

# Scripting, Databaser og Systemarkitektur, E2007

## Forelæsning

- ▶ Publicering og Annoncering af web-sites
  - ▶ Exponering af web-site med søgemaskiner
- ▶ User Tracking: Hvor kommer brugerne fra?
  - ▶ Statistik fra access-log
  - ▶ Banner-add click-throughs, kort
- ▶ Gennemgang af prøveeksamen F2007 (VoteAboutIt.com)

## Mange muligheder:

- ▶ Bannerreklamer — specielt i søgemaskiner (køb af ord)
  - ▶ <http://www.google.com>
- ▶ Prime time TV
- ▶ Aviser/blade
- ▶ Web-biblioteker
  - ▶ <http://www.submit-it.com>
- ▶ Søgemaskiner

En søgemaskine består typisk af tre dele:

- ▶ web-crawler
- ▶ database of URL's
- ▶ query processor

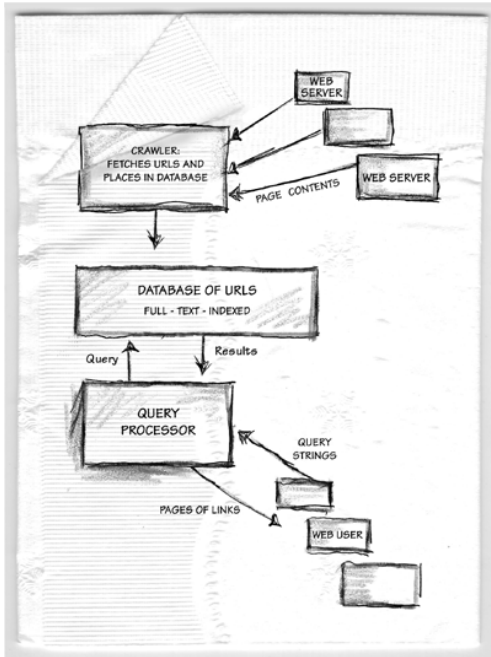
Databasen indeholder bl.a. frekvenstabeller over hvor tit ord forekommer i URL'er:

F.eks. vil teksten "**Home page for the course**"

generere følgende frekvenstabel:

<b>Ord</b>	<b>Frekvens</b>
Home	1
page	1
course	1

**Mange brugere vil komme fra søgemaskiner!**



# Du kan forbedre dine chancer for at stå øverst

- ▶ Det gælder om at stå øverst i søgemaskinens resultat
- ▶ Søgemaskiner sælger ud af ord til dem der vil betale penge for at stå øverst i søgeresultater

Hvis du ikke har råd til at købe ord kan du sørge for at der er indhold på de sider du vil have indekseret

Søgemaskiner forstår ikke billeder — endnu!

*Uærlig forbedring af dine chancer:*

```
<META name="keywords" content="sex sex money fast money money money money money money money fast fast fast">
```

Det kan nok ikke betale sig at fylde sin side med tonsvis af *keywords*

*Ærlig forbedring af dine chancer:*

```
<META name="description" content="Journal for sophisticated Web publishers, specializing in RDBMS-backed sites.">
```

*Nogle gode grunde til at skjule indhold:*

- ▶ Mirror-sites
- ▶ Privat dokument delt blandt få personer
- ▶ Mapper med filer
- ▶ Intranettet
- ▶ Administrationsmodulet
- ▶ Ressourcetunge sider der skifter hyppigt (søgning)

# Skjul indhold for robotterne

## *Hvordan man skjuler indhold med vilje:*

Det er muligt at instruere web-crawlere om ikke at søge i bestemte filer.

The Robots Exclusion Protocol: (`robots.txt`, som skal ligge i roden af dit site):

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /admin/
```

The Robots META tag:

```
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
```

Se evt: <http://www.whitehouse.gov/robots.txt>

*Bemærk:* Der er ingen garanti for at søgemaskiner overholder dine retningslinier!

# Total Exposure

*Til tider skjules indhold på trods af at det ikke var hensigten:*

- ▶ Hvis du kræver registrering
- ▶ Hvad med PHP-scripts og data i databasen?
- ▶ Flash og JavaScript genereret indhold

Løsning:

- ▶ Non-obtrusive JavaScript
- ▶ Exporter data til dummy HTML-filer som web-crawlere kan se
- ▶ Konstruer indexes fra disse sider til relevante PHP-scripts, således at brugerne ser den rette information.

Kom på forsiden af

- ▶ `www.cnn.com`
- ▶ `www.newyorktimes.com`
- ▶ `www.dr.dk`
- ▶ ...

Forskellige webservere (Apache, AOLserver, ...) benytter samme standardformat for access-logs.

Følgende oplysninger kan hentes fra en web-servers access-log:

- ▶ Typen af browser en bestemt bruger benytter
- ▶ Antal brugere som har efterspurgt ikke-eksisterende filer — og hvor de har URL'erne fra
- ▶ Antal brugere der efterspørger en bestemt fil
- ▶ Tiden en bruger i gennemsnit bruger på en fil før brugeren fortsætter med en anden fil
- ▶ Antal brugere der klikker på bestemte banner-adds
- ▶ Kommer en bruger tilbage?

# User Tracking: Hvor kommer brugerne fra og hvor mange hits er der på mit web-site?

Se webserverens access-logs:

```
74.6.29.25
```

```
[27/Nov/2007:00:36:14 +0100]
```

```
"GET /index.php HTTP/1.0"
```

```
200
```

```
9624
```

```
"http://www.google.com/search?q=greenpeace.ppt&btnG=Pesquisar&lr="
```

```
"Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com...)"
```

```
XX.XX.XXX.XXX
```

```
[27/Nov/2007:00:36:47 +0100]
```

```
"GET /subpage.php3?page=http://utenti.lycos.it/ HTTP/1.1"
```

```
404
```

```
7314
```

```
"_"
```

```
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET...)"
```

# Ikke-eksisterende filer

*Eksempel:* bruger indtaster en forkert URL direkte i browserens "location-bar" Søg efter 404 (File Not Found) i access-log:

```
130.226.141.250
```

```
[17/Feb/2000:15:51:29 +0100]
```

```
"GET /temperatur.html HTTP/1.1"
```

```
404
```

```
212
```

```
"_"
```

```
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET...)"
```

- ▶ Filnavnet `temperatur.html` skulle have været `temperature.html`
- ▶ Brugeren benytter IE (MSIE 7.0) på en Windows XP maskine
- ▶ Med `nslookup` kan det ses at 130.226.141.250 svarer til `stud250.itu.dk`:

```
# nslookup 130.226.141.250
```

```
Server: ns000.worldonline.dk
```

```
Name: stud250.itu.dk
```

# Ikke-eksisterende fil p.g.a. forkert link

Vi søger igen efter 404 (File Not Found) i access-log:

```
213.237.71.166 - - [20/Mar/2007:02:10:46 +0100]
"GET /F2001/lec8/list2v.php HTTP/1.0" 404 456
"http://hug.itu.dk:8077/SlideExtractor/slide_extractor.php"
"Mozilla/4.73 [en] (X11; U; Linux 2.2.14-12 i686)"
```

- ▶ Filen `/list2v.php` skulle eks. have været `/listv2.php`
- ▶ Brugeren benytter Netscape v. 4.73 på en Linux maskine (kerne v. 2.2.14-12).
- ▶ Vi kan se hvor brugeren kommer fra (Referer)
- ▶ Fejlen skyldes et forkert link fra scriptet `slide_extractor.php`.
  - ▶ Vi kan altså se hvilken anden side brugeren kommer fra
  - ▶ Dvs: vi kan se hvilke sider der indeholder links der ikke virker

# Banner-add click-throughs

- ▶ Banner-add ejer har en anden opfattelse af antal click-throughs end din access-log indikerer.

Det er kun access-loggen ved Someone Else's Site der ved hvem der clickede et banner på din side (via "referrer").

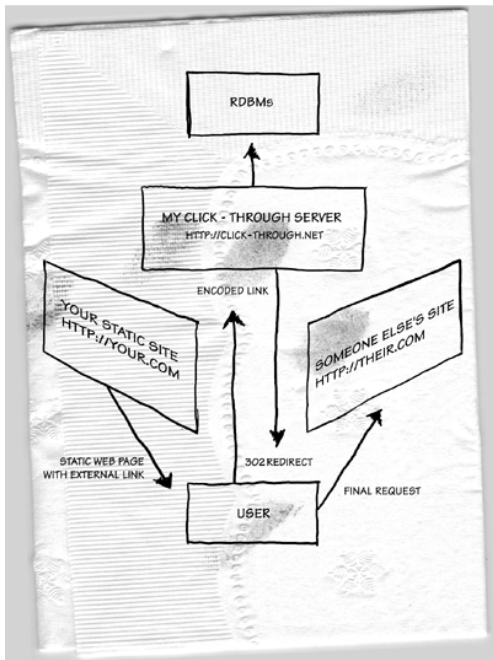
Men det er jo Someone Else's Site der skal betale dig!

- ▶ Brug en uafhængig click-through server - se figur på næste slide.

I stedet for at linke `http://their.com`, så anvend

`http://clickthrough.net/?from=itu.dk&send_to=http://their.com`.

1. Click til `their.com` fra din side registreres i databasen
2. Derefter laves en "redirect" til `their.com`.



Se

- ▶ [http://www.itu.dk/courses/DSDS/E2007/exms/f2007/exm\\_f2007.pdf](http://www.itu.dk/courses/DSDS/E2007/exms/f2007/exm_f2007.pdf)
- ▶ [http://www.itu.dk/courses/DSDS/E2007/exms/f2007/exm\\_f2007\\_vejl.pdf](http://www.itu.dk/courses/DSDS/E2007/exms/f2007/exm_f2007_vejl.pdf)

Vi gennemgår eksamenssættet og den vejledende løsning fra ovenstående links