Webcrawler (Robot)

Stand: 04.07.2022

Webcrawler haben sehr unterschiedliche Bezeichnungen und werden auch als Spider, Searchbot oder Robot, kurz Bot, betitelt. Die Crawler durchsuchen das Internet nach neuen Inhalten und werden als Programme für Suchmaschinen benutzt. Neben der Ermittlung sind außerdem eine Bewertung und die Indexierung Aufgabengebiete von Webcrawlern.

Im Laufe der Jahre haben diese Bots zunehmend an Bedeutung gewonnen und stellen rund 40 % des Traffics im Web. Sie durchsuchen systematisch Inhalte und Dokumente im Web und stellen Querverbindungen via Links her. Linkbeziehungen, interne und externe Links werden verfolgt und ebenfalls durchsucht. Alles das geschieht nach einer exakten Vorgabe (Programmierung). Diese Aufgabe findet permanent statt, sodass die Spider auch neue Inhalte regelmäßig erfassen und katalogisieren können. An seine Grenzen gerät der Webcrawler bei Informationen, die nur über ein Suchfeld oder nach einem Login auf einer Seite erreichbar sind.

Wo werden Robots benötigt?

Die Verwendung der Webcralwer hängt ganz von der Programmierung an. Die gewünschten Suchergebnisse könnten beispielsweise tagesaktuelle Nachrichten sein. Die Bots erstellen Indizes und ordnen Suchergebnisse zu bestimmten Themen. Sie helfen dabei, relevanten Input von Unwichtigem zu trennen. Je nach Typ werden sie unter anderem für folgende Aufgaben benötigt.

Data-Mining: Data-Mining meint die Ermittlung von bestimmten Kontaktdaten und persönlichen Informationen wie Geburtsdaten und Telefonnummern oder auch E-Mail-Adressen.

Preisvergleich: Die Abgleichung von Produkten und Waren nach Angeboten und Preisen gewinnt immer mehr an Bedeutung. Die Crawler betreiben eine genaue Produktrecherche, um für die User die besten Angebote zu ermitteln, die über Suchmaschinen schließlich gut übersichtlich präsentiert werden.

Webanalyse: Eine weitere wichtige Aufgabe ist die Beobachtung von Websites und deren Benutzung. Wenn eine neue Seite an den Start geht, wird dies von den Webcrawlern erfasst. Auch die Aufrufe werden übermittelt und analysiert. Daraus ergeben sich Bewertungen für Beliebtheit und Relevanz von Webseiten.

Bots via Robots.txt steuern

Wer die Robots auf der eigenen Seite kontrollieren will, nutzt dazu eine Datei mit dem Namen“Robots.txt”. Dabei können Steuerungen geregelt und Überlastungen vermieden werden. Der Webcrawler arbeitet mit dem Robots-Exclusion-Standard-Protokoll. Der Robot wird nach Aufforderung nur vorgegebene Pfade beschreiten und bestimmte Seiten kontrollieren. Der Webcrawler kann ebenso gestoppt werden und seine Indexierungs- und Kontrollarbeit einstellen, wenn es Umbauten gibt oder eine Seite neu aufgesetzt wird.

Auf diese Weise verhindern Anwender falsche oder ungenaue Daten und sorgen für eine korrekte Bearbeitung durch den Webcrawler. Sollte der Benutzer dem Robot untersagen, bestimmte Inhalte nicht zu durchsuchen, werden diese dennoch von Suchmaschinen wahrgenommen und indexiert. Dem wirkt der Anwender entgegen, indem er sich dem Canonical-Tag, Noindex-Tag oder Meta-Tags bedient.