Freie Universität Berlin


Black-Hat Search Engine Optimization (SEO) Practices for Websites

Damla Durmaz - 29. Januar. 2009

Proseminar Technisch Informatik

Leitung: Georg Wittenburg

Betreuer: Norman Dziengel

Fachbereich Informatik

Inhalt

Einleitung und Motivation

- **Black – Hat SEO:** Methoden der Suchmaschinenoptimierung (SEO), die gezielt gegen die Richtlinien einer Suchmaschine eingesetzt werden

- führen **kurzfristig** zu guten Platzierungen, aber bei Entdeckung im schlimmsten Fall Rauswurf aus dem Index

- Kenntnis über Black-Hat Techniken wichtig, um …
  - sie nicht **versehentlich** einzusetzen
  - sich vor Spammern zu **schützen**

- Warum Google?
  - hoher Arbeitsaufwand für erfolgreiches SEO nötig
  - Marktanteil von Google am größten (siehe Abb. 1)
  - von daher sinnvoll, Optimierung an Google zu richten

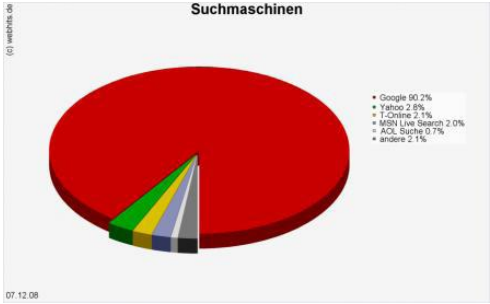

Abb. 1 – Marktanteil von Suchmaschinen[1]

Grundlegendes über Suchmaschinenoptimierung

- Suchmaschinenoptimierung (kurz SEO):
  - Techniken, die auf Webseiten oder auf Teile der Webseiten angewendet werden, um diese zu optimieren und höher zu platzieren[2]

- Ermitteln der Platzierung durch Google:
  - Menge an kleinen automatisch gesteuerten Programmen, die Daten aus dem WWW archivieren (Robots, auch Spider oder Crawler genannt)
  - Indexierung (sortieren und ablegen der Daten in Datenbank - Index)
  - Ermitteln der Platzierung (Linkpopularität, PageRank-Algorithmus, …)

- wichtige SEO – Begriffe:
  - **Keyword**: Begriff, der vom Benutzer in eine Suchmaschine eingegeben wird
  - **Keyworddichte**: Häufigkeit eines Keywords im Verhältnis zur Gesamtzahl der Wörter im Dokument

Grundlegendes über Suchmaschinenoptimierung

- einige Optimierungsmethoden:
  - Onpage-Optimierung (Optimierung des Inhaltes der Seite):
    - Vorkommen von Keywords im Text und Titel
    - Keyworddichte erhöhen (sollte nicht mehr als 8% sein)[3]
    - Einsetzen von Heading-Tags (Markierung von Überschriften, Keywords)
  - Offpage-Optimierung (Optimierung der Umgebung der Seite):
    - gute und übersichtliche Verzeichnisstruktur (Root-Ebene erhält größte Relevanz)
    - Vorkommen der Keywords in der URL
  - Linkpopularität:
    - numerischer Wert: Anzahl aller Links von anderen Seiten auf die eigene Seite (Backlinks) + Anzahl aller internen Links + Qualität der Links
    - Nachteil: leicht manipulierbar (mehrfaches Erstellen sinnloser interner Links)

Der PageRank-Algorithmus

- Googles Reaktion: PageRank-Algorithmus
  - Iterativer Algorithmus
  - Betrachtet nicht nur Linkpopularität, sondern auch **Qualität der Inhaltes**

- Page-Rank-Algorithmus - Grundidee:

$$PR(A) = (1-d) + d\left(\frac{PR(D_1)}{C(D_1)} + \frac{PR(D_2)}{C(D_2)} + \cdots + \frac{PR(D_n)}{C(D_n)}\right)$$

$PR(A)$	PageRank-Wert des Webdokuments A berechnet aus allen eingehenden n Verweisen
A	*zu bewertendes Webdokument*
d	ein Dämpfungsfaktor, der zwischen 0 und 1 liegt(erfahrungsgemäß um 0,85)
$PR(D_1)$	Der PageRank des Dokuments D1, der auf A verweist
$(1-d)$	Wahrscheinlichkeit, dass ein Besucher die Seite verlässt
$C(D_n)$	Anzahl der Verweise, die von D_n ausgehen

- Iteration wird so oft durchgeführt, bis sich die Werte kaum verändern

Der PageRank-Algorithmus

- Beispiel (siehe Abb. 2)[4]:
  - Dämpfungsfaktor d = 0,5
  - PageRank-Werte der Seiten zu Beginn bei 1

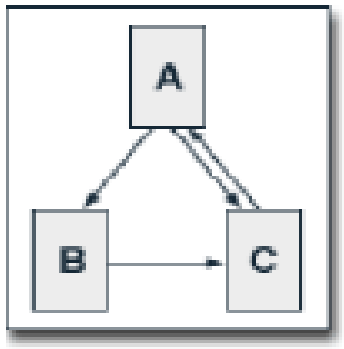

Abb. 2 – Beispiel-WWW

  - erste Iteration:

$$PR(A) = 0.5 + 0.5\,PR(C)$$
$$PR(B) = 0.5 + 0.5\,(\frac{PR(A)}{2})$$
$$PR(C) = 0.5 + 0.5\,(\frac{PR(A)}{2} + PR(B))$$

  - PageRank-Werte nach 12 Iterationen:

Iteration	PR(A)	PR(B)	PR(C)	Sum
0	1	1	1	3
1	1	0.75	1.125	2,875
2	1.0625	0.7656	1.14844	2,977
…				
11	1.07692307	0.76923077	1.15384615	~3
12	1.07692308	0.76923077	1.15384615	~3

  - **aber:** ausgehende Links können den eigenen PageRank **verschlechtern**[5]
    - Summe aller PageRank-Werte bleibt konstant
    - wenn sich ein Wert erhöht, muss der andere automatisch verringert werden

Black-Hat Suchmaschinenoptimierung
Keyword-Stuffing

- Idee: Webmaster wiederholt Keywords übermäßig oft im Dokument, um die Keyworddichte zu erhöhen
  - z.B. in Überschriften
  - Text gleiche Farbe wie Hintergrund

- Wie verhindert Google den Einsatz?
  - Analyse des **Verhältnisses** von Substantiven und Nichtsubstantiven
  - setzen einer **Grenze** für die Keyworddichte (bei ca. 8%)

- Popularität und Effektivität von Keyword-Stuffing:
  - wird schnell entdeckt: schon bei Indexierung wird Keyworddichte betrachtet
  - leider immernoch häufig verwendet:
    - große Keywordlisten, die angeblich zur Produktbeschreibung dienen
    - schreiben von Keywords in extern gespeicherte CSS-Dateien → Googlbot kann diese nicht lesen

Black-Hat Suchmaschinenoptimierung
Doorway-Pages

- Idee: speziell auf Keywords optimierte Seiten (auch Brückenseiten genannt), die nicht für Besucher gedacht sind, sondern der Platzierung der eigenen Seite helfen sollen
  - Googlebot analysiert Doorway-Page und vergibt hohen PageRank
  - anschließend wird ein Link von dieser auf die eigentliche Homepage gesetzt ➔ Homepage erhält besseren PageRank
  - normaler Besucher wird dagegen automatisch auf die Homepage weitergeleitet

- Wie verhindert Google den Einsatz?
  - zeigt z.B. nur zwei Seiten einer Domain in den Ergebnislisten an

- Popularität und Effektivität:
  - haben meistens sinnlosen Inhalt und schlechte Struktur
  - deshalb Erstellen von mehreren 1000 Brückenseiten nötig
  - trotzdem populär (Seiten mit JavaScript können so hochplatziert werden, da Googlebot JavaScript nicht interpretieren kann)

- Idee: Webserver erkennt anhand der Useragent-Informationen und IP-Nummer, ob es sich um einen menschlichen Besucher oder dem Googlebot handelt
  - **menschlicher Besucher:** eigentliche Homepage wird angezeigt
  - **Googlebot:** stark optimierte Seite wird angezeigt

- Wie verhindert Google den Einsatz?
  - bei Verdacht auf IP-Cloaking besucht Googlebot die Seite erneut, aber als menschlicher Besucher getarnt
  - Benutzen von Referrern: sagen, von welcher Seite aus der Server aufgerufen wird

- Popularität und Effektivität:
  - wird immernoch verwendet, leider aufwendig, da IP‘s von Bots oft verändert werden

- Idee: Eine für Google optimierte HTML-Seite wird bei Google angemeldet. Nachdem diese im Index vorhanden ist, wird die Seite mit einer Seite ausgetauscht, die für menschliche Besucher stark optimiert ist (rudimentäre Form von IP-Cloaking)

- Wie verhindert Google den Einsatz?
  - kann nicht von Google erkannt werden
  - Google kann nicht sehen, ob es sich um einen Manipulationsversuch oder wirklich um eine Aktualisierung einer Webseite handelt

- Popularität und Effektivität
  - früher sehr beliebt, mittlerweile aber nutzlos
  - sobald der Googlebot die Seite erneut besucht, wird ein neuer PageRank ermittelt und die Seite wird wieder nach unten platziert
  - **außerdem**: Goolgebots werden immer häufiger versendet

- Idee: Webmaster besuchen über eigene Domain andere Seiten, um ihre Informationen in den Logfiles der Server zu hinterlassen
  - Statistiken der Server-Logfiles werden oft veröffentlicht
  - besuchende Seiten werden als echter Link angegeben ➔ neuer Backlink

- Wie verhindert Google den Einsatz?
  - Google kann Logfile-Spam nicht direkt erkennen und bietet Möglichkeit, **rel=„nofollow"-Atrribut** zu setzen (allerdings sind dann alle Links betroffen)
  - **Einsatz der .htaccess-Datei:** mithilfe eines Moduls können alle typischen Keywords von Logfile-Spammern angegeben werden ➔ enthält ein URL diese, wird sie blockiert

- Popularität und Effektivität:
  - sehr populäre und häufig verwendete Methode, da der Schutz mithilfe der .htaccess-Datei sehr aufwendig ist

- Idee: Stehlen von anderen Webseiteninhalten, um nicht selbst den hohen Arbeitsaufwand einer Optimierung zu leisten
  - spart Zeit
  - unter anderen: wenn mithilfe von Google AdSense Umsatz gemacht werden soll

- Wie verhindert Google den Einsatz?
  - mithilfe von Gleichungen, die zwei Seiten auf doppelten Inhalt überprüfen
  - Webmaster kann mithilfe der .htaccess-Datei etwas tun
    - leider zeitaufwendig, von daher: ein kostenloses PHP-Skript auf www.bot-trap.de erhältlich, der sich selbst mit Spammerkmalen aktualisiert

- Popularität und Effektivität:
  - sehr beliebte, häufige Methode ➔ spart viel Zeit
  - nervend für Betreiber: Überflutung der Logfiles und Manipulation der Rankings

Googles Reaktion
Schutz vor versehentlichem Einsatz von Black-Hat SEO

- zuerst ist es wichtig, **Black-Hat Methoden zu kennen**

- viele kennen Keyword-Stuffing nicht und verwenden viele Keywords
  - kann mit **Tools** umgangen werden (Überprüfung der Keyworddichte einer Seite, Analyse der Häufigkeit der Keywords)

- komplette SEO-Pakte (SEO-Tool IBP):
  - Untersuchung der Keyworddichte
  - Vergleichen von zwei Seiten auf Content-Spam
  - Link-Optimierung
  - Logfile-Analyse

Googles Reaktion
Schutz vor Black-Hat SEO

- guter Schutz vor Logfile-Spam, Content-Spam usw.:
  - Einsatz der .htaccess-Datei

- gute Methode: Benutzen von Google Sitemaps[6]
  - Google Sitemaps liefert Statistiken auf Spam-Verdacht
  - beim Rauswurf sogar detaillierte Gründe, warum es dazu kam

- **Zusammenfassung:**
  - einige Methoden sehr **effektiv** und interessant, trotzdem **auf lange Sicht uneffektiv** ➔ Suchmaschinen passen sich ständig an
  - Konsequenzen für Einsatz enorm hoch (dauerhafter **Bann** aus Index)
  - das Wissen über Black Hat SEO trotzdem enorm wichtig, um sich zu **schützen** oder nicht selbst aus Versehen diese Techniken zu verwenden

Quellen

[1] http://www.webhits.de/deutsch/index.shtml?webstats.html Zugriff am 26.12.2008 um 15:00Uhr

[2] J. Winkler, *Suchmaschinenoptimierung*, Franzis Verlag GmbH: Poing, 2007 , S. 7

[3] http://www.tecchannel.de/webtechnik/entwicklung/1766517/

google_optimierung_die_schmutzigen_tricks/index3.html, Zugriff am 1.1.2009 um 17:22Uhr

[4] http://www.aifb.uni-karlsruhe.de/Lehre/Sommer2006/kdtm/stuff/

Google-PageRank-V1.0.pdf, Zugriff am 10.1.2009 um 08:230Uhr, S. 20

[5] http://www.aifb.uni-karlsruhe.de/Lehre/Sommer2006/kdtm/stuff/

Google-PageRank-V1.0.pdf, Zugriff am 11.1.2009 um 13:08Uhr, S. 38

[6] https://www.google.com/webmasters/tools/docs/de/about.html, Zugriff am 26.01.2009 um 22:36Uhr

Freie Universität Berlin


Black-Hat Search Engine Optimization (SEO) Practices for Websites

Damla Durmaz - 29. Januar. 2009

Proseminar Technisch Informatik

Leitung: Georg Wittenburg

Betreuer: Norman Dziengel

Fachbereich Informatik