Crawler

Stand: 13.10.2022

Das Internet ist im Prinzip eine große Ansammlung an Daten. Ganz gleich, ob Texteinträge, Videos oder Produktbilder, alle Elemente hinterlegen Informationen im Quellcode. Viele Unternehmen nutzen sie, indem sie sie sammeln und analysieren. Jede Zeile per Hand auszulesen, würde allerdings enorm lang dauern. Mithilfe spezieller Programme erleichtern Sie sich die Arbeit. Für solche Zwecke eignen sich unter anderem Crawler.

Was ist ein Crawler?

Ein **Crawler** ist ein **automatisch** arbeitendes Programm, das Daten aus dem Websiteangebot im Internet sammelt. Er wird auch als *Searchbot*, *Webcrawler* oder *Spider* bezeichnet. Während des Vorgangs **gruppiert** er die Informationen zusätzlich, um sie geordnet **wiederzugeben**.

Ursprünglich sollten Crawler das Wachstum des Internets ermitteln. Aus dieser Idee heraus entwickelten Programmierer in den 90ern eine Suchmaschine, aus der sich die heutigen Crawler herausgebildet haben, die vielfältige Einsatzbereiche bedienen. Sie sind äußerst aktiv: **Über ein Drittel** der Webseitenzugriffe entstehen durch die Programme von Suchmaschinen.

Funktionsweise eines Crawlers

Bevor der Crawler die erste Website durchsucht, muss ein Mitarbeiter ihn zunächst **konfigurieren**. Dadurch legt er fest, **welche URLs** das Programm durchgehen soll und **welche Daten** wichtig sind. Von dort aus folgt es jeder Weiterleitung. Es berücksichtigt dabei den dazugehörigen **HTML-Code** sowie **Hyperlinks**. Die Suchaufträge wiederholt der Crawler, sofern gewünscht, fortlaufend in einem festgelegten Intervall.

Das Programm agiert eingeschränkt, da es für jede Seite ein **Zeitkontingent** zur Verfügung hat. Daher ist es möglich, dass es Aktualisierungen erst beim erneuten Crawlvorgang erfasst. Darüber hinaus können Crawler das Internet aufgrund der Datenmenge noch nicht vollständig erschließen.

Verschiedene Crawler-Arten

Nicht jeder Crawler befasst sich mit den gleichen Seiten. Die bekanntesten Programme nutzen Suchmaschinen, wenn sie ihren **Index** aufstellen. Dabei trägt der Crawler alle nötigen Daten zusammen, die die Suchmaschine für die Speicherung sowie Weiterleitung auf die jeweilige Website benötigt.

Für kleinere Aufgabenbereiche reichen **Personal-Website-Crawler** aus, die Websitebetreiber beispielsweise nutzen, um die Aufrufbarkeit von Links zu überprüfen.

Andere Crawler-Arten nehmen einen Teil der Daten in den Blick. Sie erstellen beispielsweise **eine E-Mail-Liste**, um einen Verteiler aufzubauen. Soll ein Crawler nur bestimmte Themen beim Sammeln von Daten beachten, kommen sogenannte **Focused Crawler** zum Einsatz.

Crawler vs. Scraper

Sowohl ein Crawler als auch ein **Scraper** nehmen eine **Datenzusammenstellung** vor. Während ein Crawler die Daten bei sich aufbewahrt, um sie bei Bedarf abzurufen, verfolgt ein eingesetzter Scraper eine andere Strategie.

Scraper nehmen die Daten nicht nur auf, sie **kopieren** sie darüber hinaus auf eine separate Website. Dadurch eignen sie sich fremde Inhalte an. Teilweise erstellen Vergleichsportale so ihre Listen. Manche Akteure verwenden die Methode jedoch für schädliche Zwecke. Im Unterschied zu Crawlern steuern Scraper meist **einzelne Websites bewusst** an, ohne sich durch Links weiterleiten zu lassen.

Einsatzgebiete eines Crawlers

Ziel eines jeden Crawlers ist es, eine Datensammlung aufzubauen. Sie unterscheiden sich darin, mit welcher **Absicht** sie die Informationen schließlich festhalten. Für Suchmaschinen erstellen sie einen Index, den Nutzer durch bestimmte Suchbegriffe aufbereitet ansehen können.

Im Bereich des **Warenverkaufes** tragen sie **Produkteigenschaften** oder **Preise** zusammen, sodass ein Portal sie beispielsweise analysieren und miteinander **vergleichen** kann. Daneben gibt es weitere Gebiete, in denen Unternehmen Crawler einsetzen, wie bei **Linkstrukturen** in der Webanalyse.

Suchmaschinen überprüfen mithilfe neuer Daten die Beschaffenheit ihres **Indexes**. So stellen sie sicher, dass der Algorithmus neuen Content übernimmt und gelöschte Seiten entfernt, sodass User stets aktuelle Inhalte vorfinden.

Steuerung eines Crawlers und Bedeutung für die SEO

Wenn Sie nichts anderes bei Ihrer Website einstellen, können theoretisch **alle Crawler** auf die alle Seiten zugreifen. Um das zu verhindern, müssen Sie einen spezifischen Crawler in Ihrer robots.txt-Datei **sperren**. Suchmaschinen können die Informationen weiterhin einsehen und verwenden, wenn Sie sie nicht mittels „noindex“ davon abhalten. Jeder Crawler hat eine **unveränderliche Bezeichnung**, den User Agent. Bei Suchmaschinen ist meist deren Name darin enthalten, wie beim Googlebot von Google oder dem Bingbot von Bing.

Um einen Bot nicht gänzlich von einer Website auszuschließen, können Sie ihn in der **robots.txt-Datei** einschränken, sodass er beispielsweise keine Bilder indexieren oder Werbung folgen soll.

Die Crawler, die die großen Suchmaschinen einsetzen, sind allerdings förderlich für die **Sichtbarkeit** einer Website, da sie sie in ihre Suchergebnisse aufnehmen. Daher sollten Sie sich überlegen, ob Sie sie einschränken möchten. Für den Crawlingvorgang der gesamten Website, auch **Deep Crawl** genannt, benötigen die Programme eine gewisse Weile. Gleichzeitig können sie eine Website in Abhängigkeit ihrer Größe nur innerhalb einer bestimmten Zeitspanne crawlen. Vereinfachen Sie ihre Webseitenhierarchie, um das volle **Crawl Budget** auszuschöpfen.

Das verfügbare Budget steigt, **je wichtiger** die Suchmaschine die Website einstuft. Der Status hängt unter anderem von **Backlinks** ab, also davon, wie viele fremde Seiten auf die URL hinweisen. Es ist daher ein enormer Vorteil, wenn eine Seite viele Backlinks aufweist. Verknüpfen Sie andere Seiten mit Ihrer eigenen, kann der Crawler von dort aus dem nächsten Link folgen. Sorgen Sie darüber hinaus dafür, dass der Server **online** ist und **nicht zu lange lädt**, damit das Programm problemlos darauf zugreifen kann.

Neben der robots.txt-Datei ist ebenfalls die Google Search Console für die Steuerung wesentlich. Dort befindet sich die **XML-Sitemap** Ihrer Website und Sie sehen, welche Bereiche der Crawler erfasst. Für einen gewinnbringenden Webauftritt sollten möglichst viele Ihrer Unterseiten in den Suchergebnissen auftauchen. Dadurch sprechen Sie viele User an, die auf der Suche nach einem bestimmten Produkt oder Ähnlichem sind. Zudem klammern Sie durch eine gezielte Steuerung für Crawler und Suchmaschinennutzer irrelevante Seiten aus.