

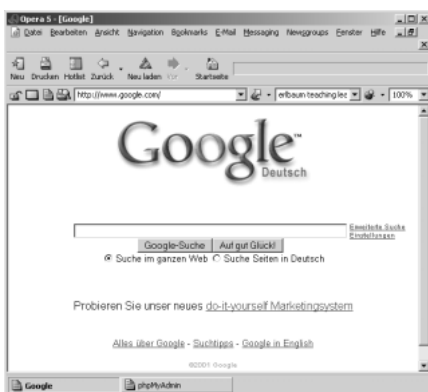
Robots, Spiders und Internetseiten

und ihr Einfluß auf die Besucherstatistik Ihrer Internetseiten

Für diese Ausgabe habe ich gedacht, nehme ich mir ein Thema vor, das von größter Wichtigkeit für alle ist, die Internet-Suchmaschinen benutzen. Gleichzeitig ist es aber auch etwas, von dem wir gewissermaßen keine Ahnung haben: Wie gelangen Internetseiten eigentlich in die Verzeichnisse von Suchmaschinen?

Die Kurzfassung für alle, die keine Zeit zum weiterlesen haben, lautet: Suchmaschinen vom Typ 'Freitextsucher' (im Gegensatz zu denjenigen die mit Registern und strukturierten Verzeichnissen arbeiten) bedienen sich hoch spezialisierter Programme, die ausgesandt werden, um Internetseiten aufzusuchen und die Informationen, die sie dort finden, an ihren Ausgangspunkt zurückzumelden. Die neuen Informationen werden in die Suchmaschine aufgenommen sobald deren Index für die Benutzer aktualisiert wird. OK, das war's. Sie können jetzt zum nächsten Artikel weiter blättern.

Oh, Sie sind ja noch da! Tja, wenn das so ist, können wir uns den ganzen Vorgang ja noch ein wenig näher anschauen: Diese speziellen Programme werden oft Robots, Spiders oder Crawler genannt. Wie bereits beschrieben, schwärmen sie aus und greifen sich Seiten aus dem Internet. Wenn es sich



um eine neue Seite handelt oder um eine Seite, die seit ihrem letzten Besuch verändert worden ist, kopieren sie sich diese neuen Daten. Sie finden Internetseiten entweder, weil sich der Urheber der Seiten bei einer Suchmaschine gemeldet und darum gebeten hat, in ihren Index aufgenommen zu werden oder weil der Roboter einem Link auf die Seite von einer anderen Internetseite aus gefolgt ist. Aus dem Gesagten ergibt sich Folgendes: Wenn der Urheber einer Internetseite den Suchmaschinen nichts von der Existenz der Seite mitteilt, und wenn auch keine Links

existieren, die auf die Seite hinführen, ist es höchst unwahrscheinlich, daß diese Internetseite gefunden wird.

Beware of the Robot

Robots arbeiten ununterbrochen. Die Robots z.B., die für AltaVista arbeiten, greifen auf ungefähr 10.000.000 Seiten am Tag zu. Wenn Ihre Internetseite im Verzeichnis einer Suchmaschine steht, können Sie davon ausgehen, daß zu irgendeinem Zeitpunkt ein Robot Ihre Seiten besucht hat. Dieser Robot ist dann allen Links, die Sie auf Ihren Seite haben gefolgt, und hat somit auch alle weiteren daraufhin gefundenen Einzelseiten kopiert. Vielleicht hat er das ja nicht auf einen Rutsch getan. Wenn Ihre Internetseiten z.B. besonders umfangreich sind, könnte der Vorgang den Server nämlich ziemlich stark belasten, und das würde Ihren EDV-Leuten wohl nicht so recht gefallen. Aus diesem Grund verteilen die Robots die Besuche auf Ihren Internetseiten über mehrere Tage, verzeichnen dabei jedes Mal nur ein paar Seiten und wiederholen den Vorgang so lange, bis sie alles kopiert haben, was sie nur können. Noch etwas: falls Sie jemals eine Ihrer Seiten einer Suchmaschine zur Aufnahme in deren Index vorgeschlagen haben, und wenn diese Suchmaschine dann behauptet, daß sie die Seite sofort registrieren und verzeichnen wird, tut sie dies deshalb noch lange nicht. In Wahrheit wird die Suchmaschine Ihrer Seite nur einen vorläufigen Besuch abstatten und sich eine Notiz machen. Zu einem späteren Zeitpunkt wird sie noch einmal zurückkommen und sich den Rest Ihrer Daten dann abgreifen.

Wie kann man nun wissen, ob einer dieser Robots die eigene Internetseite aufgesucht hat? Die Antwort lautet, wie bei den meisten Fragen, die mit dem Internet zu tun haben: 'Das kommt ganz darauf an...' Der naheliegendste Weg ist, selbst eine Suchmaschine aufzusuchen und eine Suche nach den eigenen Internetseiten zu starten. Falls Ihre Seiten gefunden werden, wurden sie zuvor in das Verzeichnis der Suchmaschine aufgenommen. Am einfachsten macht man dies bei einer Suchmaschine wie z.B. AltaVista folgendermaßen:

Host: URL Ihrer Seite, also z.B. host: philb

Wenn Sie auf diese Art Anfrage eine Antwort bekommen, wissen Sie, daß Ihre Internetseite im Verzeichnis dieser Suchmaschine enthal-

ten ist. (Es könnte sich vielleicht lohnen, gleich einmal Ihre Seiten einzeln zu überprüfen, nur um sicher zu gehen, daß die Suchmaschine alle Ihre Seiten verzeichnet hat, und daß es auch die aktuellen Versionen sind).

Wie auch immer, dies ist eine ziemlich umständliche Methode. Viel vernünftiger und einfacher ist es, auf die Log Files, die ja automatisch für Ihre Internetseiten geführt werden, zuzugreifen. Sie wissen vielleicht, daß jedesmal, wenn Sie eine Internetseite aufrufen, Ihr Browser eigentlich Dateien anfordert, und daß die Einzelheiten dieser Transaktionen von dem Hostserver bei dem die entsprechende Seite aufliegt, in einem Log File mit protokolliert und gespeichert werden. Diese Datei kann mit der entsprechenden Software, normalerweise 'Access Analy-



ser' genannt, eingesehen werden. Diese Datei kann Informationen liefern wie z.B. die IP-Adresse der Maschine, welche die Anfrage gemacht hat, welche Seiten angeschaut worden sind, das Betriebssystem des Anfragenden, den Domain-Namen, das Land aus dem die Anfrage kam und so weiter und so fort. Genauso wie jeder normale Browser diese kleine Spur hinterläßt, hinterlassen auch die Roboterprogramme der Suchmaschinen diese Spuren.

Wenn man also weiß, wie die Robots oder wie die Spider heißen, die von den verschiedenen Suchmaschinen eingesetzt werden, sollte es eine relativ einfache Aufgabe sein herauszubekommen, welche Suchmaschinen Ihre Internetseiten verzeichnen. Nun ja - und hier haben wir die bekannte Antwort wieder - , 'Es kommt halt darauf an...' Wenn Ihre Seiten sehr beliebt sind, sind ihre Log Files enorm groß und es kann ziemlich viel Zeit kosten, sich durch sie hindurch zu arbeiten bis man ein paar Namen findet, die einem

bekannt vorkommen. Hinzu kommt noch, daß über 250 verschiedene Robots am Werk sind, und daß von denen ein paar einzelne, viele oder gar keine Ihre Seite irgendwann einmal aufgesucht haben könnten. Also ist dies auch nicht der ideale Weg, um Suchmaschinen zu identifizieren. Außerdem werden ständig neue Robots eingeführt, und alte Modelle ändern Ihren Namen. Dies erschwert das Verfahren dann noch zusätzlich.

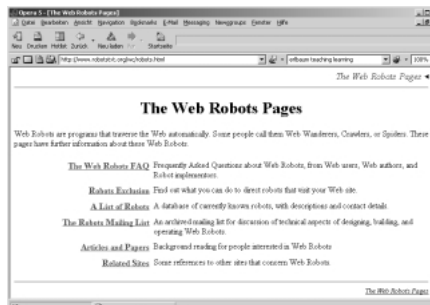
Ban the Bot!

Es gibt allerdings auch eine einfachere Lösung und die ist erfrischend anders! Bevor ich die allerdings verrate, hier noch schnell eine kleine Abschweifung. Es ist ja auch durchaus möglich, daß sie als der Urheber von Internetseiten es gar nicht wünschen, daß alle Ihre Seiten von den Robots besucht werden. Vielleicht sind ja ein paar Seiten noch gar nicht fertig oder einige Seiten enthalten persönliche Daten, die eigentlich nicht jeder lesen sollte, oder eine der Seiten soll nur vorübergehend ins Netz gestellt werden. Was immer auch der Grund sein mag, man will die Seiten nicht von einer Suchmaschine verzeichnen lassen. Die Macher der Suchmaschinen haben dies erkannt und so hat man eine Lösung gefunden, die es dem Urheber einer Internetseite erlaubt, den Robots mitzuteilen, bestimmte Seiten nicht zu verzeichnen, bestimmte Links nicht zu verfolgen oder bestimmte Unterverzeichnisse auf den Servern in Ruhe zu lassen. Dieser 'Robot Exclusion Standard' kann auf zweierlei Arten funktionieren. Entweder, indem Sie die Meta-Tag Bestandteile Ihrer Internetseite nutzen (eine Meta-Tag ist eine spezielle Art von HTML Tag, die in eine Internetseite eingebaut wird, und die nur von Suchmaschinen und nicht von einem ganz normalen Browser wie z.B. Netscape oder Internet Explorer gelesen werden kann. Anm. des Übersetzers: Diese Aussage trifft nur für den Meta-Tag 'robot' zu. Es gibt durchaus Meta-Tags wie z.B. 'expire' oder 'redirect' die auch von ganz normalen Webbrowsern gelesen werden), oder indem Sie eine kleine Textdatei im entsprechendem Verzeichnis Ihrer Internetseiten plazieren.

Uns interessiert eigentlich nur diese zweite Methode. Weil die Suchmaschinenrobots wissen, daß manche Urheber von Internetseiten nicht wollen, daß sie alles verzeichnen, suchen sie bei Ihren Besuchen zuerst nach dieser Textdatei, die 'robot.txt' heißt, um zuallererst einmal herauszufinden, ob sie überhaupt die ganze Seite verzeichnen dürfen. (Für das Verständnis dieses Artikels ist es nicht weiter wichtig, im Detail darzustellen was diese robot.txt Datei enthalten oder nicht enthalten soll. Für diejenigen, die es näher

interessiert, gibt es eine gute Beschreibung bei AltaVista (http://doc.altavista.com/adv_search/ast_haw_avoiding.html) Wenn Sie der Sache allerdings ganz auf den Grund gehen wollen, möchten Sie sich vielleicht 'A Standard for Robot Exclusion' (<http://info.webcrawler.com/mak/projects/robots/norobots.html>) anschauen. Zugegebenermaßen suchen nicht alle Roboter nach dieser Datei, aber die meisten tun es und halten sich auch an die Anweisungen, die sie vorfinden.

Also gut, das war also die kleine Abschweifung, jetzt zurück zum Artikel. Wenn Sie sich Ihre Internetstatistik anschauen, sollten Sie speziell auf Anfragen nach den 'robot.txt'-Dateien achten, denn nur Robots und Spider suchen nach diesen bestimmten Dateien. Kein normaler Browser würde danach suchen. Auf diese Weise sollte es dann viel



einfacher sein, herauszufinden, welche Suchmaschine Ihre Internetseite bis zu einem bestimmten Zeitpunkt aufgesucht hat. Wenn Sie sehen, daß 'Scooter' die Datei angefordert hat, können Sie ihn dann bis zu AltaVista zurückverfolgen. Vorausgesetzt natürlich, daß sie wissen, daß Scooter zur Suchmaschine AltaVista gehört.

Eine sehr nützliche Internetseite, die wissenswertes zu über 250 Robots und Spiders verzeichnet findet sich unter: 'The Web Robots Page' mit der URL: <http://info.webcrawler.com/mak/projects/robots/robots.html>. Schon das Namensverzeichnis klingt ziemlich faszinierend. Da gibt es z.B. ein Programm, das sich Ariadne nennt (Anm. des Übersetzers: Ariadne heisst auch die Zeitschrift, in der dieser Artikel zuerst erschienen ist) und das Teil eines Forschungsprojekts der TU München ist, ein anders nennt sich Dragon Bot (das sammelt Seiten, die etwas mit Südostasien zu tun haben) es gibt auch Googlebot (also ich setze keinen Preis für denjenigen aus, der mir sagen kann, woher der Name wohl kommt!)

Facts or Artefacts?

Gibt es irgendwelche Nachteile, wenn man diesen Ansatz verwendet? Klar, Sie haben es natürlich schon vermutet, die Antwort lautet

abermals: 'Es kommt darauf an...' Wenn Sie nur in sehr begrenztem Umfang Zugriff auf das statistische Material Ihrer Seiten haben, ist es möglich, daß sie eine sehr viel höhere Anzahl von Besuchen auf Ihren Einzelseiten zählen, als dies in Wirklichkeit von 'richtigen' Besuchern der Fall war. Wenn Sie nicht mehr Informationen aus Ihrer Statistik entnehmen können, ist es sehr schwer, die 'wirklichen Besucher' von den Spidern zu unterscheiden. Manche Experten behaupten, daß die Probleme, welche die Spider verursachen, von der 'Bandbreite' herrühren, die sie verwenden um Ihre Daten zu sammeln, speziell wenn sie die 'Schnellfeuermethode' verwenden, mit der sie in ganz kurzer Zeit ganz viele Daten abzugreifen versuchen. Diese Methode führt für den Normalnutzer, der die Seiten ja wirklich anschauen will, zu langen Antwortzeiten. Und weil jeder mit genügend Kenntnissen einen Robot oder Spider basteln kann, wird ihre Zahl in Zukunft eher noch zunehmen. Und obwohl die 'robot.txt'-Datei in diesem Fall hilfreich sein kann, gibt es keine Verpflichtung oder keinen Standard, an die oder an den sich die Robots halten müssen. Einige Robots werden sie höchstwahrscheinlich sogar vollkommen ignorieren. Dieses Problem wurde vor einiger Zeit in einem interessanten Artikel von Martijn Koster behandelt. Der Artikel ist nun zwar schon ein paar Jahre alt, er ist aber immer noch hilfreich und interessant zu lesen. <http://www.cs.biu.ac.il/home/computing/manuals/web/robots/threat-or-treat.html>

Nun, sei es wie es sei. Man kann Sie mögen, verabscheuen, oder sie können einem völlig egal sein, Spiders sind eine der wichtigsten Methoden, die wir haben, um an die Informationen, die sich da draußen befinden, heranzukommen.

Phil Bradley, Independant Internet Consultant, Feltham, UK, philb@philb.com

[Mit fr. Genehmigung vom Autor aus Ariadne Issue 27: <http://www.ariadne.ac.uk/issue27/search-engines/intro.html>]

Aus dem Englischen von *Sabine Buroh*, Freiburg

Wenn Sie nun wissen wollen, was SIE tun können, damit Ihre Webseiten von den Suchmaschinen möglichst gut gefunden werden, lesen Sie folgenden Aufsatz von Dr. Joachim Schuhmacher: <http://www.multimedia-beratung.de/artikel/suchmaschinengerechte-seitengestaltung.htm>