

Контролиране процеса на обхождане на страниците

проф. д-р инж. Христо Вълчанов

<http://cs.tu-varna.bg>

Обхождане и индексирание

- **Обхождане** - анализ на web страниците и следване на линковете в тях.
- **Индексирание** - събиране на информация за страница, така, че тя да бъде видима в резултатите от търсенето.

Необходимост от контрол

- Не е желателно да се позволява на crawler-ите да имат достъп до определени области на сървъра.
- Да се появяват резултантните страници без snippet.
- Ако потребителите не трябва да виждат кеширана версия на страница.

Контролиране процеса на обхождане - Robots Exclusion Protocol

- Протоколът е дефиниран от файла *robots.txt*, който трябва да се намира в контролирания сайт.
- Файлът включва директиви, които съдържат указания към web crawler-ите до кои директории или файлове на сайта да имат достъп.
- Местоположението на файла е в кореновата директория на съответния web сайт.

Формат на robots.txt

- Текстов файл в ASCII или UTF-8 формат.
- Съдържа последователност от записи (редове), разделени със символите за край на ред CR, CR/LF или LF.
- Всеки запис включва поле, двоеточие и стойност

`<field> : <value> <#optional-comment>`

БНФ на robots.txt

```
robotstxt ::= { entries }
entries ::= startgroupline { startgroupline } { groupmemberline | nongroupline | comment } |
          nongroupline | comment
startgroupline ::= user-agent : agentvalue [comment] EOL
groupmemberline ::= ( pathmemberfield : pathvalue | othermemberfield : textvalue )
                    [comment] EOL
nongroupline ::= ( urlnongroupfield : urlvalue | othernongroupfield : textvalue ) [comment] EOL
comment ::= # { anychar }
agentvalue ::= textvalue
pathmemberfield ::= disallow | allow
othermemberfield ::=  $\epsilon$ 
urlnongroupfield ::= sitemap
othernongroupfield ::=  $\epsilon$ 
pathvalue ::= / path
urlvalue ::= absoluteURI
textvalue ::= { valuechar | SPACE }
valuechar ::= <any UTF-8 character except ("#" CTL)>
anychar ::= <any UTF-8 character except CTL>
EOL ::= CR | LF | CR LF
```

Видове записи

- Записите са категоризирани в няколко типа в зависимост от типа на елемента *field*.
 - **start-of-group**;
 - **group-member**;
 - **non-group**.

start-of-group : елементът *user-agent* се използва да укаже за кой *crawler* дадената група е валидна

Видове записи - 2

`user-agent: a`
`disallow: /c`

`user-agent: b`
`disallow: /d`

`user-agent: e`
`user-agent: f`
`disallow: /g`

Директиви

- Записите тип **group-member** са известни като *директиви* и се задават във вида:

`<element> : <path>`

Директиви - 2

- **disallow** – указва директория или файл, до които е забранен достъп за зададения web робот.
- **allow** – позволява прилагането на по-прецизен контрол на достъпа.
- **crawl-delay** – изисква роботите да изпълняват пауза между последователните заявки за страници. Паузата се задава в секунди, като се препоръчва да е в диапазона от 1-10sec.
- **sitemap** – Задава URL на XML файл, който съдържа информация за структурата на даден web сайт. Тази директива е от типа **non-group** и като такава се обработва независимо от група.

Шаблони

- * - задава нула или повече появявания на произволен символ;
- \$ - задава край на URL

Примери

Позволява на всички роботи да обхождат сайта

User-agent: *

Disallow:

Забранява достъпа на всички роботи

User-agent: *

Disallow: /

Забранява достъпа на всички роботи до директориите

/myfolder, /mytest и техните поддиректории

User-agent: *

Disallow: /myfolder/

Disallow: /mytest/

Забранява достъпа на всички роботи с изключение на GoogleBot

User-agent: googlebot

Disallow:

User-agent: *

Disallow: /

Примери

```
# Забранява достъпа на всички роботи до /scripts  
# с изключение на "page.php"  
User-agent: *  
Disallow: /scripts/  
Allow: /scripts/page.php
```

```
# Указва използване на sitemap файл  
User-agent: *  
Disallow:
```

```
Sitemap: http://www.example.com/sitemap.xml
```

Примери

```
# Забранява достъпа на робот GoogleBot до URL,  
# съдържащ "page"  
User-agent: googlebot  
Disallow: /*page
```

```
# Забранява достъпа на робот GoogleBot до всички  
# asp файлове. Няма да се изключат файлове или  
# директории със съдържащи се запитвания заради "$"  
User-agent: googlebot  
Disallow: /*asp$
```

```
# Изключва се: /pretty-wasp  
# Изключва се: /login.asp  
# Не се изключва: /login.asp?forgotton-password=1
```

Контролиране процеса на индексиране чрез мета тагове

Указват дали дадената страница да бъде индексирана и как тя да се появи в резултатите от търсенето. Таговете се поставят в секцията <head> на страницата.

```
<!DOCTYPE html>
<html>
<head>
    <meta name="robots" content="noindex" />
    ...
</head>
<body> ... </body>
</html>
```

Мета тагове

- Забрана само на стандартния робот на Google да индексира страницата:

```
<meta name="googlebot" content="noindex" />
```

- Появяване на страница в резултатите на Google, но не и в Goggle News:

```
<meta name="googlebot-news" content="noindex" />
```

- Индивидуално конфигуриране на различни crawler-и:

```
<meta name="googlebot" content="noindex" />
```

```
<meta name="googlebot-news" content="nosnippet" />
```

- Множество директиви:

```
<meta name="googlebot" content="noindex, nofollow" />
```


Мета тагове

- Сумарно използване на директивите:

```
<meta name="robots" content="nofollow">
```

```
<meta name="googlebot" content="noindex">
```

Контрол чрез хедър X-Robots-Tag

- Съдържа се в HTTP отговора от web сървъра.
- Всички директиви от мета таговете са приложими и тук.
- Необходима е настройка в конфигурационния файл на съответния web сървър (*httpd.conf*).
- Директивите ще се приложат към целия сайт.

Пример:

Включване на директиви *noindex*, *nofollow* в хедъра **X-Robots-Tag**, които ще се приложат за всички .PDF файлове в целия сайт.

```
<Files ~ "\.pdf$">  
    Header set X-Robots-Tag "noindex, nofollow"  
</Files>
```

Примери

- Crawler-ите да не индексират страницата:

```
HTTP/1.1 200 OK
Date: Tue, 27 Jan 2015 21:42:43 GMT
...
X-Robots-Tag: noindex
...
```

- Множество хедъри или списък с директиви:

```
HTTP/1.1 200 OK
Date: Tue, 27 Jan 2015 21:42:43 GMT
...
X-Robots-Tag: noindex, noarchive
X-Robots-Tag: unavailable_after: 27 Jun 2015 15:00 PST
...
```

- Прилагане само към конкретен crawler:

```
HTTP/1.1 200 OK
Date: Tue, 27 Jan 2015 21:42:43 GMT
...
X-Robots-Tag: googlebot: nofollow
X-Robots-Tag: otherbot: noindex, nofollow
...
```

Възможни директиви

- **noindex** – страницата няма да се включи в резултатите от търсенето и няма да се кешира (в резултатите няма да се изведе „Cached” линк).
- **nofollow** – не се обхождат линкове в страницата.
- **none** – еквивалентно на горните две.
- **nosnippet** – не се показва snippet в резултатите от търсенето.
- **noimageindex** – не се индексират изображенията в страницата.
- **notranslate** – да не се превежда страницата в резултатите.
- **unavailable_after** – страницата няма да се покаже в резултатите от търсенето след определените дата/време, зададени съгласно стандарта RFC850.

Въпроси?