcher Störungen minimiert werden.
3. Einsparung von Bandbreiten: Durch das Reduzieren der Crawler-Anfragen kann Bandbreite gespart werden, was besonders wichtig für Websites mit hohem Traffic oder begrenztem Hosting-Paket sein kann.
4. Kontrolle über Crawl-Raten: Website-Betreiber können gezielt steuern, wie oft Suchmaschinen ihre Seite besuchen, um die Sichtbarkeit in Suchmaschinen zu optimieren, ohne dabei Ressourcen zu verschwenden.

Insgesamt erlaubt der Crawl-Delay Website-Betreibern, ein Gleichgewicht zwischen der Aktualität ihrer Inhalte in Suchmaschinen und der optimalen Nutzung ihrer Server-Ressourcen zu finden.

Nachteile und Probleme des Crawl-Delays

Doch überall, wo es Vorteile gibt, gibt es natürlich auch Nachteile. Hierzu zähle ich mein oben genanntes Beispiel, denn der Crawl der Seite hat ewig gedauert und ist dabei nicht einmal fertig geworden. Zudem habe ich zwei Drittel der URLs mit einer 5xx-Antwort erhalten und somit war die ganze investierte Zeit quasi vergeudet. Doch welche Nachteile – und vor allem welche Probleme – hat die Einbindung eines Crawl-Delays in die robots.txt?

1. Verzögerte Indexierung: Ein zu langer Crawl-Delay kann dazu führen, dass Suchmaschinen länger brauchen, um neue oder aktualisierte Inhalte zu indexieren. Das kann die Sichtbarkeit von zeitkritischen Inhalten beeinträchtigen.
2. Unvollständige Indexierung: Bei Websites mit vielen Unterseiten kann ein zu hoher Crawl-Delay dazu führen, dass Suchmaschinen nicht alle Seiten in einem angemessenen Zeitraum erfassen können.
3. Kein standardmäßiges Verhalten: Nicht alle Crawler respektieren den Crawl-Delay, was bedeutet, dass einige Bots die Anweisungen ignorieren und die Websites weiterhin häufig abrufen könnten.
4. Fehlinterpretation: Ein falsch eingesetzter oder zu restriktiver Crawl-Delay kann von Suchmaschinen so interpretiert werden, dass die Website-Betreiber nicht möchten, dass ihre Inhalte häufig gecrawlt werden, was potenziell deren Rankings in den SERPs beeinträchtigen könnte.

Die Nutzung eines Crawl-Delays erfordert daher eine sorgfältige Abwägung, um sicherzustellen, dass die gewünschte Balance zwischen Serverleistung und Suchmaschinenpräsenz erreicht wird.

Wann ist ein Crawl-Delay sinnvoll?

Den Crawl-Delay in die robots.txt zu integrieren, ist dann sinnvoll, wenn man die Bots auf der Seite steuern will. Zudem kann man die Priorität der Bots beeinflussen, denn nicht jeder Bot ist gleichermaßen sinnvoll. Der Google-Bot und der Bing-Bot sollten beispielsweise die größte Priorität haben und darum eher nicht vom Crawl-Delay betroffen sein. Andere Bots wiederum haben nicht so eine hohe Priorität, da der Website-Betreiber von diesen nicht profitiert (zum Beispiel der ChatGPT-Bot), und diese sollten dann den Crawl-Delay haben.

Ansonsten ist der Crawl-Delay sinnvoll, wenn beispielsweise eine hohe Serverlast festgestellt wird. Diese hohe Auslastung sollte dabei nicht nur einmal auftreten, sondern

regelmäßig. Der Crawl-Delay kann dann dabei helfen, die Last zu reduzieren und sicherzustellen, dass menschliche Nutzer ohne Probleme alle URLs der Domain aufrufen können.

Es ist jedoch wichtig zu beachten, dass ein Crawl-Delay nicht immer die beste oder einzige Lösung ist. Bevor Sie einen festlegen, sollten die genaue Ursache für Serverprobleme oder andere Herausforderungen ermittelt werden. Manchmal könnte die Lösung in der Optimierung des Servers, der Website-Struktur oder anderen Faktoren liegen.

Fazit

Alles in allem ist ein Crawl-Delay ein nützliches Instrument für Website-Betreiber, um den Zugriff von diversen Crawlern und Bots auf ihrer Seite zu steuern und potenzielle Serverüberlastungen oder Bandbreitenauslastungen zu vermeiden. Besonders sinnvoll ist dieser bei Websites, die eine hohe Serverlast durch Bots erfahren oder auf einer sehr instabilen Server-Infrastruktur laufen.

Allerdings kann ein zu genauer Crawl-Delay auch negative Auswirkungen haben, insbesondere eine verzögerte oder unvollständige Indexierung in den verschiedenen Suchmaschinen wie Google, Bing und Co. Dies könnte wiederum die Rankings der Website in den SERPs beeinträchtigen.

Websites, die ständig aktualisiert werden, umfangreiche Inhalte haben oder von einer großen Anzahl von Nutzern besucht werden, sollten die Verwendung eines Crawl-Delays sorgfältig prüfen und dabei die Balance zwischen Server-Gesundheit und Suchmaschinenpräsenz finden.

Allerdings sollten kleinere Websites oder solche, die auf leistungsstarken Servern laufen und nicht von intensivem Crawling betroffen sind, möglicherweise auf die Einstellung eines Crawl-Delays verzichten, um eine maximale Sichtbarkeit in den Suchmaschinen sicherzustellen.

Kurz gesagt sollte die Entscheidung zur Implementierung eines Crawl-Delays in die robots.txt nach sorgfältiger Abwägung der spezifischen Bedürfnisse und Herausforderungen jeder Website getroffen werden.