

Internetsøgemaskiner



Gerth Stølting Brodal

Datalogisk Institut
Aarhus Universitet

Overblik

- Indledning
- Google facts
- Internetgrafen
- En søgemaskines dele
 - Crawling
 - Indeksering
 - Søgning og ranking
- Afslutning

Internettet

- Meget stor mængde ustruktureret information.
- Hvordan finder man relevant info? Søgmaskiner!

94: Lycos,...

96: Alta Vista: mange sider.

99: Google: mange sider og god ranking.

Søgemaskinernes Barndom

princess diana

Engine 1

[Princess Diana Memorial WebRing](#)

Follow the WebRing for a tour of memorial site
87% <http://www.geocities.com/RainForest/Vines/1009/diana>
1998

[Grouped results from http://www.geocities.com](#)

[FOR DIANA, PRINCESS OF HEART - Dr. K](#)

...
Dr. Kate Wachs Comments on Princess Diana T
84% <http://www.therelationshipcenter.com/diana.shtml> (Si

[Princess Diana Editorial Cartoons! Cartoons a](#)

The Professional Cartoonists Index is the most c
cartoonists o
daily cartoon
82% <http://www>

[Diana, Princess of Wales](#)

1 July 1961 - 31 August 1997 The BBC Web sit
Camera Press/Snowdon
79% <http://www.royal.gov.uk/start.htm> (Size 2.3K) Doc
[Grouped results from http://www.royal.gov.uk](#)

Relevant and
high quality

Engine 2

1. [Re: Lost in the shadow of Princess Diana](#)

[URL: www.spiceisle.com/talkshop/messages/6232.htm]
The Spicelander TalkShop. [Follow Ups] [Pos
The Spicelander TalkShop] Date: September
00:54:03 From: Sno,...
Last modified 12-Sep-97 - page size 4K - in English [Tran

2. [Re: Princess Diana's gown auction](#)

[URL: www.elle.com/textes/blablaba/forum/messages/1/15]
Re: Princess Diana's gown auction. [Follow Ups
Followup] [Elle International - Blablaba] Posted
September 07, 1997 at 02:15:26:..
Last modified 30-Mar-98 - page size 2K - in English [Tran

3. [Re: Princess Diana](#)

[URL: spicyhot.com/gaynet/messages/1053.html]
Re: Prince
Maine Ga
Novembe
Last modifi

4. [Re: Princess Diana - Queen of Hearts](#)

[URL: www.elle.com/textes/blablaba/forum/messages/1/28]
Re: Princess Diana - Queen of Hearts. [Follow U
Followup] [Elle International - Blablaba] Posted
on August 31, 1997 at..
Last modified 30-Mar-98 - page size 4K - in English [Tran

Relevant but
low quality

Engine 3

1. [Free Passwords To Adult Sites ...](#)

99% - **Articles & General info:** Free Passwords
Sites warez princess diana demi moore
magazine kathy ireland lingerie jennifer ariston cook
warez princess diana demi moore... 03/09/98
Commercial site: <http://www.purient.com/warez>

2. [SEX CHAT XXX NUDE PORNO PLAYBOY P](#)

[AMERICAN PORN FREE PICTURES WOMEN](#)
99% - **Articles & General info:** SEX CHAT XXX
PORNO PLAYBOY PAMELA ANDERSON P
PICTURES WOMEN ADULT MUSIC CHAT B
BROTICA SHOT MCCARTHY LEHURSE SA
CREDIT CRAWFORD STEE GIBL... 03/09/98
Personal page: <http://www.connix.com/~wgonzo/sex/slidesuperall.htm>

3. [Ro](#)

[SEX CHAT XXX NUDE PORNO PLAYBOY PAMELA ANDERSON PICTURES WOMEN ADULT MUSIC CHAT BROTICA SHOT MCCARTHY LEHURSE SA CREDIT CRAWFORD STEE GIBL... 03/09/98](#)
Personal page: <http://www.octet.com/~gonzo/jy>

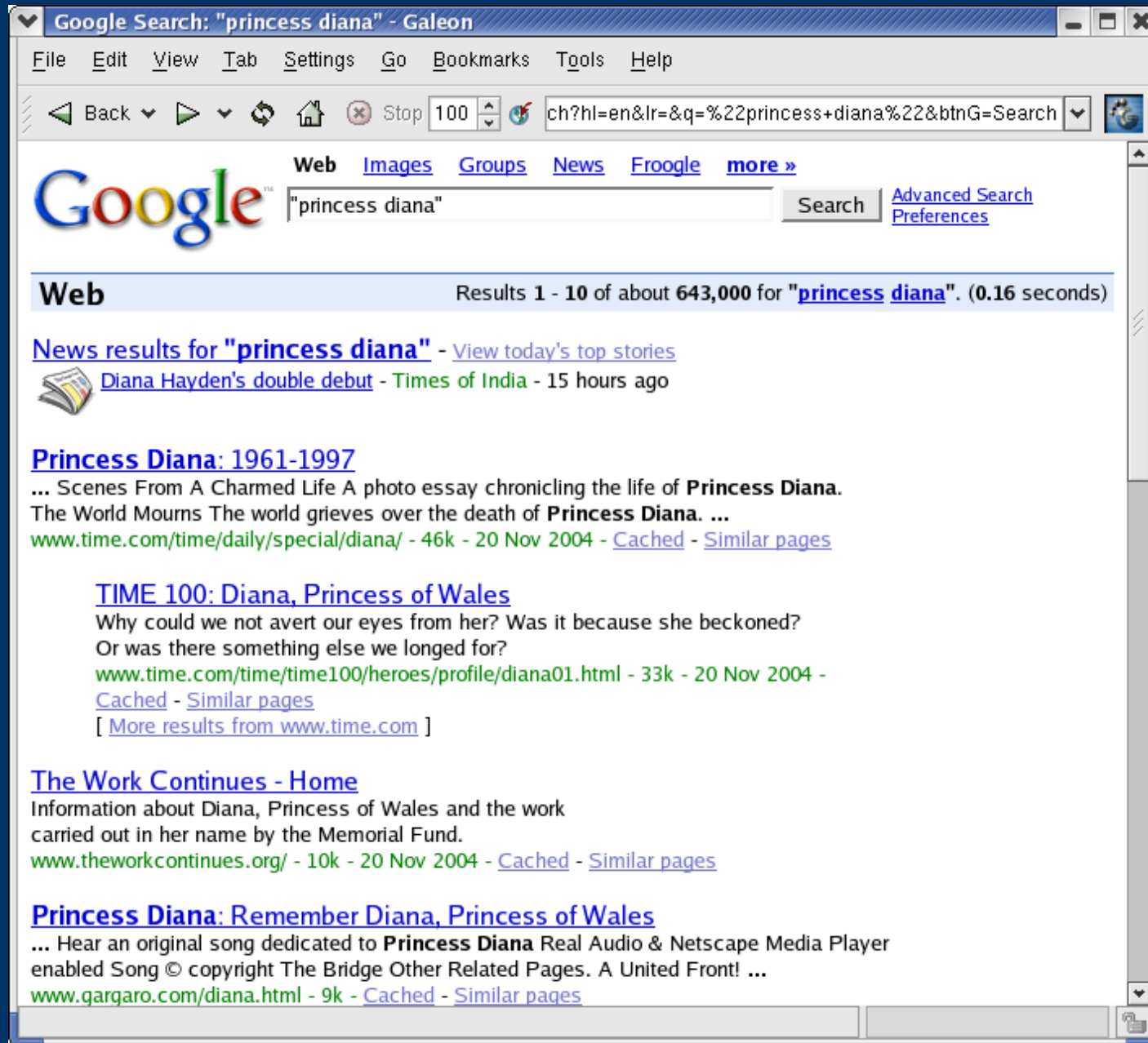
4. [Sunday, 18-Jan-98](#)

99% - **Articles & General info:** Sunday, 18-Jan-
CHAT XXX NUDE PORNO PLAYBOY PAME

Not relevant
index pollution


[Fra: Henzinger, 2000]

... og Søgemaskiner i 2004



The screenshot shows a Galeon browser window with the title "Google Search: 'princess diana' - Galeon". The address bar contains the search URL: "ch?hl=en&lr=&q=%22princess+diana%22&btnG=Search". The search results page displays the Google logo and the search query "princess diana". The results are categorized under "Web" and show "Results 1 - 10 of about 643,000 for 'princess diana'. (0.16 seconds)".

Web Results 1 - 10 of about 643,000 for "princess diana". (0.16 seconds)

[News results for "princess diana"](#) - [View today's top stories](#)
 [Diana Hayden's double debut](#) - [Times of India](#) - 15 hours ago

[Princess Diana: 1961-1997](#)
... Scenes From A Charmed Life A photo essay chronicling the life of **Princess Diana**.
The World Mourns The world grieves over the death of **Princess Diana**. ...
www.time.com/time/daily/special/diana/ - 46k - 20 Nov 2004 - [Cached](#) - [Similar pages](#)

[TIME 100: Diana, Princess of Wales](#)
Why could we not avert our eyes from her? Was it because she beckoned?
Or was there something else we longed for?
www.time.com/time/time100/heroes/profile/diana01.html - 33k - 20 Nov 2004 -
[Cached](#) - [Similar pages](#)
[[More results from www.time.com](#)]

[The Work Continues - Home](#)
Information about Diana, Princess of Wales and the work
carried out in her name by the Memorial Fund.
www.theworkcontinues.org/ - 10k - 20 Nov 2004 - [Cached](#) - [Similar pages](#)

[Princess Diana: Remember Diana, Princess of Wales](#)
... Hear an original song dedicated to **Princess Diana** Real Audio & Netscape Media Player
enabled Song © copyright The Bridge Other Related Pages. A United Front! ...
www.gargaro.com/diana.html - 9k - [Cached](#) - [Similar pages](#)

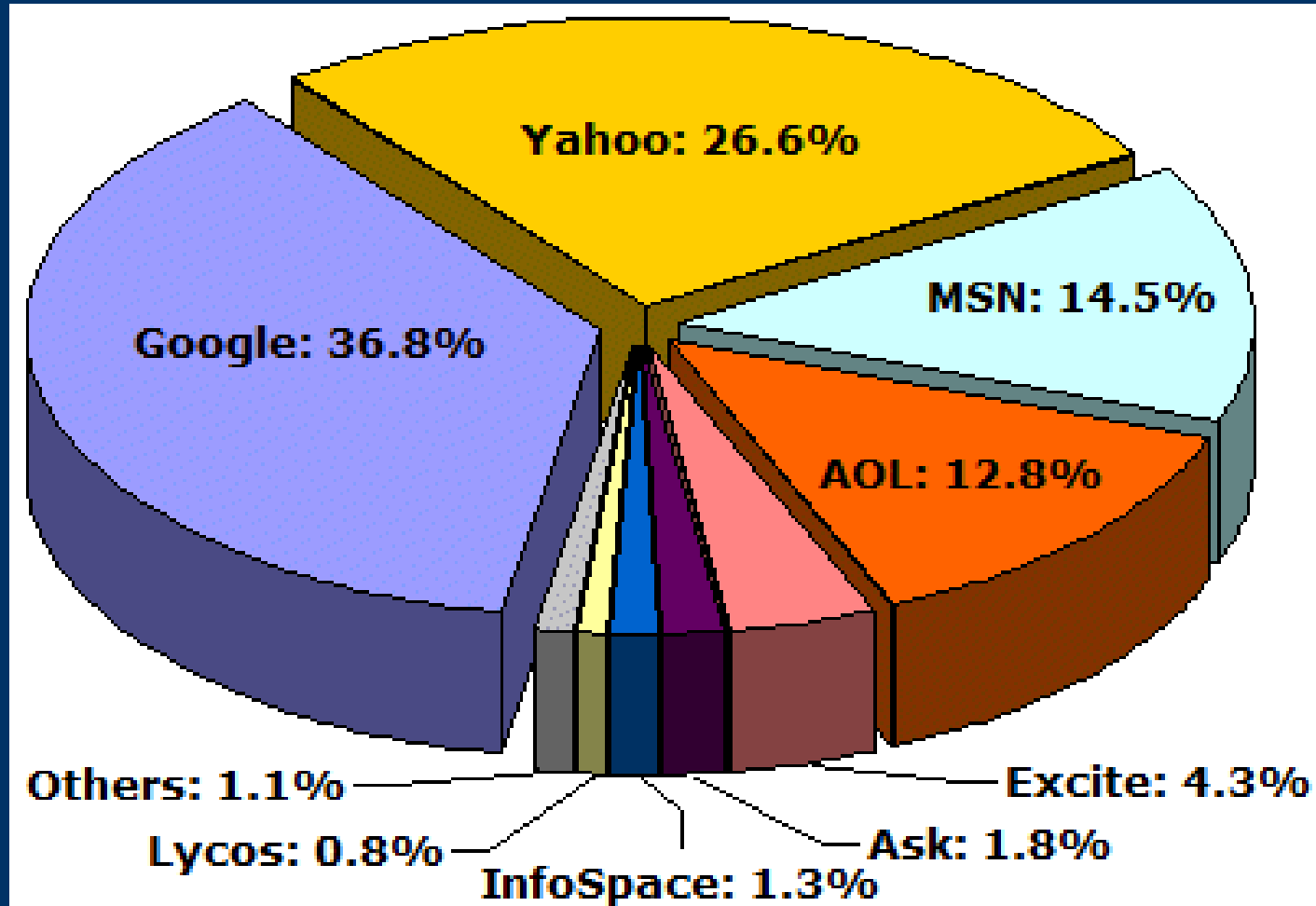
Moderne Søgemaskiner

Imponerende performance. F.eks. Google:

- Søger i $8 \cdot 10^9$ sider.
- Svartider $\approx 0,1$ sekund.
- 1000 brugere i sekundet.
- Finder relevante sider.

I'm Feeling Lucky

Internetsøgninger i USA (maj 2004)



[Fra: www.searchenginewatch.com]

Overblik

✓ Indledning

- Google facts
- Internetgrafnen
- En søgemaskines dele
 - Crawling
 - Indeksering
 - Søgning og ranking
- Afslutning

Google™

- Startet i 1995 som forskningsprojekt ved Stanford University af ph.d. studerende **Larry Page** og **Sergey Brin**
- Privat firma grundlagt 1998
- +5.000 medarbejdere
- Ansvarlig for ca. halvdelen af alle internet-søgninger
- Hovedsæde i Silicon Valley



google \approx googol = 10^{100}

Søgemaskine

- Hurtig
- Relevante links
- Opdateret
- Cache
- GoogleScout (lignende sider)
- Automatisk stavekontrol
- Interface til WAP og PDA
- Produktsøgninger (froogle)
- Billed søgning
- Aktiekurser, kort, ordbøger, nyheder, telefonbøger ...
- Stemmestyret teknologi

AdWords

- tekstbaseret reklame
- query afhængig

Licenser

- AOL/Netscape, Red Hat, Virgin Group, YAHOO, The Washington Post ...

Web API

Hardware + Software





- +8.000.000.000 web sider (+20 TB)
- PageRank: +3.000.000.000 sider og +20.000.000.000 links
- +35.000.000 ikke HTML sider
- +845.000.000 USENET beskeder (20 år)
- +880.000.000 billeder
- +2 Terabyte index, opdateres en gang om måneden
- +2.000.000 termer i indeks
- +150.000.000 søgninger om dagen (2000 i sekundet)
- +200 filtyper: HTML, Microsoft Office, PDF, PostScript, WordPerfect, Lotus ...
- +28 sprog



- Cluster af +10.000 Intel servere med Linux
 - Single-processor
 - 256 MB–1 GB RAM
 - 2 IDE diske med 20-40 Gb
- Fejl-tolerance: Redundans
- Hastighed: Load-balancing

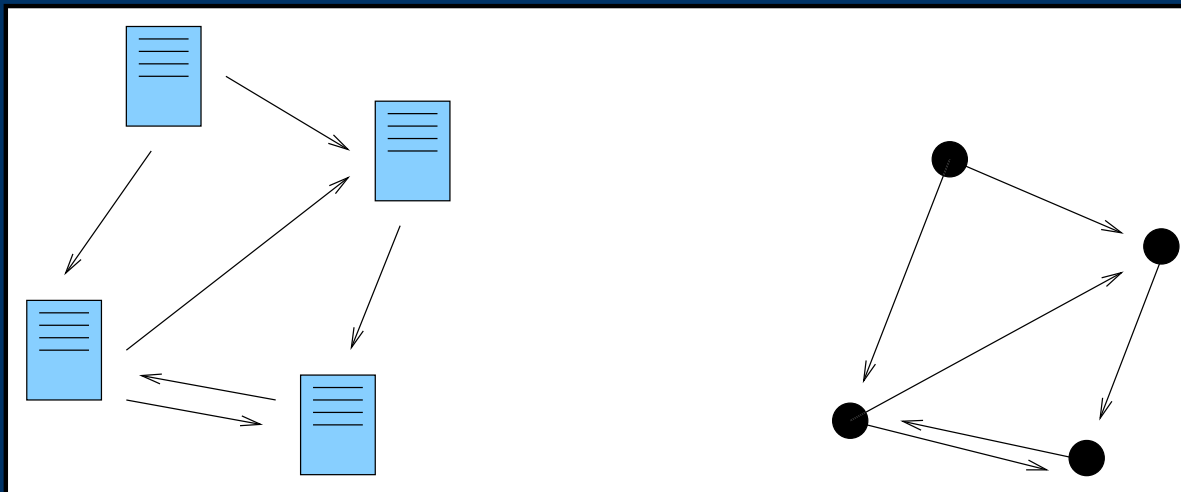


Overblik

- ✓ Indledning
- ✓ Google facts
- Internetgrafen
- En søgemaskines dele
 - Crawling
 - Indeksering
 - Søgning og ranking
- Afslutning

Internetgrafen

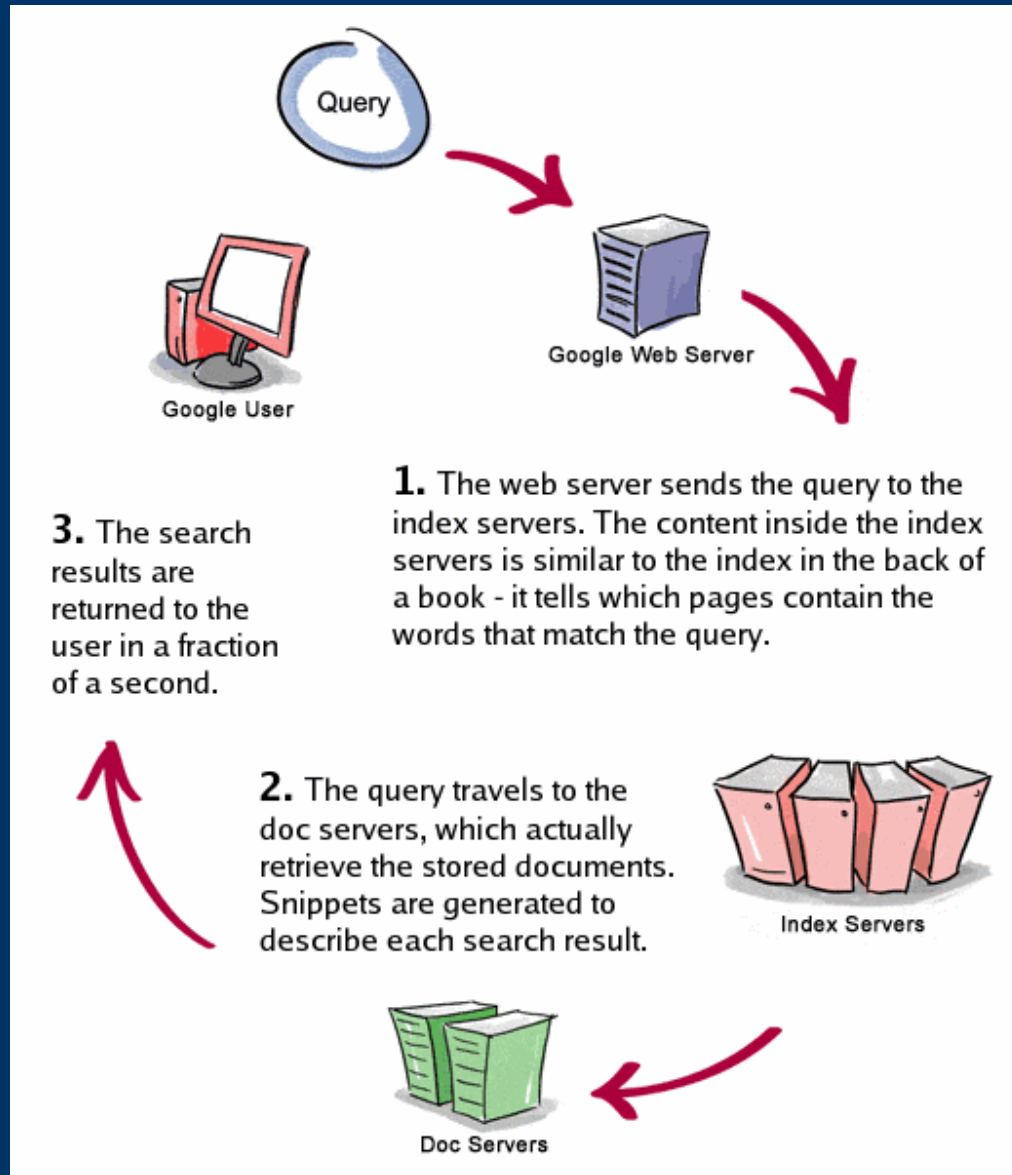
knuder = sider (URL'er)
orienterede kanter = links



Overblik

- ✓ Indledning
- ✓ Google facts
- ✓ Internetgrafen
- En søgemaskines dele
 - Crawling
 - Indeksering
 - Søgning og ranking
- Afslutning

Life of a Google Query



[Fra: <http://www.google.com/corporate/tech.html>]

En søgemaskines dele

Indsamling af data:

- Webcrawling (gennemløb af internetgrafem).

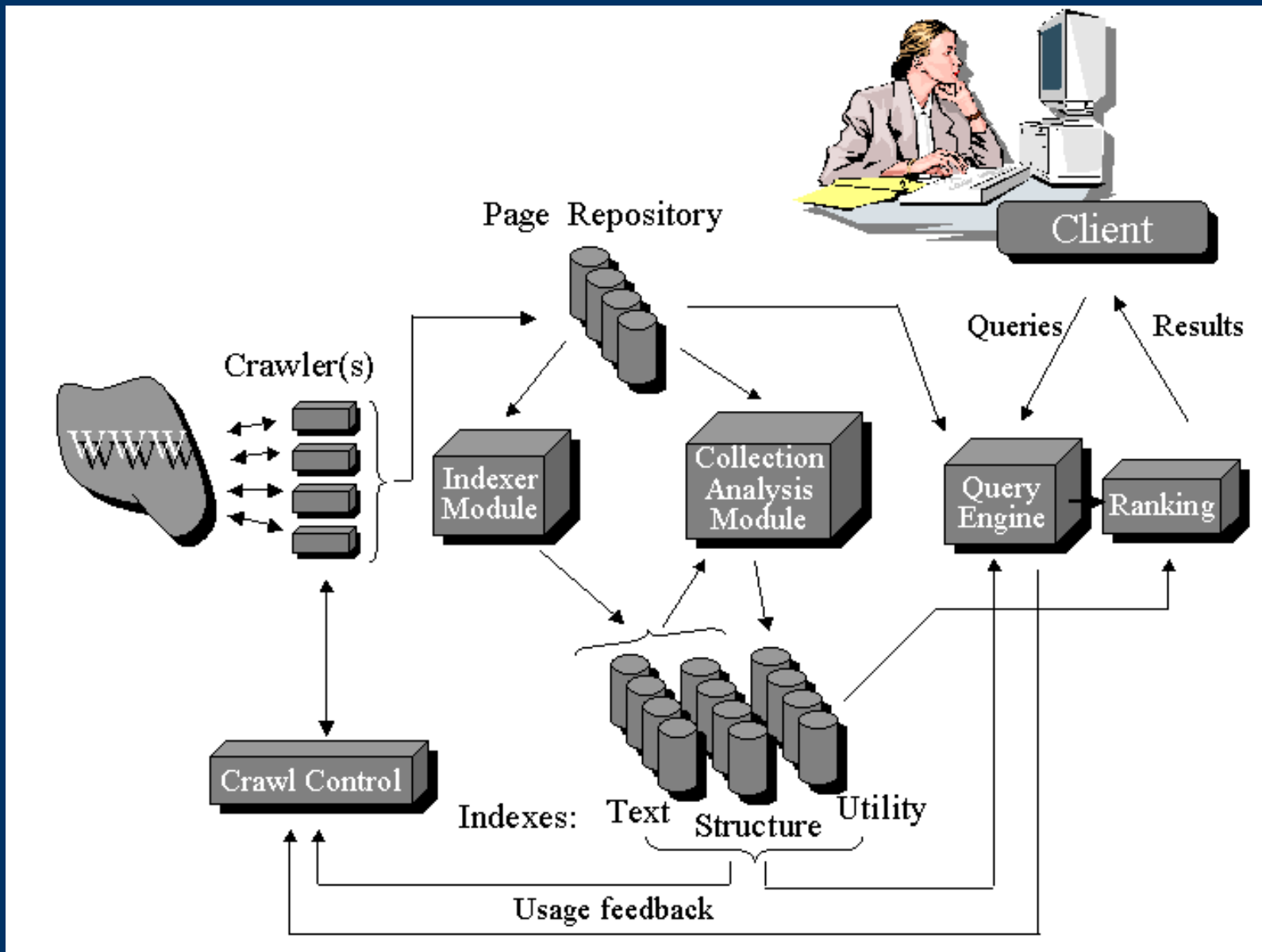
Indeksning data:

- Parsning af dokumenter.
- Lexicon: indeks (ordbog) over alle ord mødt.
- Inverted file: for alle ord i lexicon, angiv i hvilke dokumenter de findes.

Søgning i data:

- Find alle dokumenter med søgeordene.
- Rank dokumenterne.

Opbygning af en Søgemaskine



Overblik

- ✓ Indledning
- ✓ Google facts
- ✓ Internetgrafen
- ✓ En søgemaskines dele
 - Crawling
 - Indeksering
 - Søgning og ranking
- Afslutning

Crawling

Webcrawling = Grafgennemløb

$S = \{\text{startside}\}$

repeat

 fjern en side s fra S

 parse s og find alle links (s, v)

foreach (s, v)

if v ikke besøgt før

 indsæt v i S

Designovervejelser

- Startpunkt (initial S).
- Crawl-strategi (valg af s).
- Mærkning af besøgte sider.
- Robusthed.
- Ressourceforbrug (egne og andres ressourcer).
- Opdatering. Kontinuert vs. periodisk crawling.

```
 $S = \{\text{startside}\}$   
repeat  
  fjern en side  $s$  fra  $S$   
  parse  $s$  og find alle links  $(s, v)$   
  foreach  $(s, v)$   
    if  $v$  ikke besøgt før  
      indsæt  $v$  i  $S$ 
```

Output: DB med besøgte dokumenter.

DB med links i disse (kanterne i Internetgrafem)

DB med DokumentID–URL mapning

Crawl-strategier

- Breath First Search
- Depth First Search
- Random
- Priority Search

Mulige prioriteter:

- Sider som opdateres ofte (kræver metode til at estimere opdateringsfrekvens).
- Efter vigtighed (kræver metode til at estimere vigtighed, f.eks. PageRank).

BFS virker godt

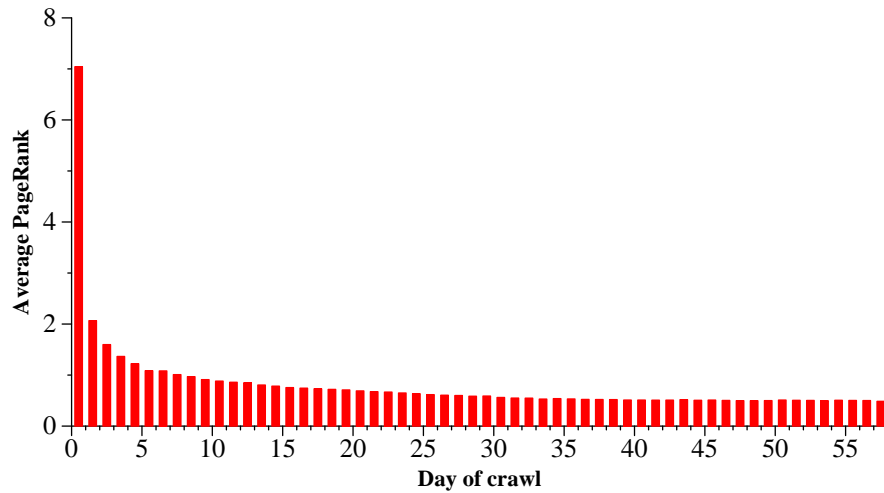


Figure 1: Average PageRank score by day of crawl

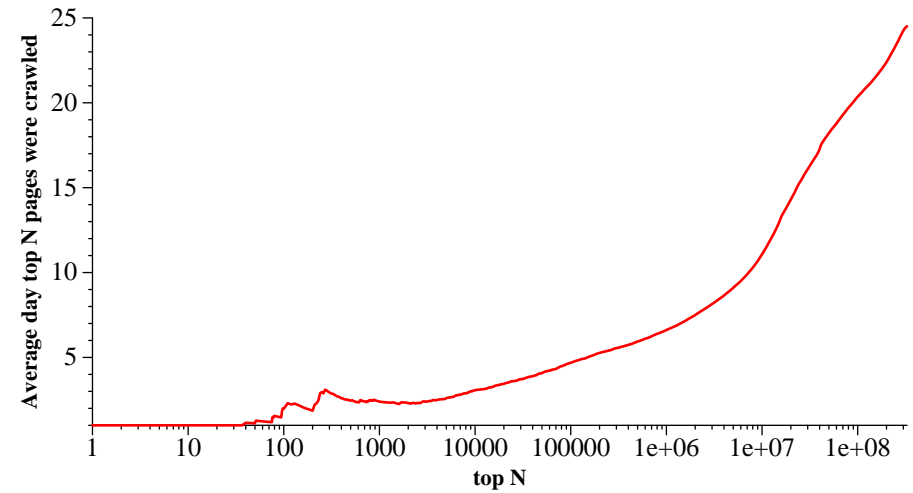


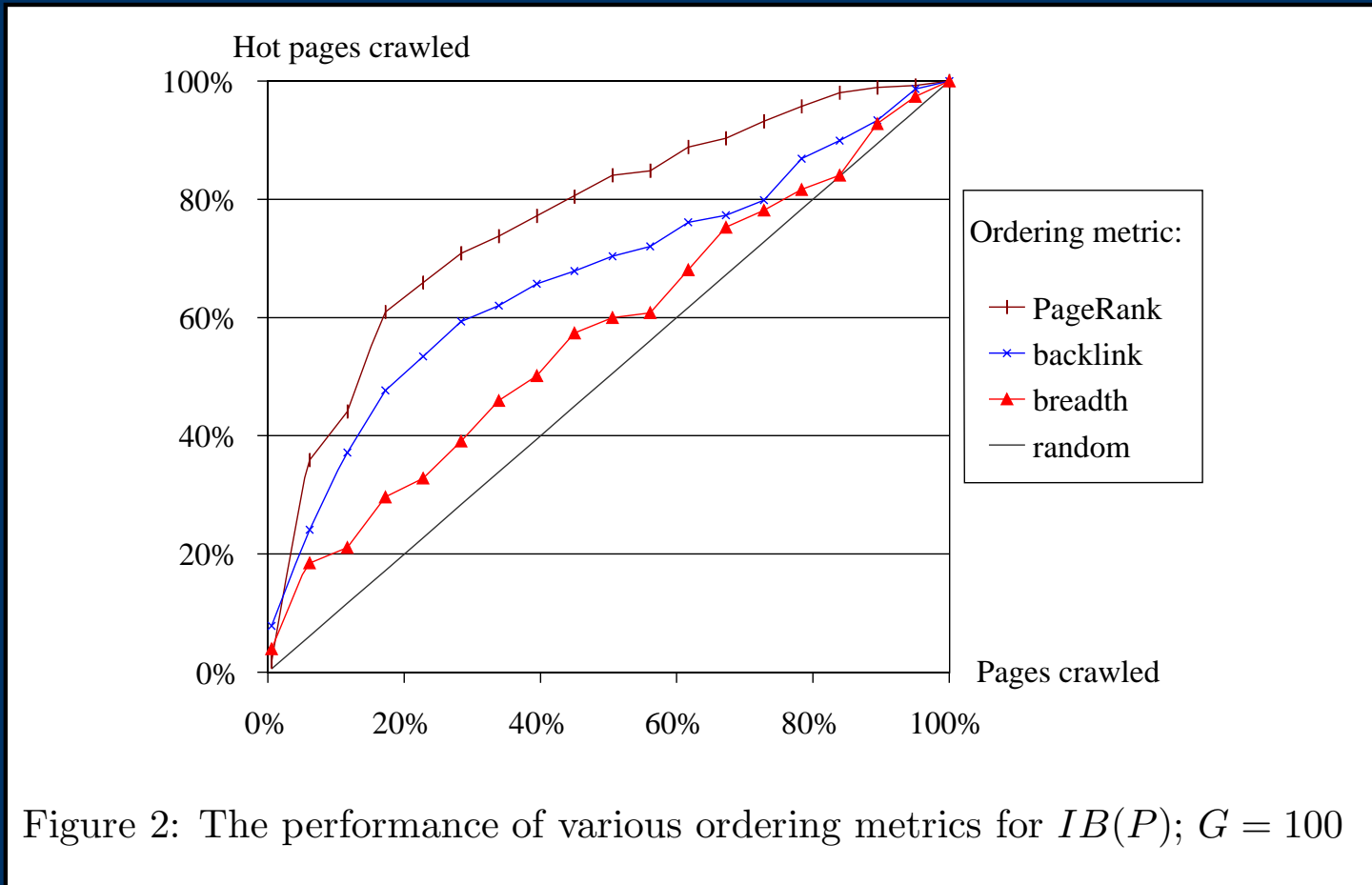
Figure 2: Average day on which the top N pages were crawled

[Fra: Najork and Wiener, 2001]

Fra et crawl af 328 millioner sider.

PageRank prioritet er endnu bedre

(men mere beregningstung...)



[Fra: Arasu et al., 2001]

Fra et crawl af 225.000 sider på Stanford University.

Robusthed

- Normalisering af URLer.
- Parsning af malformet HTML.
- Mange filtyper.
- Forkert content-type fra server.
- Forkert HTTP response code fra server.
- Enorme filer.
- Uendelige URL-løkker (crawler traps).
-

Robusthed

- Normalisering af URLer.
- Parsning af malformet HTML.
- Mange filtyper.
- Forkert content-type fra server.
- Forkert HTTP response code fra server.
- Enorme filer.
- Uendelige URL-løkker (crawler traps).
- ⋮

Vær konservativ – opgiv at finde alt.
Crawling tager måneder – brug checkpoints.

Ressourceforbrug

Egne ressourcer

- Båndbredde (global request rate)
- Lagerplads (brug kompakte representationer)
- Distribuér på flere maskiner (opdel f.eks. rummet af ULR'er)

Ressourceforbrug

Egne ressourcer

- Båndbredde (global request rate)
- Lagerplads (brug kompakte representationer)
- Distribuér på flere maskiner (opdel f.eks. rummet af ULR'er)

Andres ressourcer (politeness)

- **Båndbredde** (lokal request rate). Tommelfingerregel: 30 sekunder mellem request til samme site.
- Robots Exclusion Protocol (www.robotstxt.org).
- Giv kontakt info i HTTP-request.

Erfaringer ang. effektivitet

- Brug caching (DNS opslag, robots.txt files, senest mødte URL'er).
- Flaskehals er ofte I/O under tilgang til datastrukturerne
- CPU cykler er ikke flaskehals (Java og scripting languages er OK).
- En tunet crawler (på een eller få maskiner) kan crawle

200-400 sider/sek ~ 35 mio sider/dag.

Statistik

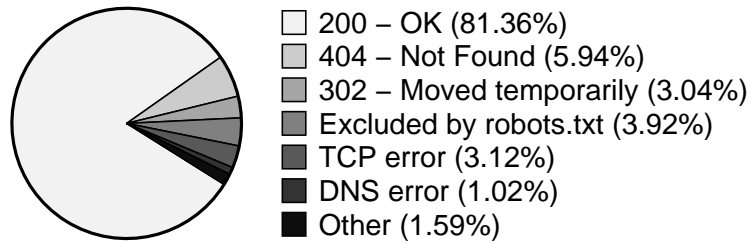


Figure 6: Outcome of download attempts

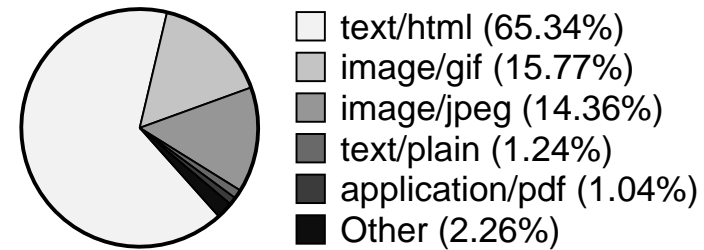


Figure 7: Distribution of content types

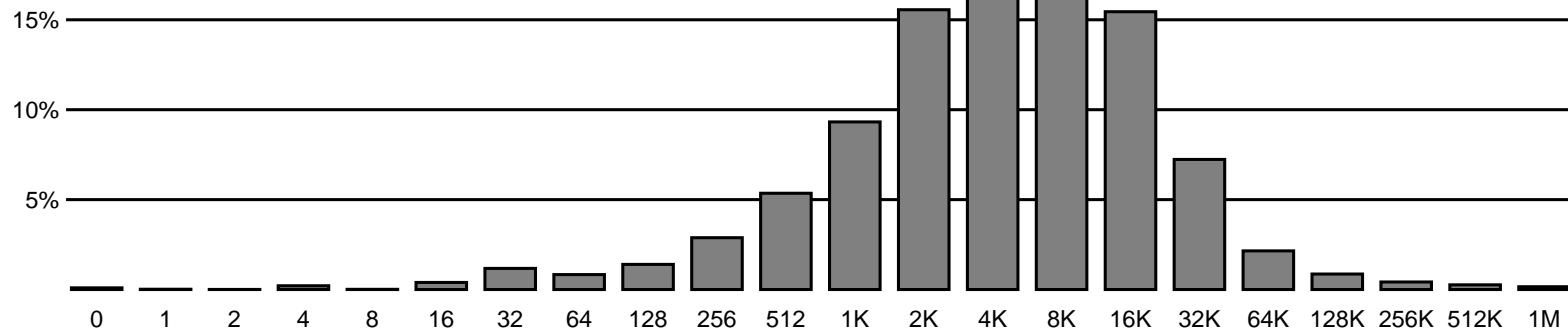


Figure 8: Distribution of document sizes

[Fra: Najork and Heydon, 2001]

Overblik

- ✓ Indledning
- ✓ Google facts
- ✓ Internetgrafen
- ✓ En søgemaskines dele
 - ✓ Crawling
 - Indeksering
 - Søgning og ranking
 - Afslutning

Indeksering af dokumenter

Opgave:

Preprocessér en dokumentssamling så dokumenter med et givet søgeord kan blive returneret hurtigt.

Input: dokumentssamling.

Output: søgestruktur.

Indeksering: Inverted file + lexicon

- Inverted file = for hvert ord w en liste af dokumenter indeholdende w .
- Lexicon = ordbog over alle ord i dokumentsamlingen.
(key = ord, value = pointer til liste i inverted file + evt. ekstra info for ordet, f.eks. længde af listen)

For en milliard dokumenter:

Inverted files \sim totalt antal ord ≥ 100 mia

Disk

Lexicon \sim antal forskellige ord ~ 2 mio

RAM

Lexicon

Kan være i RAM, så almindelige ordbogs-datastrukturer er OK.
F.eks.:

- Binær søgning i sorteret liste af ord.
- Hash tabeller.
- Tries, suffix træer, suffix arrays.

Inverted File

- Simpel (forekomst af ord i dokument):

```
ord1: DocID, DocID, DocID  
ord2: DocID, DocID  
ord3: DocID, DocID, DocID, DocID, DocID, . . .  
⋮
```

- Detaljeret (*alle* forekomster af ord i dokument):

```
ord1: DocID, Position, Position, DocID, Position. . .  
⋮
```

- Endnu mere detaljeret:

Forekomst annoteret med info (heading, boldface, anchor text, . . .). Kan bruges under ranking.

Komprimer inverted file

- Specifikke metoder
 - Gem differencen mellem DocID'er (ikke absolutte DocID'er).
 - Kod denne difference effektivt.
- Generiske værktøjer (zip,...)
 - Komprimer hver liste.
 - Opdel lister i blokke, komprimer hver blok.

Parsning af dokumenter

- Find ord
 - Fjern mark-up, scripts, . . .
 - Definition af ord? (sekvens af alfanumeriske tegn, længde max 256, max 4 digits).
 - Lowercase
 - Tegnsæt? ascii, latin-1, Unicode,
- Stemming? (“funktion”, “funktionalitet”, . . . → “funktio”).
- Stop ord? (udelad hyppige ord som “og”, “er”, . . .).

Bygning af index

```
foreach dokument  $D$  i samlingen  
  Parse  $D$  og identificér ord  
  foreach ord  $w$   
    Udskriv (DocID,  $w$ )  
    if  $w$  ikke i lexicon  
      indsæt  $w$  i lexicon
```



$(1, 2), (1, 37), \dots, (1, 123)$, $(2, 34), (2, 37), \dots, (2, 101)$, $(3, 486), \dots$

Disk sorting ↓

$(22, 1), (77, 1), \dots, (198, 1)$, $(1, 2), (22, 2), \dots, (345, 2)$, $(67, 3), \dots$

≈ inverted file

Overblik

- ✓ Indledning
- ✓ Google facts
- ✓ Internetgrafen
- ✓ En søgemaskines dele
 - ✓ Crawling
 - ✓ Indeksering
 - Søgning og ranking
- Afslutning

Søgning og Ranking

Query: computer AND science:

1. Slå computer og science op i lexicon. Giver adresse på disk hvor deres lister starter.
2. Scan disse lister og “flet” dem (returnér DocID'er som er med i begge lister).

```
computer: 12, 15, 117, 155, 256, . . .  
science: 5, 27, 117, 119, 256, . . .
```

3. Udregn rank af fundne DocID'er. Hent de 10 højst rank'ede i dokumentsamling og returnér URL samt kontekst fra dokument til bruger.

OR og NOT kan laves tilsvarende. Hvis lister har ord-positioner kan frase-søgninger (“computer science”) og proximity-søgninger (“computer” tæt på “science”) også laves.

Tekstbaseret ranking

Vægt forekomsten af et ord med f.eks.

- Antal forekomster i dokumentet.
- Ordets typografi (fed skrift, overskrift, . . .)
- Forekomst i META-tags.
- Forekomst i tekst ved links som peger på siden

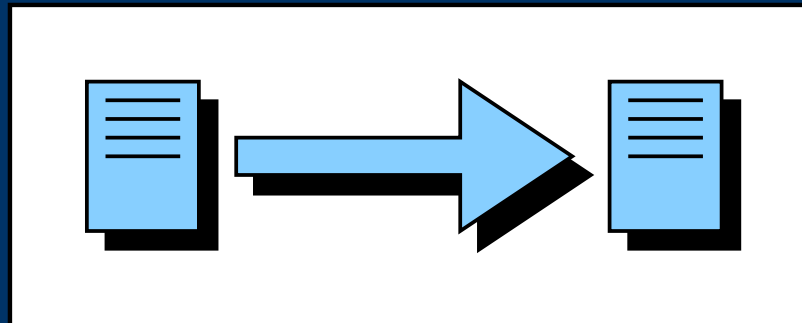
Forbedring, men ikke nok på Internettet (rankning af f.eks. 100.000 relevante dokumenter).

Let at spamme (fyld siden med søge-ord).

Linkbaseret ranking

Idé 1: Link til en side \approx anbefaling af den.

Idé 2: Anbefalinger fra vigtige sider skal vægte mere.



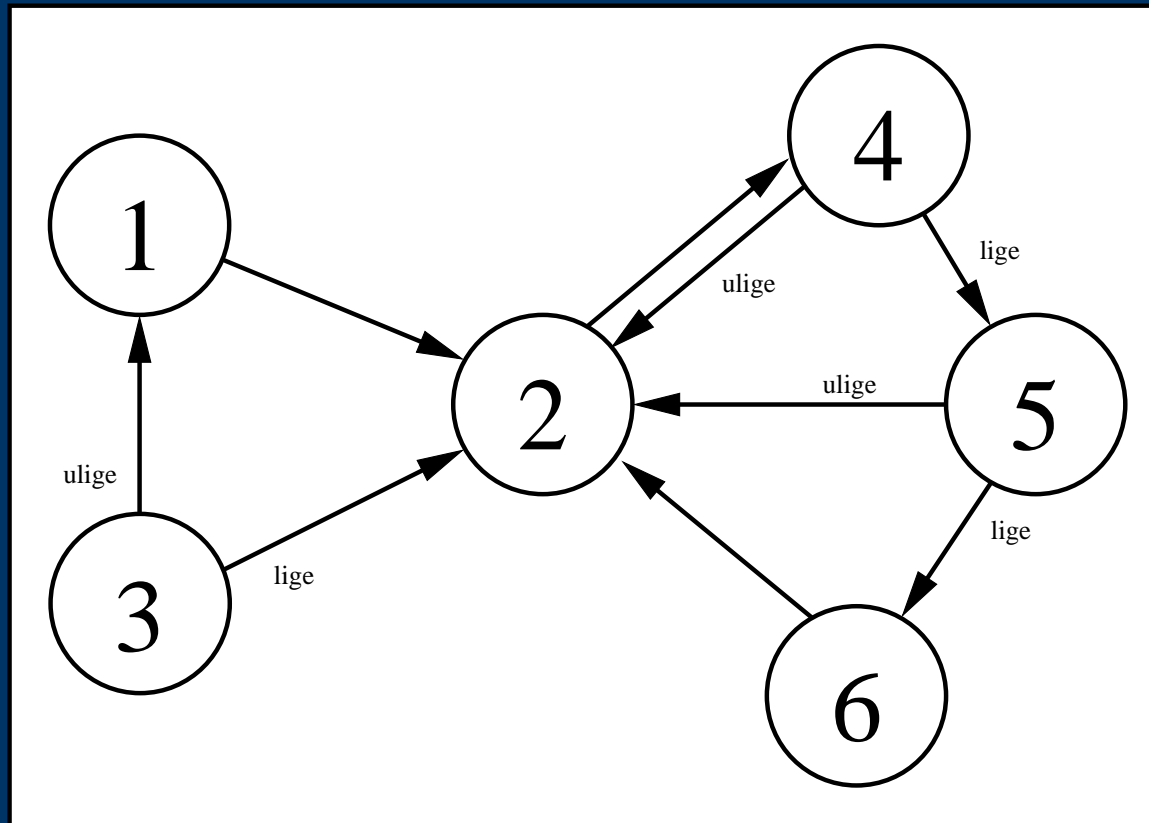
Google PageRank™ \approx websurfer

PageRank beregning kan opfattes som en websurfer som (i uendelig lang tid) i hver skridt

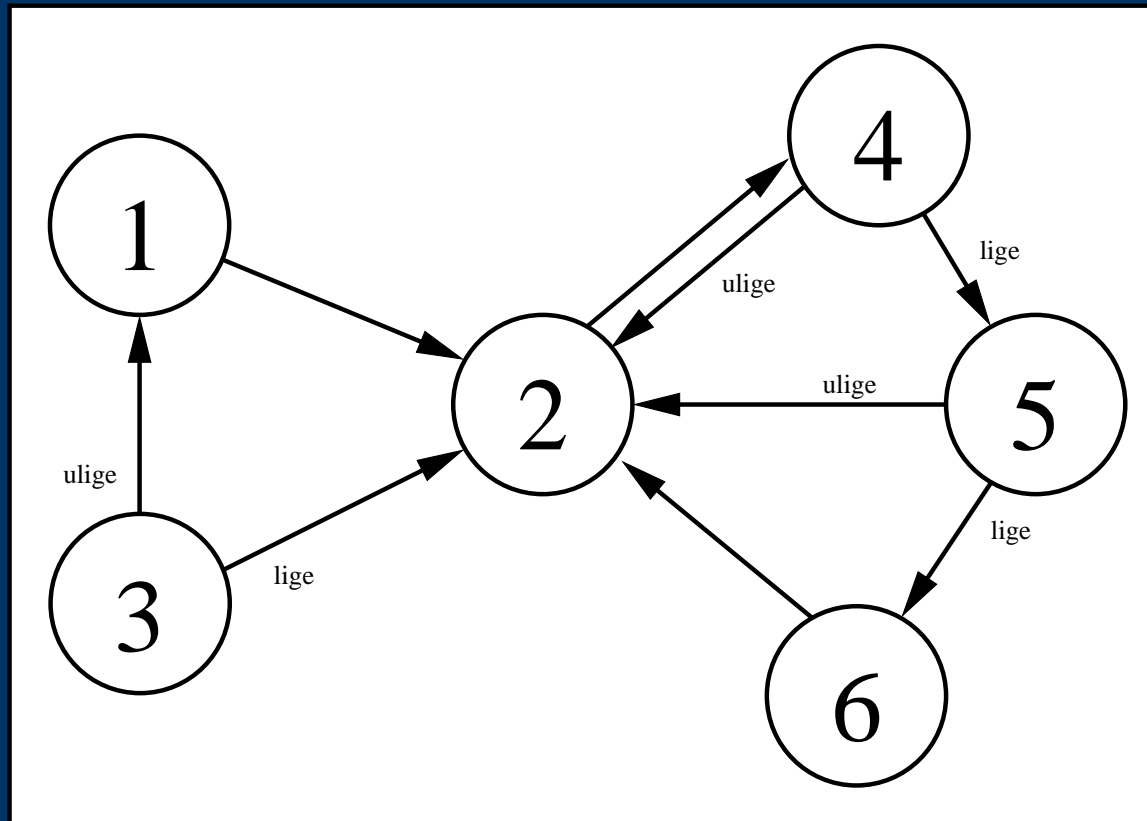
- med 85% sandsynlighed vælger at følge et tilfældigt link fra nuværende side,
- med 15% sandsynlighed vælger at gå til en tilfældig side i hele internettet.

PageRank for en side x er lig den procentdel af hans besøg som er til side x .

Simpel graf



Simpel graf



Metode RandomWalk

Start på knude 1

Gentag **25** gange:

Kast en terning:

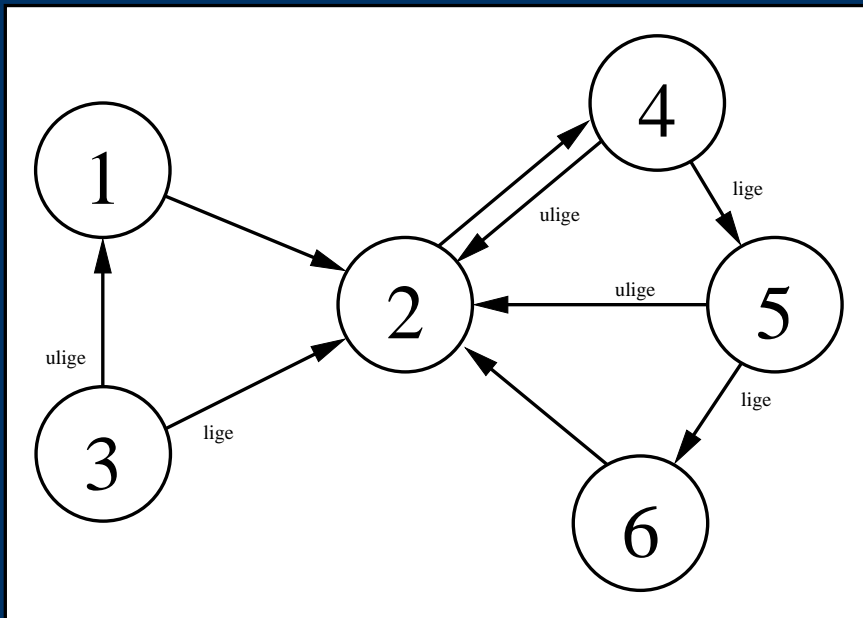
Hvis den viser 1-5:

Vælg en tilfældig pil ud fra knuden
ved at kaste en terning hvis ≥ 2 udkanter

Hvis den viser 6:

Kast terningen igen og spring hen til den
knude som terningen viser

Simpel graf — Sandsynlighedsfordeling



Skridt	1	2	3	4	5	6
0	1.000	0.000	0.000	0.000	0.000	0.000
1	0.028	0.861	0.028	0.028	0.028	0.028
2	0.039	0.109	0.028	0.745	0.039	0.039
3	0.039	0.432	0.028	0.118	0.338	0.044
4	0.039	0.299	0.028	0.388	0.077	0.169
5	0.039	0.406	0.028	0.277	0.189	0.060
6	0.039	0.316	0.028	0.366	0.143	0.107
7	0.039	0.373	0.028	0.291	0.180	0.087
8	0.039	0.342	0.028	0.339	0.149	0.103
9	0.039	0.361	0.028	0.313	0.169	0.090
10	0.039	0.348	0.028	0.329	0.158	0.098
11	0.039	0.357	0.028	0.318	0.165	0.094
12	0.039	0.351	0.028	0.325	0.160	0.096
13	0.039	0.355	0.028	0.320	0.163	0.094
14	0.039	0.352	0.028	0.323	0.161	0.096
15	0.039	0.354	0.028	0.321	0.163	0.095
16	0.039	0.353	0.028	0.323	0.162	0.095
17	0.039	0.354	0.028	0.322	0.162	0.095
18	0.039	0.353	0.028	0.322	0.162	0.095
19	0.039	0.353	0.028	0.322	0.162	0.095
20	0.039	0.353	0.028	0.322	0.162	0.095
21	0.039	0.353	0.028	0.322	0.162	0.095
22	0.039	0.353	0.028	0.322	0.162	0.095
23	0.039	0.353	0.028	0.322	0.162	0.095
24	0.039	0.353	0.028	0.322	0.162	0.095
25	0.039	0.353	0.028	0.322	0.162	0.095

Beregning af PageRank

PageRank vektoren \vec{r} er egenvektor for nabomatricen A for internetgrafen (normaliseret, d.v.s. indgangene i række i divideret med udgraden af side i)

$$\vec{r} = \vec{r}A$$

Matematisk teori (ergodisk sætning om random walks):

For vilkårlig startvektor x :

$$\vec{x}A^k \rightarrow r \quad \text{for} \quad k \rightarrow \infty$$

hvis A opfylder visse betingelser.

Beregning af PageRank

For at opfylde betingelser i PageRank: erstat A med

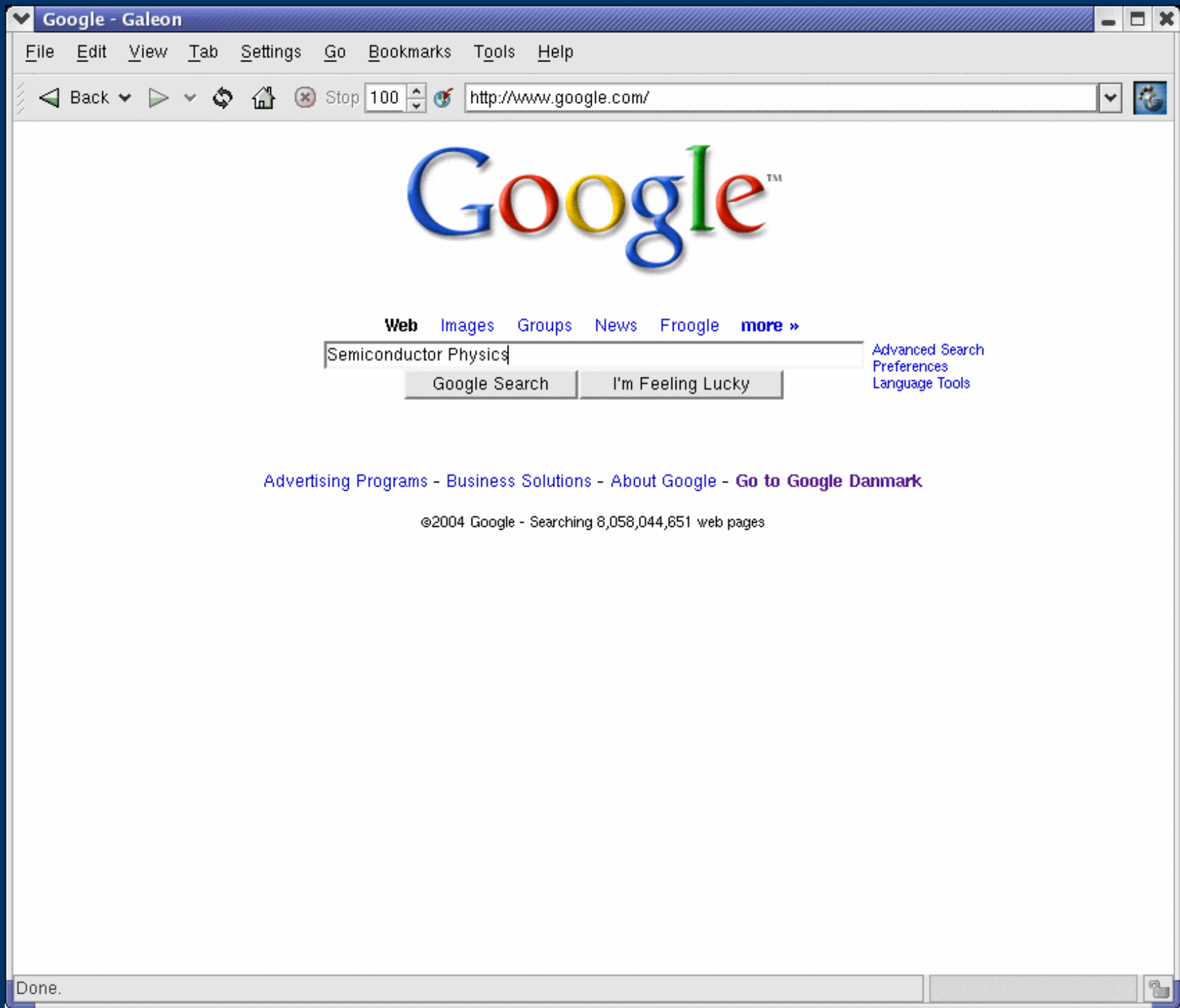
$$0.85A + 0.15E ,$$

hvor E er en (normaliseret) nabomatrice som indeholder kanter fra alle sider til alle sider. Vægtningen 85–15% er valgt ud fra at den har vist sig god i praksis.

Beregning: Gentag

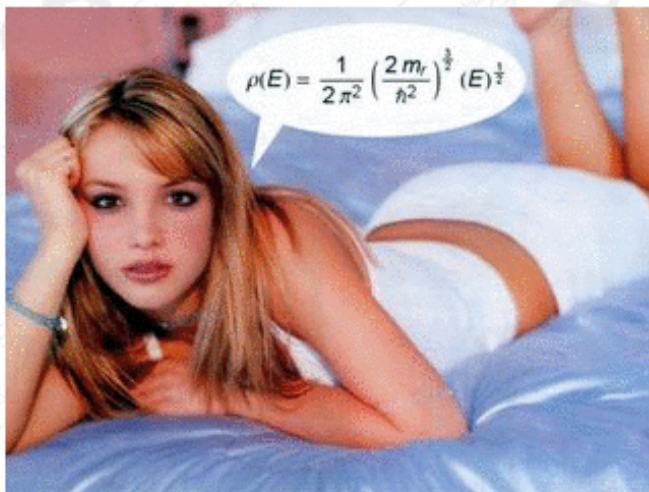
$$\vec{r}_{ny} = \vec{r}_{gl.}(0.85A + 0.15E)$$

I praksis: 20-50 iterationer er nok.



[Home] [Picture Galleries] [Britney Spears guide to Semiconductor physics]
 [Links] [Lyrics] [Guestbook] [Stuff] [Chat] [Link to us] [Awards] [Newsfeed]

Britney's Guide to Semiconductor Physics



It is a little known fact, that Ms Spears is an expert in semiconductor physics. Not content with just singing and acting, in the following pages, she will guide you in the fundamentals of the vital laser components that have made it possible to hear her super music in a digital format.

- * [Introduction](#)
- * [The Basics of Semiconductors](#)
- * [Semiconductor Crystal Structures](#)
- * [Semiconductor Junctions](#)
- * [Photonic Crystals](#)
- * [Crystal Growth, Fabrication and Processing](#)
- * [Photolithography](#)
- * [Semiconductor and](#)



Search
 Search WWW Search BritneySpears.ac



BritneySpears.ac:

[Click here](#) to donate food to the starving people of the world.



Overblik

- ✓ Indledning
- ✓ Google facts
- ✓ Internetgrafen
- ✓ En søgemaskines dele
 - ✓ Crawling
 - ✓ Indeksering
 - ✓ Søgning og ranking
- Afslutning

SAS-hoteller sortlistet efter Google-fusk

Verdens mest populære søgemaskine, Google, har boykottet SAS-koncernens nordiske hoteller og konferencecentre, efter de har brugt skjulte websider til at opnå en god placering i søgeresultaterne. Metoden er udviklet af danske Netpointers, som risikerer en bombe under sit forretningsgrundlag.

[Fra: www.computerworld.dk, 9. november 2004]

Gør-det-selv

Programmeringsprojekt i kurset *Algorithms for Web Indexing and Searching* (Gerth S. Brodal, Rolf Fagerberg), efteråret 2002.

- Opgave: lav en søgemaskine for domæne `.dk`.
- 15 studerende.
- 4 parallelt arbejdende grupper (crawling, indexing, PageRank, søgning/brugergrænseflade).
- Erfaring: Rimelig vellykket søgemaskine, hvor rankningen dog kræver yderligere finjustering...

References

- Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan, *Searching the Web*. ACM Transactions on Internet Technology, 1, p. 2-43, 2001.
- Sergey Brin and Larry Page, *The Anatomy of a Search Engine*, 1998.
<http://www-db.stanford.edu/pub/papers/google.pdf>
- Monika Rauch Henzinger, *Web Information Retrieval*. Proceedings of the 16th International Conference on Data Engineering, 2000.
- Marc Najork and Allan Heydon, *High-Performance Web Crawling*. Compaq SRC Research Report 173.
- Marc Najork and Janet L. Wiener, *Breadth-First Search Crawling Yields*. In Proceedings of the Tenth International World Wide Web Conference, 114-118, 2001.