Roboter-Erkennung für Nutzungsstatistiken

Vorgehen für die Filterung von automatisierten Zugriffen auf Dokumente in Repositorien für Nutzungsstatistiken

Version 1.1, Juli 2013

Matthias Hitzler, Marco Recke

Computer- und Medienservice der Humboldt-Universität zu Berlin
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Saarländische Universitäts- und Landesbibliothek Saarbrücken
Universitätsbibliothek Stuttgart
und
Verbundzentrale des GBV (VZG)

Auf Initiative und in Abstimmung mit der
Deutschen Initiative für Netzwerkinformation e. V.

1 Einleitung

Open-Access-Repositorien sind für jeden Zugriff offen. Ein erheblicher Anteil an Hits, oft 30% und mehr wird dabei durch automatisch arbeitende Computerprogramme erzeugt. Das sind unter anderem Suchmaschinen, die Websites suchen und in ihre Indizes aufnehmen, aber auch Programme, die das Internet nach bestimmten Inhalten von Websites wie E-Mail-Adressen durchsuchen. Diese so genannten Robot Hits müssen erkannt und ausgeschlossen werden, da sie die Nutzungsstatistik verfälschen.

Nur ein Teil der Roboter gibt sich in den Logfiles direkt zu erkennen. Bei den übrigen kann man aufgrund der vagen Merkmale nicht mit Sicherheit entscheiden, ob es sich um einen Roboter oder einen menschlichen Benutzer handelt.

Da Roboter-Zugriffe nie vollständig aus der Statistik eliminiert werden können, kommt es vor allem darauf an, gemeinsame Kriterien zu entwickeln, die bei der Filterung der Zugriffe anzuwenden sind, um die Statistiken untereinander vergleichbar zu machen.

2 Roboter

Roboter sind Computerprogramme, die im Internet verschiedenste Aufgaben erfüllen. Sie werden unter anderem für statistische Analysen von Websites (Anzahl der Dokumente auf einem Server, Häufigkeit von Dateitypen etc.) oder der Pflege von Seiten (Auffinden von „dead links“) genutzt. Die meisten Roboter existieren jedoch für unterschiedliche Zwecke, zumeist für Suchmaschinenindizierung, also um Ressourcen im Internet zu entdecken und zu erfassen.

Aus dem funktionalen Vorgehen von Roboter lässt sich folgende formale Definition ableiten:

„A robot is a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced.” (Koster 1999)

Hierbei kann hinsichtlich der Intention und des Verhaltens zwischen „freundlichen“ und „bösen“ Robotern unterschieden werden. „Freundliches“ Verhalten schließt den durch das Roboter-Exclusion-Protokolls (REP)[1] festgelegten Zugriff auf eine im Stammverzeichnis des Webservers abgelegte Datei (robots.txt) ein. In dieser Datei wird festgelegt, ob bzw. welche Unterordner und -seiten der Roboter besuchen darf und welche nicht. Die Implementierung dieses Protokolls ist jedoch freiwillig, so dass viele Roboter, insbesondere die „böswilligen“, auf diesen Zugriff verzichten bzw. dessen Inhalt ignorieren.

Für Repositorien, insbesondere Open-Access-Repositorien, spielt das Verhindern von Robot-Zugriffen mittels der robots.txt keine große Rolle. Vielmehr ist es in der Regel erwünscht, dass Suchmaschinen die Seiten indizieren und somit ein Auffinden der Informationen erleichtern. Für die Berechnung von Nutzungsstatistiken jedoch sind Robot-Zugriffe äußerst relevant.

[1] http://www.robotstxt.org/orig.html.

3 Identifikation von Robot-Zugriffen

Zur Identifikation von Roboterzugriffen auf Basis von Log-Files gibt es verschiedene Ansätze (vgl. Geens u.a. 2006), welche hier kurz vorgestellt werden sollen.

Die klassische und am häufigsten genutzte Methode ist die Identifikation anhand des user agent. Allgemein wird der user agent als Parameter in einem http-Header festgelegt und lässt Rückschlüsse auf das zugreifende Programm zu. Bei menschlichen Nutzern besteht der Parameter in der Regel aus Browsertyp und -version sowie Hinweise auf das Betriebssystem. Bei automatisierten Zugriffen werden häufig spezifische Bezeichnungen verwendet. Anhand dieser Bezeichnungen können sie von menschlichen Zugriffen unterschieden werden. Ein offensichtlicher Nachteil dieses Systems ist die Beliebigkeit der user agent Bezeichnung. Roboter können einen frei gewählten user agent erhalten, so dass etwa auch eine Tarnung als menschlicher Zugriff möglich ist.

Eine weitere Methode ist die Identifikation anhand von IP-Adressen. Bekannte Roboter werden mit ihrer entsprechenden IP-Adresse in eine Liste eingetragen und können dann in einem Abgleich mit dem Log-File erkannt werden. Probleme macht dieser Ansatz durch wechselnde IP-Adressen der Roboter, sowie Proxy-Servern, welche die ursprüngliche IP verschleiern. Eine Pflege und Aktualisierung einer solchen Liste ist zudem nicht erst seit IPv6 kompliziert und mühsam.

Die Entdeckung von Roboterzugriffen durch deren Verhalten auf dem Webserver ist eine weitere Methode. Hierbei können verschiedene Verhaltensmuster auf den Zugriff eines Roboters hindeuten:

1. Zeitmuster: Zugriff erfolgt oft in schnellen oder gleichmäßigen Abständen auf einzelne Dateien des Webservers
2. Dokumentabhängiger Zugriff: Bilder werden von Robotern häufig ignoriert, da diese keinen nennenswerten Nutzen für sie haben
3. Verweis: Roboter ignorieren häufig den referrer Parameter (der Ort von dem auf die aktuelle Datei referenziert wurde)
4. Zugriff auf robots.txt: Manche Roboter greifen als erstes auf die Datei robots.txt zu
5. Methode des Aufrufs: Roboter greifen häufig nur per HEAD auf Dateien zu

Die Wahrscheinlichkeit, dass ein menschlicher Zugriff fehlerhaft als Roboter-Zugriff deklariert wird, ist bei diesen Herangehensweisen zur Identifikation jedoch wesentlich höher als bei der Filterung nach user agent oder IP-Adresse. Heuristische Methoden versuchen die oben genannten Verhaltensmuster zu verzahnen und sich einer möglichst geringen Fehlerquote anzunähern. (vgl. Stassopoulou u.a. 2007; Tan u.a. 2002; Geens u.a. 2006) Geens u.a. (2006) haben hierzu bei einer konkreten Untersuchung herausgefunden, dass die optimale Verknüpfung wie folgt aufgebaut ist: “Robots.txt OR IP address list OR Robotic user agent OR HEAD method OR (Unassigned referrer AND No images)”. Hierbei hat die Analyse eine Trefferquote von 97,51% bei einer Fehlerquote von 10,65% ergeben. Diese Daten sind jedoch an einem Einzelfall erarbeitet und tendenziell veraltet.

4 Umsetzung der Roboterfilterung in OAS und Strategie

Die im Projekt Open-Access-Statistik vorgenommene Filterung der Roboterzugriffe basiert derzeit auf dem Prinzip der oben beschriebenen Identifikation nach dem user agent. Eine Liste, welche bekannte Roboter bzw. deren regulären Ausdruck enthält, wird mit dem user agent des jeweiligen Zugriffs

abgeglichen. Die auf der Website von COUNTER bereitgestellte Liste[2] wurde im Projekt Open-Access-Statistik um etliche Eintragungen erweitert, welche aus der frei zugänglichen Liste von robotstxt.org[3] sowie aus der Open-Source-Statistiksoftware AWStats gespeist wurde.

Um vergleichbare Nutzungsstatistiken zu erhalten, muss eine solche Liste der Forschergemeinde frei zugänglich sein[4]. Ein angestrebtes Ziel im Projekt OA-Statistik war es, die zukünftige Verwaltung einer solchen Roboterliste auszuarbeiten und umsetzen. Ein Zwischenschritt sind die in den KE Guidelines festgehaltenen Anforderungen an eine solche Roboterliste[5]:

1. Es muss möglich sein, die Liste mit einem 'timestamp' zu versehen, so dass der jeweils aktuelle Stand der Liste stets ersichtlich ist.
2. Die Liste muss in einem maschinenlesbaren Format, bevorzugt XML, zur Verfügung gestellt werden.
3. Die Liste muss von COUNTER akzeptiert werden.
4. Die Quelle der jeweiligen Eintragungen muss erkennbar sein.
5. Die Liste muss anhand eines persistenten URI verfügbar sein.
6. Es muss möglich sein, verschiedene Versionen der Roboterliste zu verwalten.

Diese Liste sollte um folgende Punkte ergänzt werden[6]:

7. Die Liste kann und soll erweitert werden.
8. Neue Versionen der Liste werden regelmäßig veröffentlicht.
9. Zur leichteren Erkennung verschiedener Versionen müssen zusätzlich zum 'timestamp' der Liste Versionsnummern verwendet werden.
10. Die Basis Liste besteht aus den Listen COUNTER, PLOS und NEEO.
11. Neue Einträge können von der Allgemeinheit vorgeschlagen werden.
12. Neue Einträge werden vor Aufnahme in die Liste überprüft.
13. Die dauerhafte Verfügbarkeit der Liste muss gewährleistet werden.

Bisher nicht bekannte user agents können idealerweise anhand ihres Namens erkannt werden. Falls dies nicht möglich ist, können oben beschriebene Verhaltensmuster einer Identifikation dienlich sein. Sollten solche Zugriffe darauf hindeuten, dass es sich um einen Roboter handelt, muss aber überprüft werden, ob der damit verknüpfte user agent nicht tendenziell auch von menschlichen Zugriffen genutzt wird.

Eine solche allgemeine Roboter-Liste wurde während des Workshops “Usage Statistics And Beyond” (22.-23. April 2013) ausgiebig diskutiert und soll im Rahmen der COAR-Arbeitsgruppe “Usage Data And Beyond” eine konkrete Ausgestaltung erfahren.

[2] http://www.projectcounter.org/code_practice.html.
[3] http://www.robotstxt.org/db.html.
[4] Einen kommerziellen Dienst, welcher kostenpflichtig eine regelmäßig aktualisierte Roboterliste zur Verfügung stellt ist, wird von IAB/ABCe angeboten: http://www.iab.net/1418/spiders.
[5] http://wiki.surf.nl/display/standards/KE+Usage+Statistics+Guidelines#KEUsageStatisticsGuidelines-5.2.3.Robotlistschema.
[6] Diese Punkte waren Teil der Diskussion der Breakout Session „Roboter-List“ beim Workshop Usage Statistics and Beyond.

5 Literatur

Dikaiakos M., Stassopoulou A., Papageorgiou L.: An investigation of web crawler characterization and metrics. In: Computer Communications, 28(8), S. 880-897. Online: http://www.sciencedirect.com/science/article/pii/S0140366405000071.

Koster, Martijn (1999): “What is a WWW robot?” In: The Web Rob ots Pages. Online: http://webdoc.gwdg.de/ebook/aw/1999/webcrawler/mak/projects/robots/faq.html#what.

Geens N., Huysmans J., Vanthienen J.: Evaluation of Web Robot Discovery Techniques: A Benchmarking Study. Advances in Data Mining, 6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14-15, 2006. Proceedings, Buchreihe Lecture Notes in Computer Science, Volume 4065/2006. Online: http://www.springerlink.com/content/271124758k874484/.

Tan P.-N., Kumar V.: Discovery of web robot sessions based on their navigational patterns. In: Data Mining and Knowledge Discovery, 6 (1), S. 9-35. Online: http://link.springer.com/article/10.1023%2FA%3A1013228602957.