Denne guide er oprindeligt udgivet på Eksperten.dk

COMPUTERWORLD

Kontroller søgemaskinernes adgng

Denne artikel bekriver kort og overskueligt, hvorledes man kan bestemme hvilke dele af en hjemmeside søgemaskiner skal scanne, samt hvilke de skal holde sig fra

Skrevet den **15. Apr 2009** af **arkanoid** I kategorien **Generelt IT / Søgemaskiner** | ★★★☆☆

Der findes en skjult, almægtig kraft, der kravler rundt på nettet i en konstant søgen efter nye informationer. Jeg taler selvfølgelig om de såkalde web-crawlere, programmer der automatisk scanner hjemmesider og indekserer indholdet til brug for søgemaskiner.

Disse web-crawlere, eller botter som de også kaldes, kan både komme fra søgemaskiner med reelle hensigter, men desværre også fra samvittighedsløse spammere, der scanner efter e-mailadresser og andre informationer, de kan bruge til at reklamere.

Botterne (eller i det mindste de godartede af dem) kan dog styres, så de kun scanner de dele af din hjemmeside, som du ønsker at få lagt ud på søgemakinernes sider. De finder ud af, hvad de skal gøre v.h.a. en lille tekstfil på din server. En tekstfil ved navn *robots.txt*

I det efterfølgende, vil jeg give en introduktion til hvordan *robots.txt* filen virker.

Lad os starte med det helt grundlæggende. Åbn din yndlings-teksteditor (*Notepad* eller *vi* er udmærkede til dette formål) og opret en fil. Kald filen for *robots.txt*. Det er vigtigt at navnet er præcis det samme som her, eller er det bare en værdiløs tekstfil.
Dernæst skal filen ligges i dit roddrev på serveren. Dvs. *http://minside.dk/* er rigtigt, hvorimod den ikke vil virke, hvis den ligger i f.eks. *http://minside.dk/subfolder*

Godt. Når det er klaret, kan vi gå i gang med at skrive selve indholdet af filen. Der er to elementer som du skal bruge i første omgang.
Det første er **User-agent**, den definerer hvilke crawlere der skal følge oplysningerne nednunder. F.eks. kunne vi skrive:
User-agent: *
Det ville betyde at samtlige crawlere skal følge de regler, der efterfølgende bliver udstukket ved hjælp af elementet **Disallow**. Disallow skal efterfølges af stien til de mapper eller filer, man ønsker at crawleren skal undgå.

Eksempel 1:
User-agent: *
Disallow: /

Dette eksempel ville fortælle **alle** crawlere, at hele siden skal udelukkes fra crawling (gennemsøgning).

Det er dog sjældent at vi ligefrem ønsker hele siden udelukket fra gennesøgning. Lad os f.eks. sige at vi ønsker mappen *private* udelukket. Så vil det se således ud:

Eksempel 2:
User-agent: *
Disallow: /private/

Det er vigtigt at huske skråstregen på begge sider af mappenavnet

Er der nu flere mapper på serveren, som vi ønsker at søgemaskinerne skal undgå, laver vi blot en ny linje og skriver den næste ind:

Eksempel 3:
User-agent: *
Disallow: /private/
Disallow: /images/

Man kan også vælge at udelukke enkelte filer. Det gøres på samme måde:

Eksempel 4:
User-agent: *
Disallow: /private/
Disallow: /images/
Disallow: /admin/login.php

Bemærk at der **ikke** skal bruges en skråstreg efter et filnavn.

Nå, nu er det vist på tide at begynde at diskriminere lidt. Selv om at alle webmastere elsker Google, er det langt fra alle, der bryder sig om Googles billedsøgning, da denne direct-linker til billeder, og derfor koster ejeren båndbredde, hver gang nogen søger på ens billeder. Det vi så gør, er at bede botten *Googbot-Image* om at udelukke hele siden fra søgning:

Eksempel 5:
User-agent: Googlebot-Image
Disallow: /

De forskellige kommandoer kan sagtens kombineres som man lyster, og på samme måde, som man kan stille flere betigelser op, for en søgemaskine, kan man også stille betingelser for flere søgemaskiner:

Eksempel 6:
User-agent: *
Disallow: /
User-agent: Scooter
Disallow: /private/
Disallow: /not_for_scooter/priv.php

Det interessante ved ovenstående stykke kode, er at vi kan se, at elementerne ikke nedarves. Det første element, der beskriver alle botter, bliver altså ignoreret for Googlebot's tilfælde. Det sætter os i stand til at lave nogle ret komplicerede betingelser.

Man kan rent faktisk også, v.h.a. af *Disallow* sætte en søgemaskine i stand til at indeksere det hele:

Eksempel 7:
User-agent: *
Disallow: /
User-agent: Scooter

Disallow:

Ved at lade feltet efter *Disallow:* være tomt bliver alt sat til *Allow*, altså tilladt for botten Scooter's (Altavistas crawler) tilfælde.

Derudover er der nogle få crawlere, der understøtter et tredje element. Elementet hedder **Allow** og bladt dem der understøtter det, er Googles Googlebot

Eksempel 8:
User-agent: *
Disallow: /
User-agent: Googlebot
Allow: /

Dette er, ifølge

Googles FAQ for webmastere

den foretrukne måde at undgå alle crawlere undtagen Google.

Hvis man, af en eller anden grund, ikke ønsker at bruge robots.txt filen, er der også en anden mulighed. Man kan i stedet sætte oplysningerne ind i et meta-tag. Som med alle andre Meta-tags skal dette placeres mellem <head> og </head> tags'ne og skrives eksempelvis således:
<meta name="robots" content="noindex,nofollow" />
Bemærk at *name* attributten er "*robots*" og *content* består af *(no)index* og *(no)follow*

Metatags:

1) Denne sørger for at botter ikke indekserer siden og heller ikke følge links
<meta name="robots" content="noindex,nofollow" />

2) Denne sørger for at siden bliver indekseret, men at link ikke bliver fulgt
<meta name="robots" content="index,nofollow" />

3) Denne sørger for at siden ikke bliver indekseret, men at link blvier fulgt
<meta name="robots" content="noindex,follow" />

4) Til sidst er der en lidt kortere måde at deklare 1) på:
<meta name="robots" content="none">

Spørgsmål, kommentarer, ris og ros modtages gerne (det er min første artikel, så døm nu ikke for hårdt :-))

Nyttige links:

Liste over søgemaskiner og deres crawlere
Validér din robots.txt fil

Kommentar af a1 d. 30. Dec 2004 | 1

rigtig god forklaring af robots.txt....., men det bør nok nævnes at det IKKE er alle spiders der "respekterer" robots.txt, og hvis de sætter deres "user-agent" til f.eks. det samme som IE 6, har du meget svært ved at slippe for at blive indexeret på nogen måde.
Er der ikke også en: <meta name="robots" content="all"> ?

Kommentar af honne d. 30. Dec 2004 | 2

Kommentar af lenk d. 30. Dec 2004 | 3

glimrnede aftikel om robots.txt.

Kommentar af espersen d. 30. Dec 2004 | 4

absolut en fin artikel - og nødvendig ifm. hjemmeside-design/SEO. Men havde håbet på en del mere end blot Robots.txt

Kommentar af phatlasse (nedlagt brugerprofil) d. 30. Dec 2004 | 5

Rigtig gode, korte og præcise tips, som enhver der laver hjemmesider burde læse.

Kommentar af jahamnej d. 30. Dec 2004 | 6

Genialt, skulle faktisk lige til at søge på dette emne. Hehe, nogle gange er man heldig at vælte ind på eksperten på det rigtige tidspunkt. Takker!

Kommentar af morteeart d. 30. Dec 2004 | 7

ah, genial artiklen. Det er en forklaring på brugen af robots.txt som er YDERST VIGTIGT for både dig selv, og alle andre søgere på nettet.
Bør læses af alle som laver seriøse websider.

Kommentar af -kick- d. 29. Dec 2004 | 8