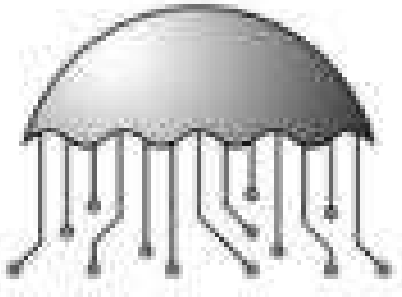


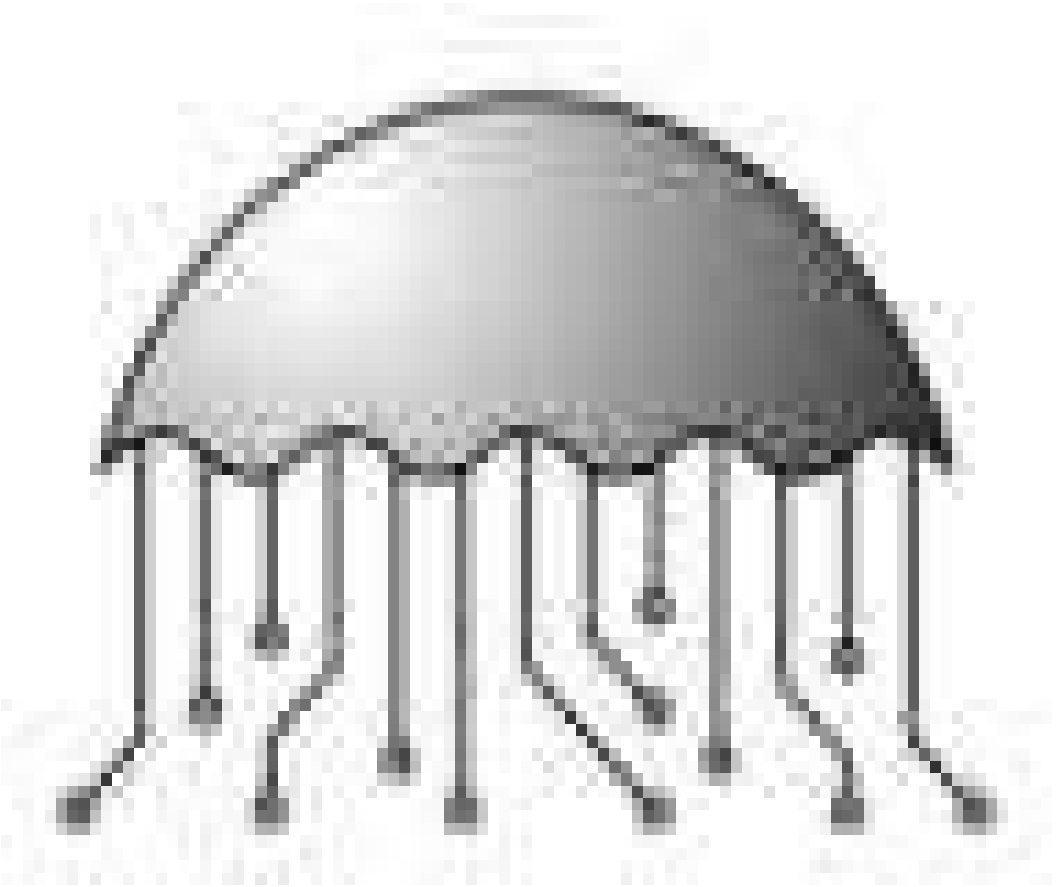
Web-Page Crawler auf der Basis von Konzepten von Cho, Molina, Page

Heinrich-Heine Universität Düsseldorf
Informationswissenschaft
Professor Stock
Sommersemster 2004
Daniel Ritter

Übersicht

1. Was ist ein Crawler ?
2. Crawlerkonzepte von Cho, Molina, Page
3. Eigenschaften des Googlebots

Was ist ein Crawler ?



Was ist ein Crawler ?

„Ein Crawler ist dermassen programmiert, dass er ständig das Web durchsurft und allen Links folgt, an denen er vorbeikommt. Wenn er neue Webseiten besucht, überprüft er seine eigene Datenbank um zu sehen ob die Seite bekannt ist. Ist sie unbekannt, fügt er sie der Datenbank hinzu, ist sie bekannt fügt er eventuelle Änderungen des Inhalts der Datenbank zu.“

Was ist ein Crawler ?

„Eine Komponente einer Suchmaschine, die das Web durchwandert, URL'S speichert und Keywords für jede besuchte Seite berechnet. Crawler können auch Robots oder Spider genannt werden.“

Was ist ein Crawler ?

„Ein Crawler ist ein Softwareprogramm, dass selbstständig das WWW durchwandert und gefundene Webseiten weiterverarbeitet und speichert. “

Was ist ein Crawler ?

•Ein Crawler ist ein Softwareprogramm

•Ein Crawler ruft automatisiert Webseiten auf und findet auf ihnen Hypertext-Links zu weiteren Webseiten, welche danach ebenfalls in einer festgelegten Reihenfolge aufgerufen werden.

•Dies findet nach vorher festgeschriebenen Regeln statt.

•Ein Crawler ist an einen Index gekoppelt, in welchen er die gefundenen Seiten einspeist.

•Ein Crawler kann auch zur Index-Pflege benutzt werden, indem er nicht mehr existierende Seiten aus dem Index entfernt oder geänderte und umgezogene Seiten im Index aktualisiert.

Was ist ein Crawler ?

Ein Crawler muss NICHT UNBEDINGT zu einer Suchmaschine gehören.

Google™ Deutschland

nutch

WebSnake

> wget

AvantGo®

Was ist ein Crawler ?

Schematischer Ablauf eines Crawlingvorgangs

Schritt 1
Übergabe einer Start-URL an den Crawler

Schritt 2
Einlesen des Inhaltes der Start-URL. Schreiben der gefundenen Links in einen Speicher. Weitere individuelle Schritte für den spezifischen Crawler

Schritt 3
Einlesen des Inhaltes einer der URL's aus dem URL-Speicher. Schreiben der gefundenen Links in den Speicher. Weitere individuelle Schritte für den spezifischen Crawler

Schritt 4
Überprüfen, ob der Speicher leer ist oder eine andere Abbruchbedingung erfüllt wurde. Falls weiter gecrawlt werden soll, Rücksprung zu Schritt 3. Ansonsten Sprung zu Schritt 5.

Schritt 5
Ende des Crawlvorgangs

Crawler Konzepte von Cho, Molina, Page



Crawler Konzepte von Cho, Molina, Page

Schematischer Ablauf eines Crawlingvorgangs

Schritt 1
Übergabe einer Start-URL an den Crawler

Schritt 2
Einlesen des Inhaltes der Start-URL. Schreiben der gefundenen Links in einen Speicher. Weitere individuelle Schritte für den spezifischen Crawler

Schritt 3
Einlesen des Inhaltes einer der URL's aus dem URL-Speicher. Schreiben der gefundenen Links in den Speicher. Weitere individuelle Schritte für den spezifischen Crawler

Schritt 4
Überprüfen, ob der Speicher leer ist oder eine andere Abbruchbedingung erfüllt wurde. Falls weiter gecrawlt werden soll, Rücksprung zu Schritt 3. Ansonsten Sprung zu Schritt 5.

Schritt 5
Ende des Crawlvorgangs

Crawler Konzepte von Cho, Molina, Page

Grundsätzliche Idee

„Wenn man nicht zufällig IRGENDEINE URL als nächstes besucht, sondern versucht die relevantesten Seiten zuerst zu besuchen, wird der Index der Suchmaschine qualitativ besser werden. Ausserdem können Ressourcen gespart werden“

Crawler Konzepte von Cho, Molina, Page

- „Ur-Crawler“ arbeiteten nicht nach dieser Idee
- Idee ist NICHT funktionabel für kleine Webs die in kurzer Zeit komplett gecrawlt werden können
- Idee ist bei nicht komplett erfassbaren Webs ein wichtiger Faktor
- Idee soll helfen Spam zu reduzieren
- Idee soll helfen relevante Dokumente zuerst zu erfassen
- Idee soll mehr relevante Dokumente erfassen
- Idee spart Speicherplatz
- Idee spart Internetbandbreite
- Idee spart Rechenleistung -> Energie
- Idee verbessert den Index bereits beim Crawlvorgang
- Idee verringert menschlichen Aufwand zur Indexpflege
- Index bleibt aktueller da Seiten häufiger besucht werden können

Crawler Konzepte von Cho, Molina, Page

Kriterien nach denen die nächste zu crawlende URL ausgesucht werden kann

1. Ähnlichkeit zu einem Suchwort für den Crawler (Similarity to driving query)

Seiten werden zuerst gecrawlt, auf die Links zeigen, die dem Suchwort ähnlich sind. Beim aktualisieren von Seiten werden solche zuerst gecrawlt, die das Suchwort (oft) enthalten.

2. Backlink Anzahl

Die Anzahl der Links die auf eine URL zeigen generieren ihren Backlinkwert

3. Pagerank

Einsatz des Pageranks, besonders interessant für bereits erfasste Seiten die aktualisiert werden sollen, bei nicht bekannten Seiten muss der PR geraten werden.

4. Location Metric

Bestimmte Eigenschaften der URL des Dokuments, z.B. wenige Unterverzeichnisse, Endung .com, bestimmtes Suchwort in URL enthalten

Crawler Konzepte von Cho, Molina, Page

3 verschiedene Crawlingmodelle

1. Crawl & Stop

In diesem Konzept besucht der Crawler eine festgelegte Anzahl von K Seiten mit einer beliebigen Start-URL. Ein ideal funktionierender Crawler hätte nach dem Besuch von K Seiten aus allen verfügbaren Seiten nur die K relevantesten extrahiert. Wären also N Seiten durch Links (K < N) verfügbar gewesen, hätte der Crawler aus diesen N verfügbaren Seiten die K relevantsten Seiten extrahiert und den Vorgang beendet. Einfach ausgedrückt könnte man sagen, dass der Crawler eine gewissen Menge von Seiten besuchen darf, die Menge der verfügbaren Seiten jedoch grösser ist und er nur die relevantesten Seiten aus der Gesamtmenge besuchen sollte.

Crawler Konzepte von Cho, Molina, Page

3 verschiedene Crawlingmodelle

2. Limited Buffer Crawl

Auch bei diesem Vorgang besucht der Crawler eine festgelegte Anzahl von K Seiten. Sein Speicherplatz in den er erfasste Seiten ablegen kann ist jedoch nur S gross (S < K). Der Crawler hat also nicht genug Platz um alle gecrawlten Dokumente unterzubringen. Hier entfernt der Crawler während des Crawlingvorgangs Dokumente aus seinem Speicher um Platz für neue zu schaffen. Bei einem ideal funktionierenden Crawler wären nach dem Erfassen von S Dokumenten nur noch die relevantesten aus der Gesamtmenge K im Speicher. Da dem Crawler die Relevanz von neue erfassten Dokumenten nicht bekannt ist muss er versuchen sie auf der Basis der bereits vorhandenen Dokumente zu „erraten“.

Crawler Konzepte von Cho, Molina, Page

3 verschiedene Crawlingmodelle

3. Crawl & Stop with Threshold (Threshold = Schwelle)

Hier wird vorgegangen wie bei der ersten Methode. Jedoch wird ein Ziel-Qualitätskriterium für das Crawlen vergeben. Nach dem Crawvorgang von K Dokumenten muss ein perfekter Crawler alle Dokumente mit dem höchsten Qualitätswert erfasst haben.

Crawler Konzepte von Cho, Molina, Page

Experimentaler Crawlvorgang

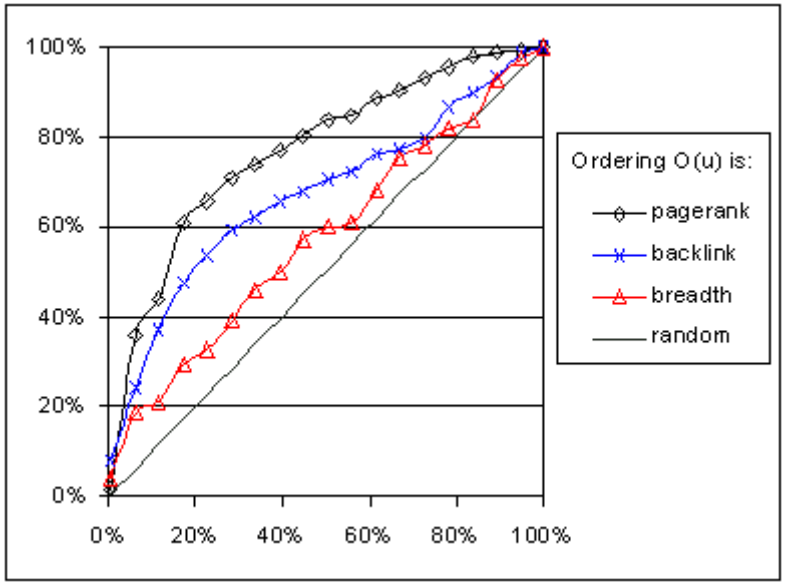
Dataset ist das Web der Standford University.

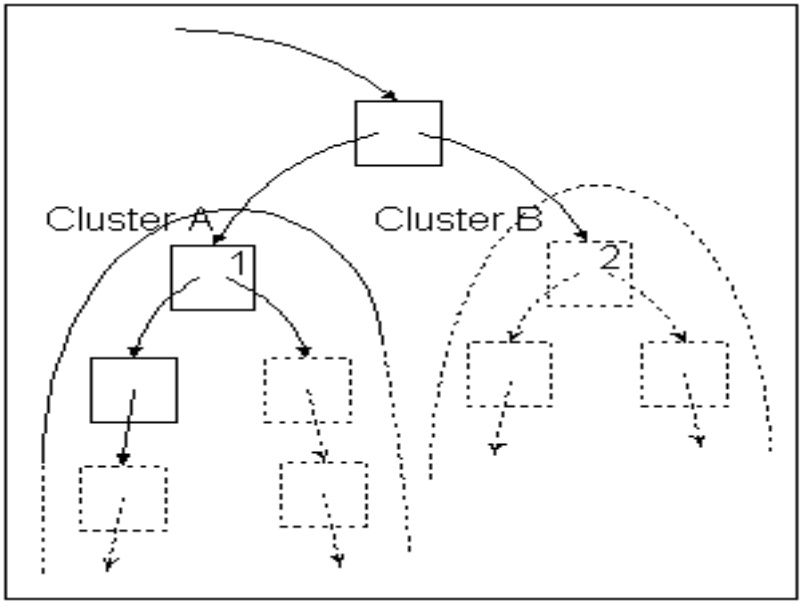
Benutzt wird ein „virtueller“ Crawler um Traffic zu sparen

Der virtuelle Crawler hat ein Web von 225.000 Dokumenten zur Verfügung

Crawler Konzepte von Cho, Molina, Page

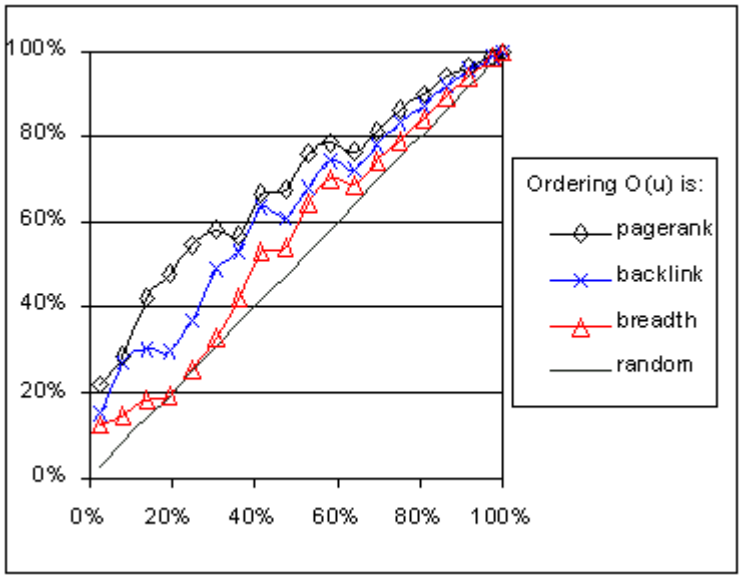
Experiment 1 : Backlink based & Threshold





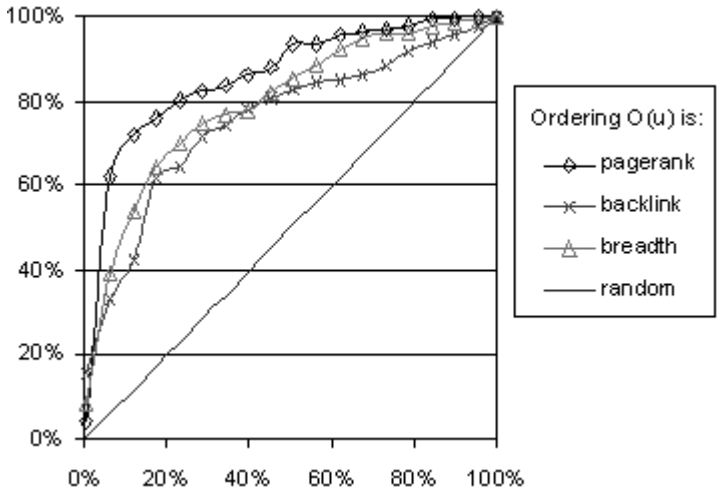
Crawler Konzepte von Cho, Molina, Page

Experiment 2: Crawl & Stop (Kein Schwellenwert)



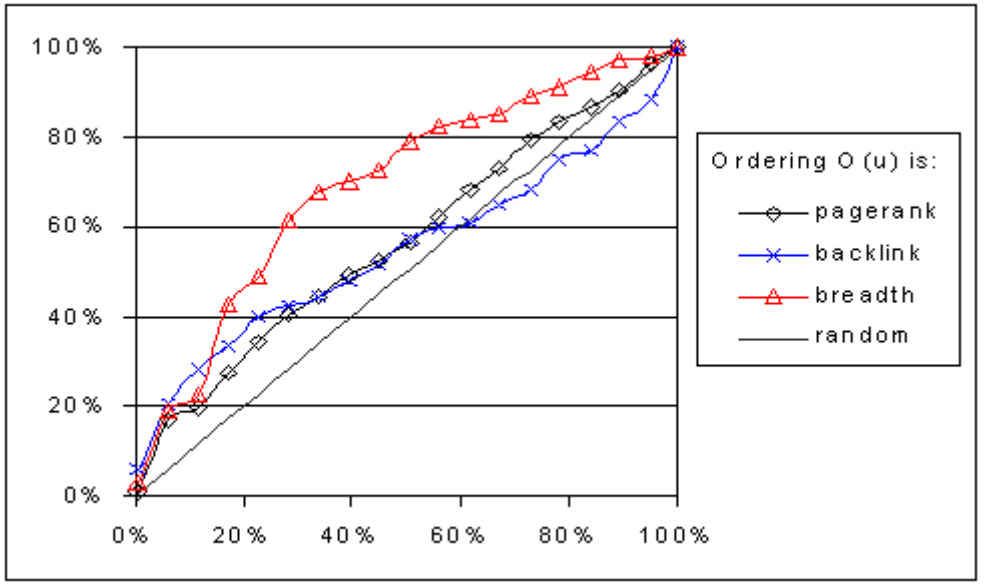
Crawler Konzepte von Cho, Molina, Page

Experiment 3: Threshold & PageRank



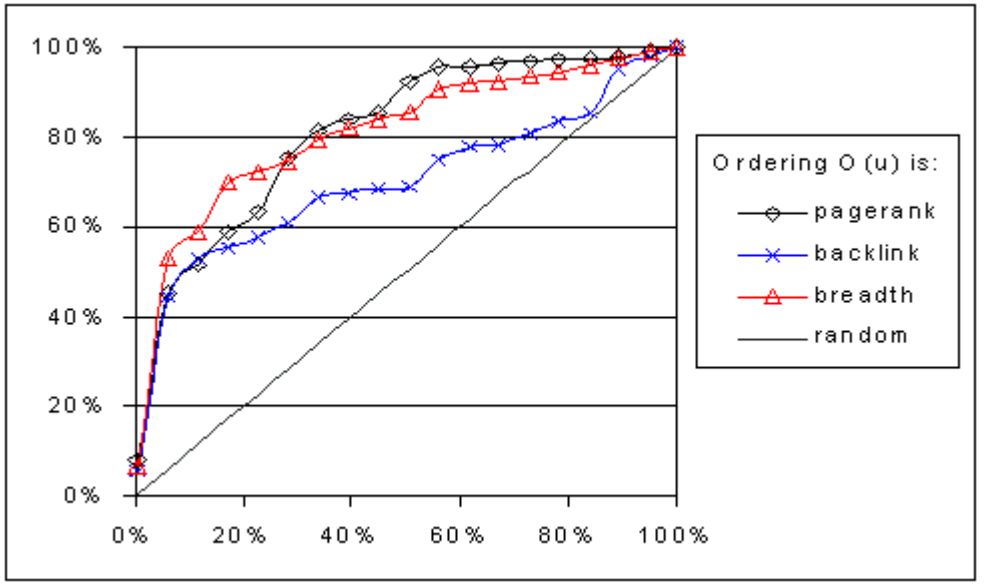
Crawler Konzepte von Cho, Molina, Page

Experiment 4: Ähnlichkeit zu Suchbegriff



Crawler Konzepte von Cho, Molina, Page

Experiment 5: Suchwort & Backlinks



Eigenschaften des Googlebots

- Google lässt sich nicht gerne in die Karten schauen
- Viele Vermutungen, wenige Fakten
- Googlebot's Code ist ständigen Veränderungen unterworfen
- Googlebot basiert mit Sicherheit zu Teilen auf den eben besprochenen Konzepten.

Eigenschaften des Googlebots

Google lässt sich nicht gerne in die Karten schauen

From:<googlebot@google.com>
Subject: Re: [#10444133] Additional Googlebot Information for a student available ?
Date:Wed, 02 Jun 2004 14:59:25 -0700
To: Daniel Ritter <ritterd@uni-duesseldorf.de>

Hi Daniel,

Thank you for your interest in Google. You can find information about Google, including search services, corporate information, services available for site owners, and much more, at http://www.google.com/about.html. As you may imagine, because Google is a privately held company, there is some information that we're unable to provide to the public at this time.

Additionally, you may be interested in receiving our free bimonthly Google Friends Newsletter, which keeps you up to date on all the latest happenings at Google. Simply visit http://www.google.com/contact/newsletter.html to sign up.

Regards,
The Google Team

Eigenschaften des Googlebots

Fakten

Googlebot ist auf mehrere Rechner verteilt.

Eigenschaften des Googlebots

Fakten

Es gibt 2 verschiedene Googlebot-Varianten

64.*.*.* (Freshbot läuft ständig)
209.*.*.* (Deepbot läuft einmal im Monat)

Eigenschaften des Googlebots

Fakten

Seiten mit höherem Pagerank werden häufiger besucht

Eigenschaften des Googlebots

Fakten

Googlebot crawlt langsamer als er könnte
um Server nicht voll auszulasten.

Eigenschaften des Googlebots

Fakten

Für lokale Google-Sites wie z.B. google.de werden lokalsprachige Dokumente anders gewichtet.

Eigenschaften des Googlebots

Fakten

Googlebot hält sich an die robots.txt

Eigenschaften des Googlebots

Robots.txt Beispiel

```
# robots.txt für http://www.meineseite.de

User-agent: *
Disallow: /meinebilder/
Disallow: /meineprivatentexte/
Disallow: /meineliebesbriefe/

Allow: /schwerzufindendesverzeichnis/

User-agent: BlödeSuchmaschine
Disallow: *
```

Eigenschaften des Googlebots

Fakten

Googlebot berücksichtig Backlinks nur bei Seiten mit einem PR > 3

Eigenschaften des Googlebots

Fakten

Freshbot besucht eine Seite sehr schnell, nachdem man sich für Ad-Sense angemeldet hat.

Eigenschaften des Googlebots

Fakten

Googlebot kann folgende Dateiformate besuchen und speichern:

html, pdf, asp, jsp, hdml, shtml, xml, cfm, doc, xls, ppt, rtf, wks, lwp, wri.

Eigenschaften des Googlebots

Fakten

Das Schalten von Ad-Words für eine Seite beeinträchtigt NICHT ihr Ranking (offizielles Google Statement)

Eigenschaften des Googlebots

Fakten

Das Ändern der URL einer Seite wird das Ranking der Seite und damit auch die Besuchsfrequenz des Bots verändern, auch wenn die Inhalte gleich bleiben.

Eigenschaften des Googlebots

Fakten

Vorsicht bei Search Engine Optimization

"Wir sind seit 1996 im Internet vertreten und waren eigentlich immer sehr vorsichtig mit unserer Site. Wir wurden dann jedoch fast ein Jahr lang mit E-Mails von sogenannten Link-Programmen bombardiert, in denen uns die Bedeutung solcher Programme für die Relevanz unserer Site angepriesen wurde.

Wir gaben schließlich nach und nahmen einen der Services in Anspruch. Leider arbeitete das Programm mit versteckten Links, und Google entfernte uns innerhalb einer Woche aus dem Index. Es kann jeden treffen, und bevor man sich versieht, ist man auf eine solche Scharlatanerei hereingefallen."

- Frank, Besitzer eines Limousinenverleihs

Eigenschaften des Googlebots

Fakten
Warum besucht Googlebot Seiten nicht ?

- Web Host nicht erreichbar
- Dynamische Seiten mit Sessions vorhanden
- Benutzung von Doorway Pages
- Benutzung von Frames
- robots.txt verbietet es Googlebot
- Googlebot hat Cloaking entdeckt

Danke
:-)

Download der Präsentation:
http://www.daniel-ritter.de/arbeiten/