



Sådan virker Google™ (måske)

Gerth Stølting Brodal

Datalogisk Institut
Aarhus Universitet

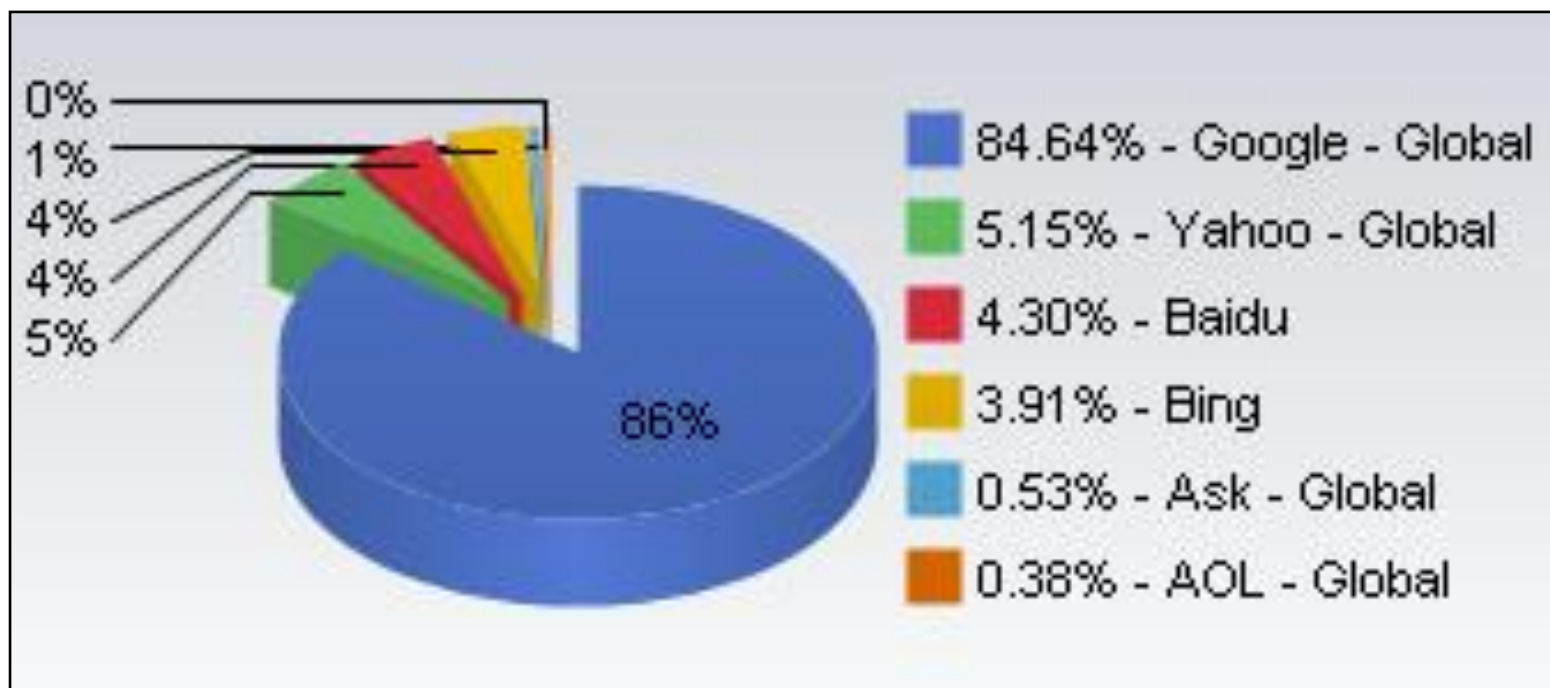
madalGO 
CENTER FOR MASSIVE DATA ALGORITHMICS

Hvornår har du sidst brugt Google?

- a) Seneste time
- b) Idag
- c) Denne uge
- d) Denne måned
- e) Aldrig

Internetsøgemaskiner

(April 2011)



marketshare.hitslink.com

Google™

Baidu 百度

msn.

HOTBOT

altavista™

Jubii
-or not to be

alltheweb
◦ • ◦ find it all ◦ • ◦

AOL Search

excite

LYCOS

YAHOO! SEARCH

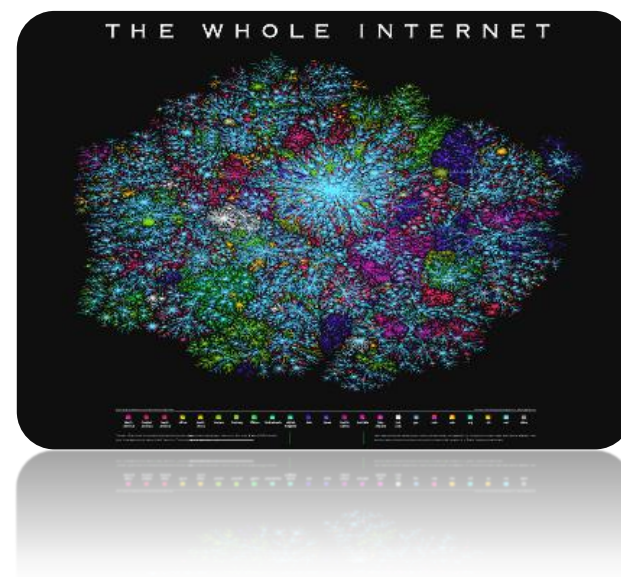
bing™

Ask™
.com

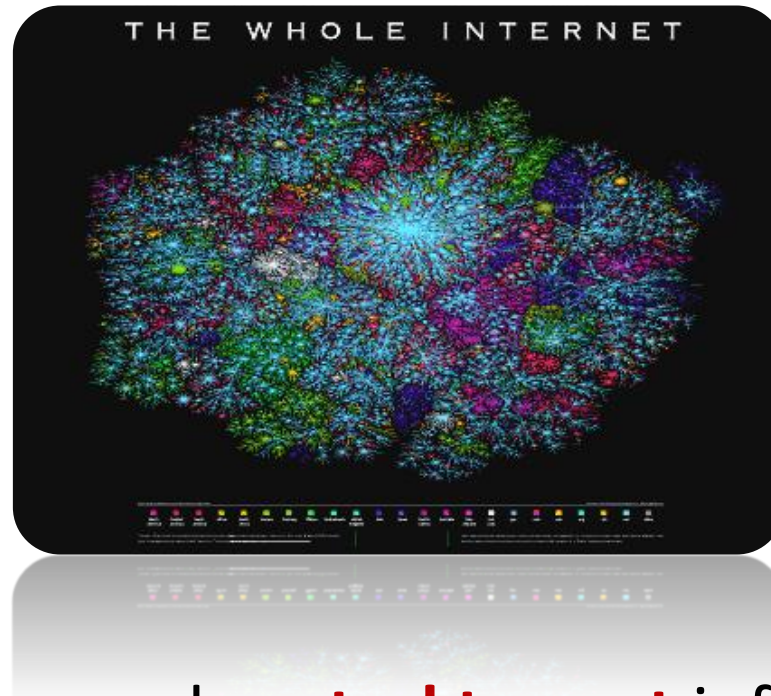
Historie

Historiske Perspektiv

- 1969 Første ARPANET forbindelser - starten på internettet
- 1971 Første email
- 1990 HTML sproget defineres
- 1991-1993 få kender til HTML
- 1993 Mosaic web-browser
- 1993 Lycos
- 1994 WebCrawler
- 1995 Yahoo!, Altavista
- **1998 Google**
- 2004 Facebook
- 2005 YouTube
-



Internettet — **W**orld**W**ide**W**eb



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

1994 Lycos,. . .

1996 Alta Vista: mange sider

1999 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 &

The Professional Cartoonists Index is the most c
cartoonists on

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 Spicelslander TalkShop. [Follow Ups] [Pos
The Spicelslander 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/blablabla/forum/messages/1/15]
Re: Princess Diana's gown auction. [Follow Ups
Followup] [Elle International - Blablabla] 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
November
Last modifi

4. Re: Princess Diana - Queen of Hearts

[URL: www.elle.com/textes/blablabla/forum/messages/1/28]
Re: Princess Diana - Queen of Hearts. [Follow U
Followup] [Elle International - Blablabla] 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 aniston 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 FIRST FREE PICTURES WOMEN
99% - Articles & General info: SEX CHAT XX
PORNO PLAYBOY PAMELA ANDERSON PE
PICTURES WOMEN ADULT MUSIC CHAT B
EROTICA JERRY MCCRARTHY LEONOR SA
CHAT CRAWFORD MISS GILLS... 03/09/98

Personal page: <http://www.connix.com/~wgonzo/sex/slidesuperall.htm>

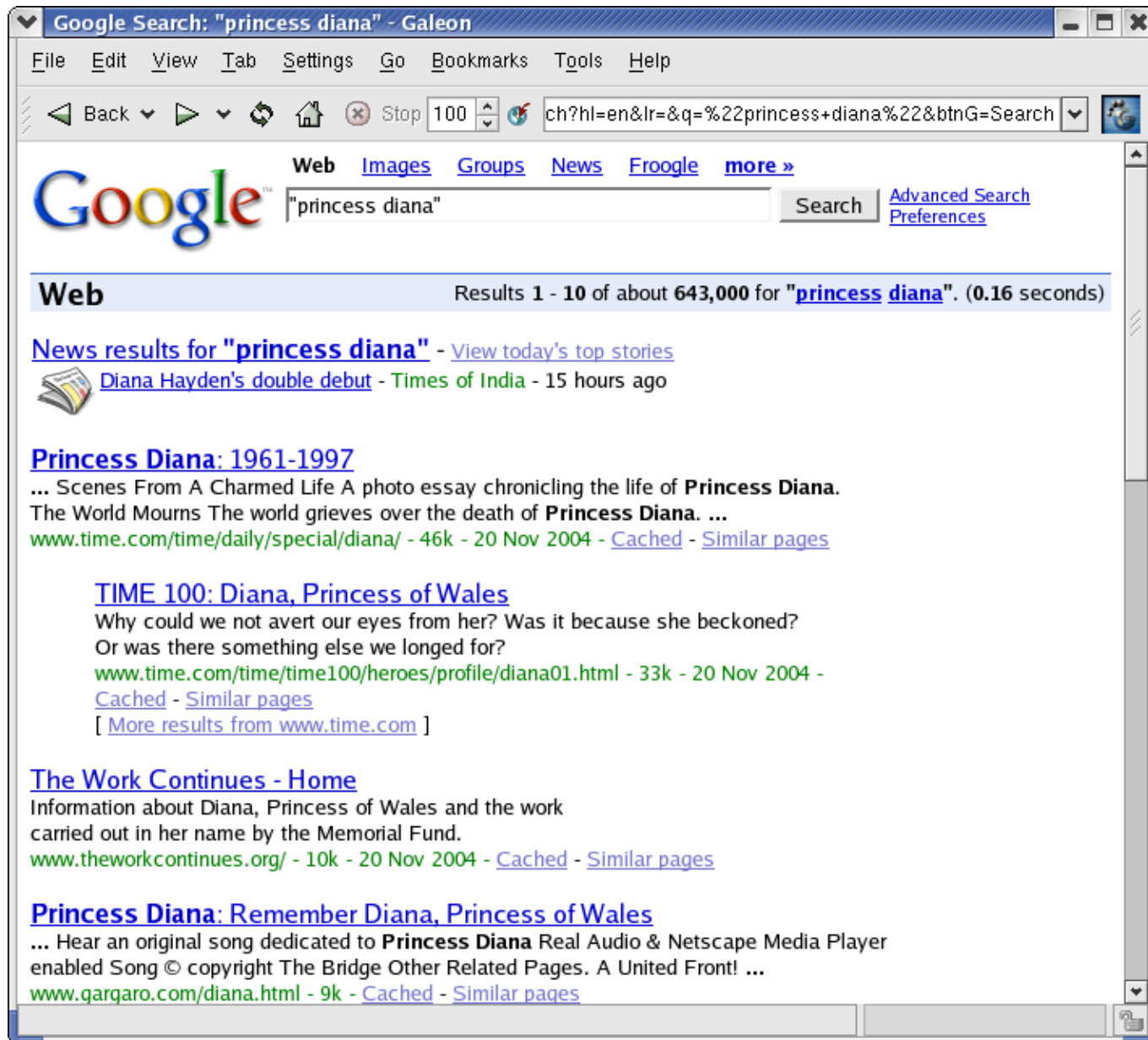
3. Ro Not relevant
index pollution

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

2004



2009

"princess diana" - Google Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

"princess diana" - Google Search

Web Images Maps Groups Scholar Blogs Gmail more ▾ Search settings | Sign in

Google "princess diana" Search Advanced Search

Web [Show options...](#) Results 1 - 10 of about 2,350,000 for "princess diana". (0.11 seconds)

[Diana, Princess of Wales - Wikipedia, the free encyclopedia](#)
Posthumously, as in life, she is most popularly referred to as "**Princess Diana**", a title she never held. Still, she is sometimes referred to (according to ...
[Early life](#) - [Royal descent](#) - [Education](#) - [Marriage](#)
[en.wikipedia.org/wiki/Diana_Princess_of_Wales](#) - [Cached](#) - [Similar](#)

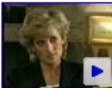
[Death of Diana, Princess of Wales - Wikipedia, the free encyclopedia](#)
"Prince Charles Implicated in Murder of **Princess Diana**" Logic dictates Princess Di was deliberately frightened into writing the incriminating letter before ...
[en.wikipedia.org/wiki/Death_of_Diana_Princess_of_Wales](#) - [Cached](#) - [Similar](#)

[Princess Diana, Princess of Wales: photos, pictures, facts, news](#)
Facts, photos, news, pictures about **Princess Diana**, **Princess Diana** of Wales, Lady Diana Spencer.
[www.princess-diana.com/](#) - [Cached](#) - [Similar](#)

Image results for "**princess diana**" - [Report images](#)

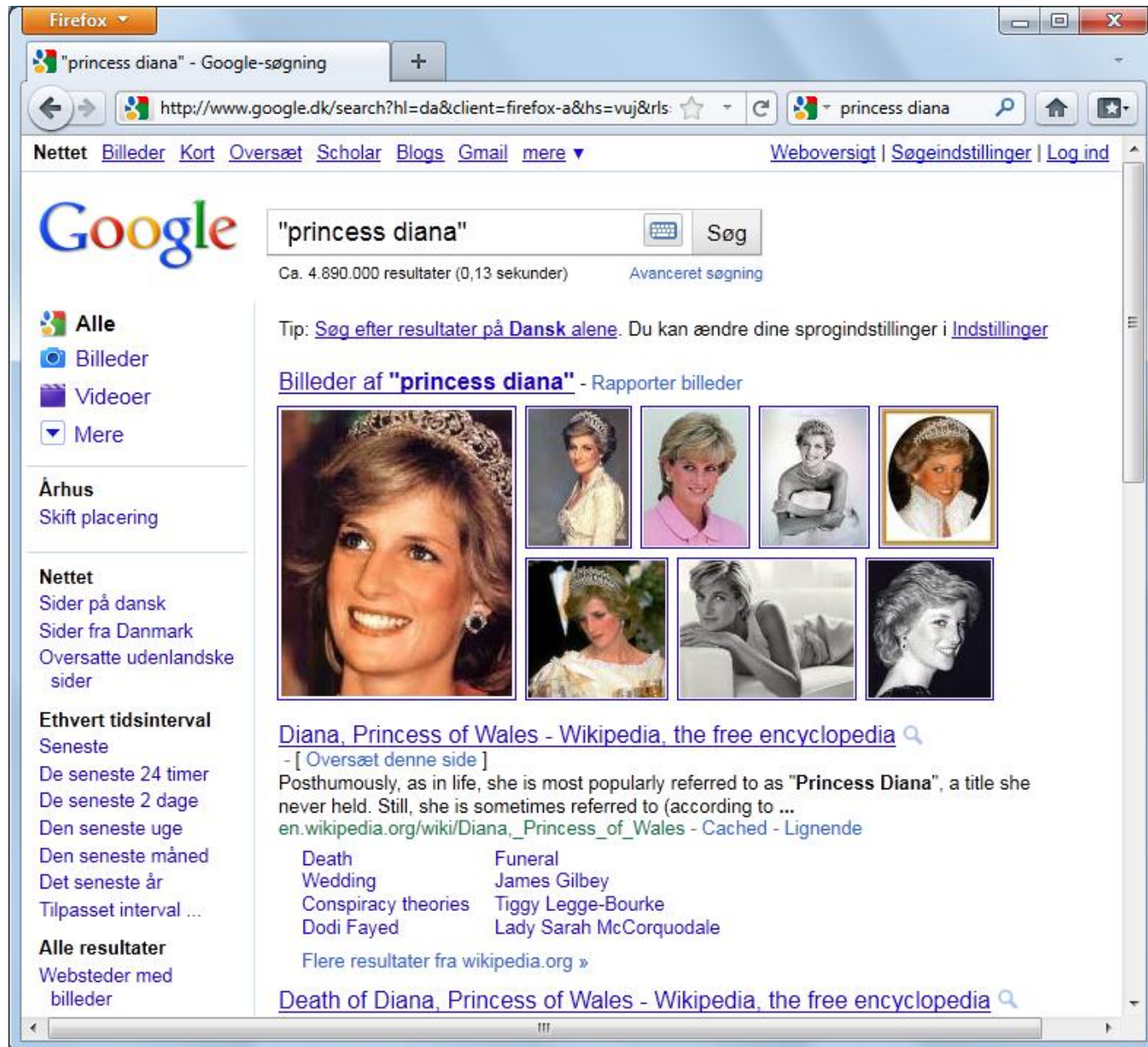


Video results for "**princess diana**"

 Candle In The Wind: A Princess Diana Tribute 4 min 13 sec www.youtube.com	 Who Killed Princess Diana? 21 min video.google.com
 PRINCESS DIANA Funeral Highlights 2 min 36 sec www.youtube.com	 Panorama interview, Princess Diana Queen Of ... 55 min video.google.com

[Princess Diana: 1961-1997](#)
1 Aug 2004 ... Photos and articles from Time Magazine covering the life and death of the Princess of Wales.
[www.time.com/time/daily/special/diana/](#) - [Cached](#) - [Similar](#)

2011





unf



Søg

Ca. 6.480.000 resultater (0,09 sekunder)

Avanceret søgning

- Alle
- Billeder
- Videoer
- Mere

Aarhus
Skift placering

Nettet
Sider på dansk
Sider fra Danmark
Oversatte udenlandske sider

Ethvert tidsinterval
Seneste
De seneste 24 timer
De seneste 2 dage
Den seneste uge
Den seneste måned
Det seneste år
Tilpasset interval ...

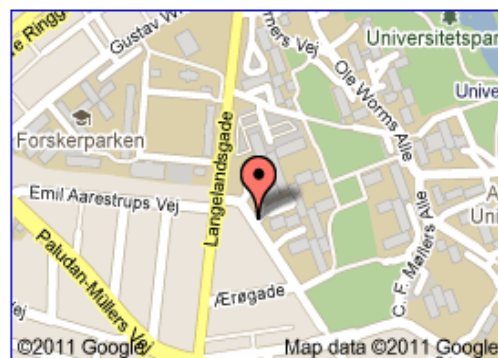
Flere værktøjer

UNF Danmark

Tilbyder foredrag, studiebesøg og nyheder indenfor alle grene af naturvidenskaben.
www.unf.dk/ - Cached - Lignende

København Om UNF
 Odense Sciencecamps
 Århus Medlemskab
 Aalborg Matematik

Flere resultater fra unf.dk »



UNF Århus

Ny Munkegade
 8000 Aarhus Municipality
 8942 3345

Bus: Langelandsgade/Kaserneboulevau
 Hent anvisninger

1 anmeldelse

UNF København

Hvad skal UNF KBH holde foredrag om d. 3. marts? [Læs beskrivelse] ...
kbh.unf.dk/ - Cached - Lignende

UNF Odense

Bag alle disse spørgsmål gemmer der sig en naturvidenskabelige forklaring ...
odense.unf.dk/ - Cached - Lignende

UNF Århus

Går du på gymnasiet eller HTX er der nu to muligheder for at komme billigt ...
aarhus.unf.dk/ - Cached - Lignende

Vis flere resultater fra unf.dk



UNF med til rumfærgeren Discoverys sidste opsendelse.
 Af René Tronsgaard Rasmussen, ...

Vil du med UNF til NASA? Af Christian Fredborg
 Brædstrup, 25. nov 2010 ...

Firefox

IP Address Geolocation to Identify Webs...

http://www.ip2location.com/

Google

Live Demo Using IP2Location™ - April 2011

IP Address	: 93.166.243.202
Location	:  DENMARK, ARHUS, ARHUS
Latitude / Longitude	: 56.158135 LATITUDE, 10.212002 LONGITUDE
Connecting through	: TDC BB-ADSL USERS
Time Zone	: UTC +01:00
Net Speed	: DSL
IDD Code	: 45
Weather Station	: DAXX0003 - ARHUS

Moderne Søgemaskiner

Imponerende performance

- Søger i 10^{10} sider
- Svartider 0,1 sekund
- 1000 brugere i sekundet
- Finder **relevante** sider

I'm Feeling Lucky

Nye Krav til Søgemaskiner



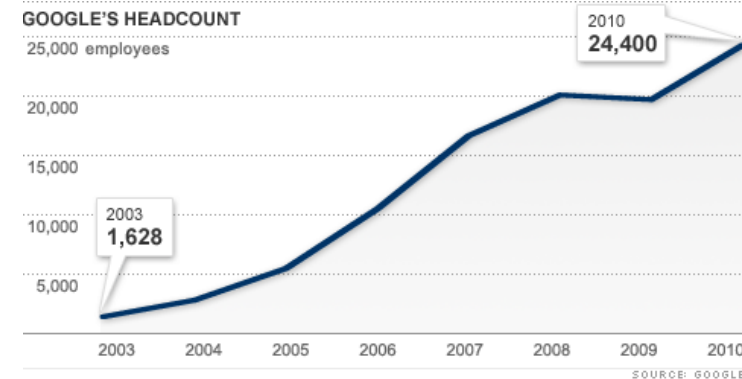
- Dynamiske websider:
Nyheder, Twitter, Facebook, ...

- Personlig ranking:
Baseret på hidtige besøgte websider,
gmail, social netværk, ...





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



google $\approx$ googol = 10^{100}

Google™



Google™ Docs



Google Chrome



Google calendar



You Tube



GMail™

by Google™



Google AdWords



Hvilken Google bygning er dette ?

- a) Arkiv over periodiske kopier af internettet
- b) Supercomputer
- c) Laboratorium for højhastighedsnetværk
- d) Distributionslager af firma løsninger



Hvilken Google bygning er dette ?

a) Arkiv over periodisk kopier af internettet



b) Supercomputer

c) Laboratorium for højhastighedsnetværk

d) Distributionslager af firma løsninger





- +8.000.000.000 web sider (+20 TB)
- PageRank: +3.000.000.000 sider og +20.000.000.000 links
- +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

Google™ (2004)



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





AU 14

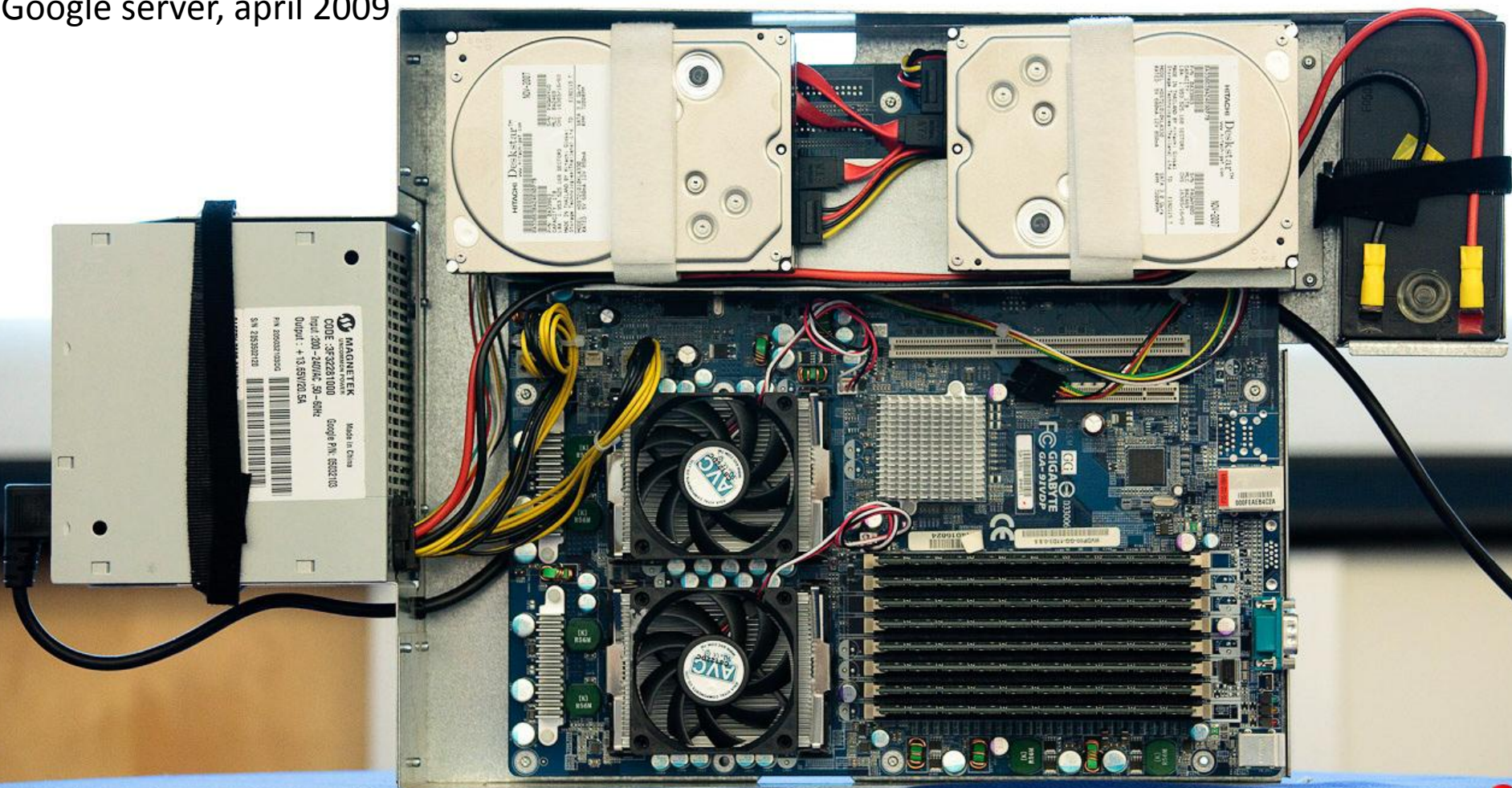
AU-14

Secure
Hewlett
Security Controls

11
WFO



Google server, april 2009



Google™ Datacentre (gæt 2007)

USA

- Mountain View, Calif.
- Pleasanton, Calif.
- San Jose, Calif.
- Los Angeles, Calif.
- Palo Alto, Calif.
- Seattle
- Portland, Oregon
- **The Dalles, Oregon**
- Chicago
- **Atlanta, Ga. (x 2)**
- **Reston, Virginia**
- Ashburn, Va.
- Virginia Beach, Virginia
- Houston, Texas
- Miami, Fla.
- **Lenoir, North Carolina**
- **Goose Creek, South Carolina**
- **Pryor, Oklahoma**
- **Council Bluffs, Iowa**

INTERNATIONAL

- Toronto, Canada
- Berlin, Germany
- Frankfurt, Germany
- Munich, Germany
- Zurich, Switzerland
- **Groningen, Netherlands**
- **Mons, Belgium**
- **Eemshaven, Netherlands**
- Paris
- London
- Dublin, Ireland
- Milan, Italy
- Moscow, Russia
- Sao Paulo, Brazil
- Tokyo
- Hong Kong
- Beijing



The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

*Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu*

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at <http://google.stanford.edu/>. To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date. Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords

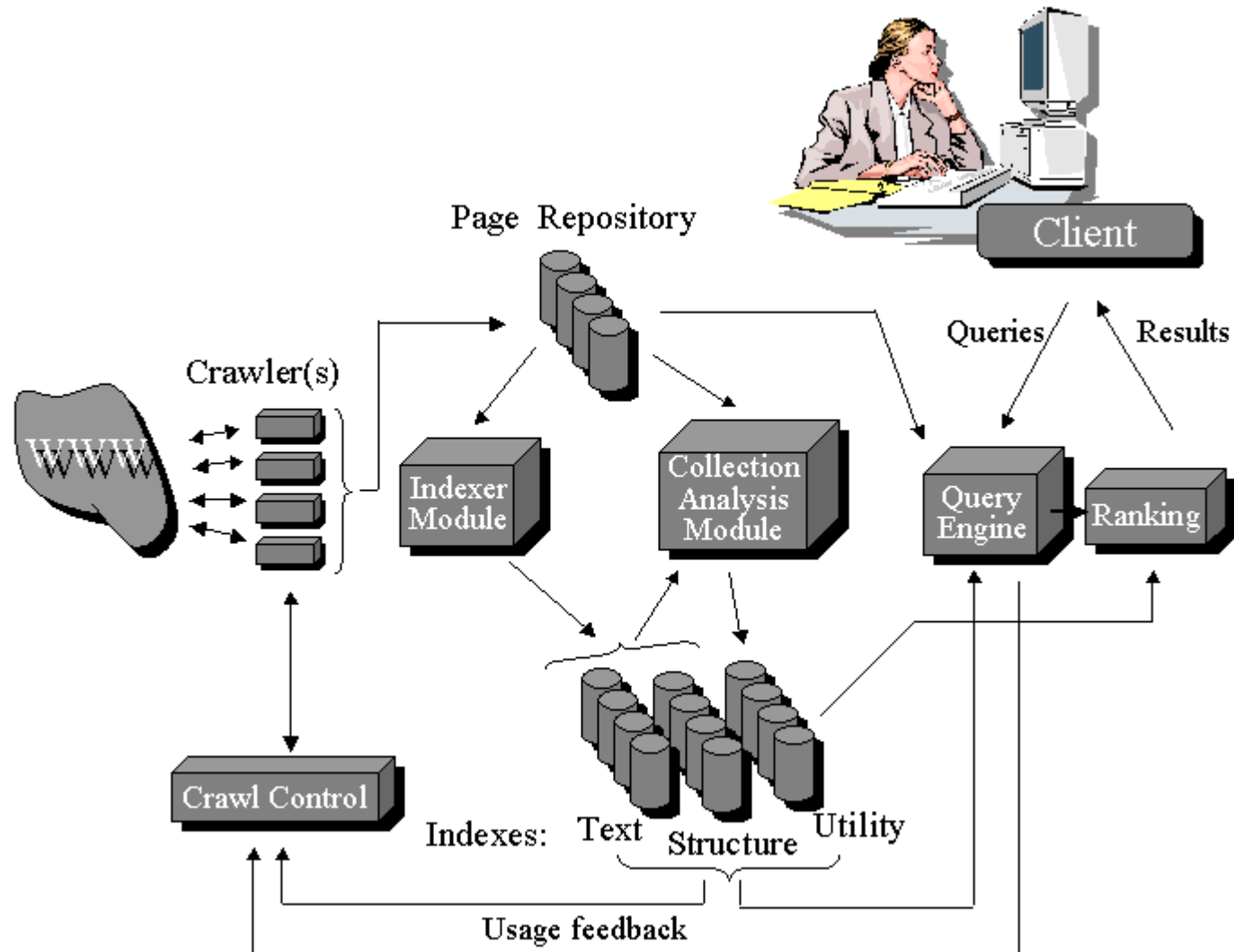
World Wide Web, Search Engines, Information Retrieval, PageRank, Google

1. Introduction

(Note: There are two versions of this paper -- a longer full version and a shorter printed version. The full version is available on the web and the conference CD-ROM.)

The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo! or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines. We have built a large-scale search engine which addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results. We chose our system name, Google, because it is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search

Opbygning af en Søgemaskine



En søgemaskines dele

Indsamling af data

- **Webcrawling** (gennemløb af internet)

Indeksering data

- **Parsning** af dokumenter
- **Leksikon**: indeks (ordbog) over alle ord mødt
- **Inverteret fil**: for alle ord i leksikon, angiv i hvilke dokumenter de findes

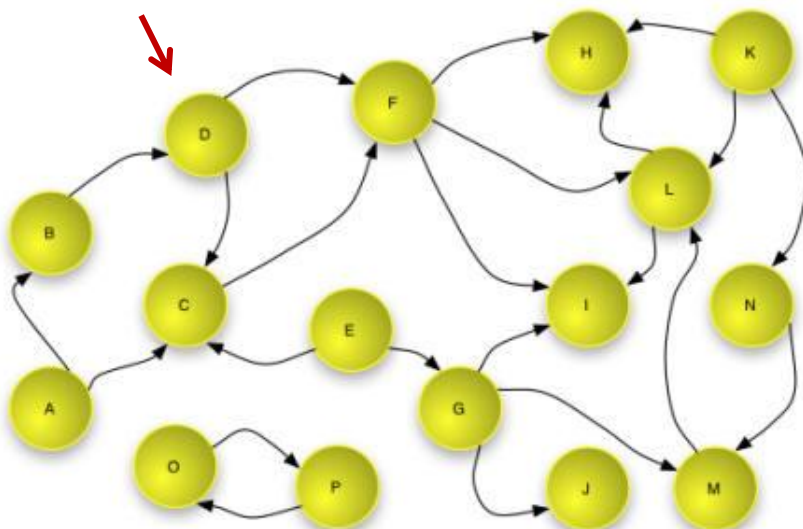
Søgning i data

- Find alle dokumenter med søgeordene
- **Rank** dokumenterne

Crawling

Webcrawling = Grafgennemløb

startside



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

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

Statistik

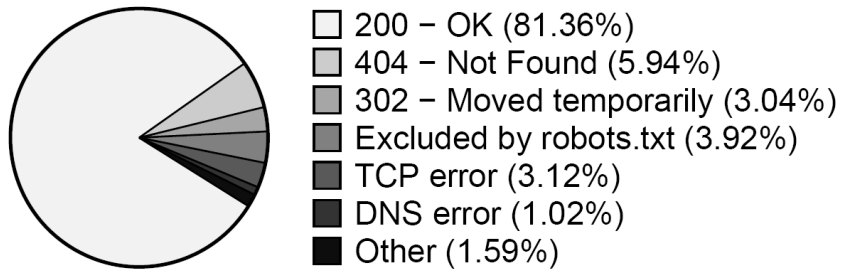


Figure 6: Outcome of download attempts

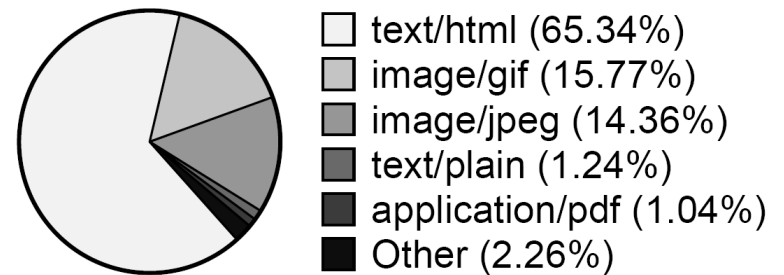


Figure 7: Distribution of content types

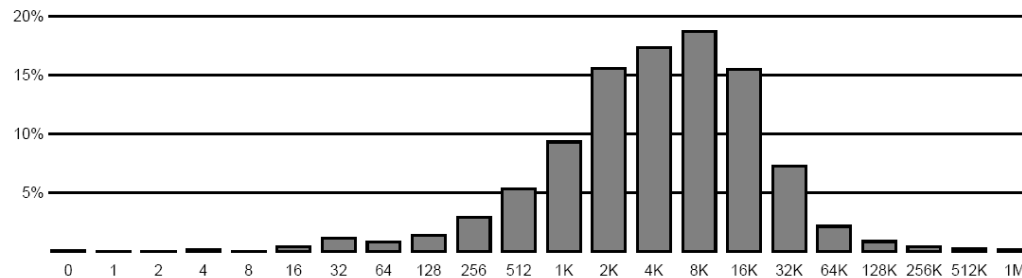
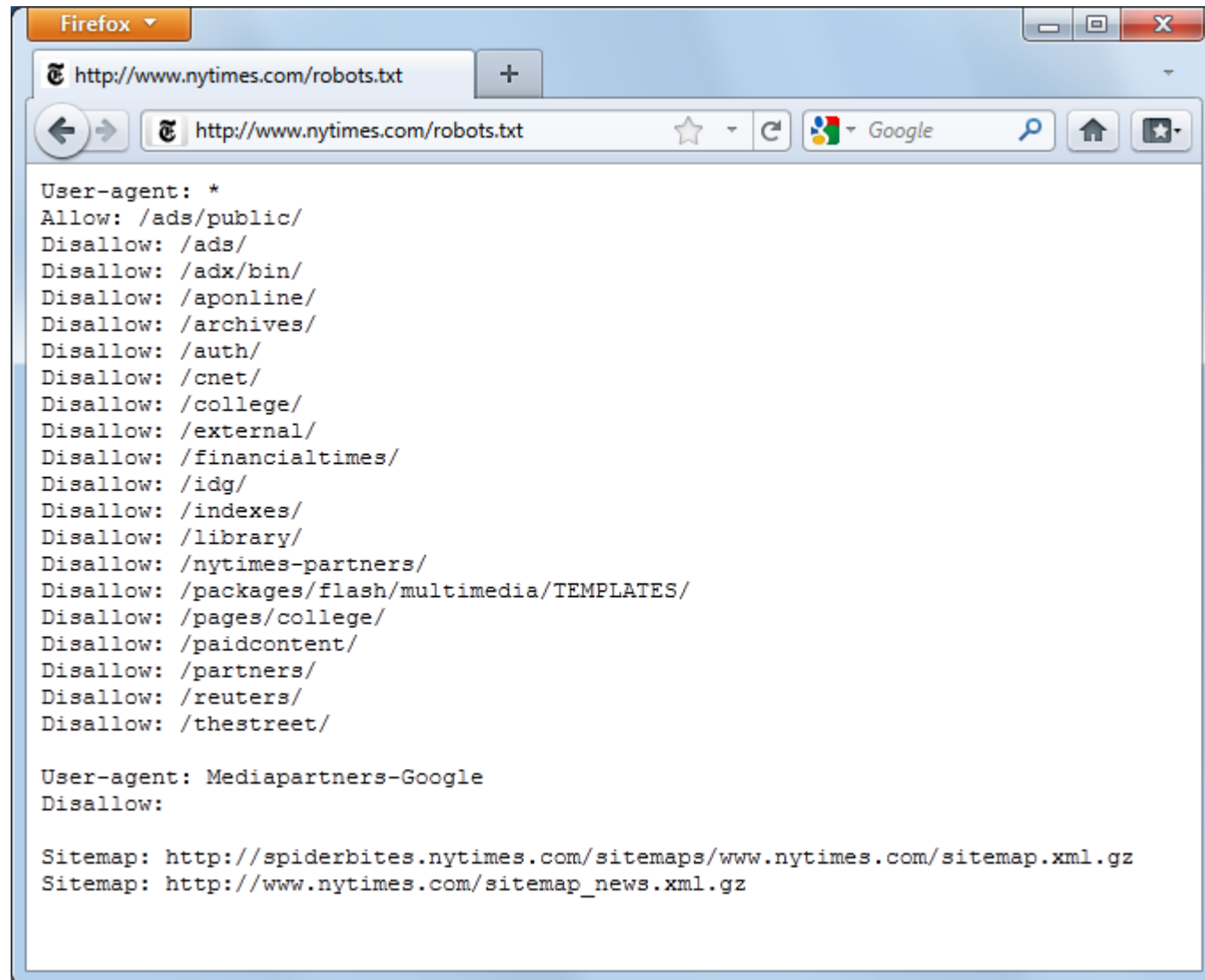


Figure 8: Distribution of document sizes

robots.txt @ The New York Times



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 kan tage måneder

Designovervejelser - Crawling

- 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 Internetgrafen)
DB med DokumentID–URL mapping

Crawling Strategi ?

- Bredde først søgning
- Dybde først søgning
- Tilfældig næste s
- Prioriteret søgning, fx
 - Sider som opdateres ofte (kræver metode til at estimere opdateringsfrekvens)
 - Efter vigtighed (kræver metode til at estimere vigtighed, fx PageRank)

```
S = {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
```

Hvilken crawling strategi er bedst?

- a) Bredde først søgning (“ringe i vandet”)
- b) Dybde først søgning (“søg i en labyrint”)
- c) Ved ikke

```
S = {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
```

Hvilken crawling strategi er bedst?



- a) Bredde først søgning (“ringe i vandet”)
- b) Dybde først søgning (“søg i en labyrint”)
- c) Ved ikke

```
S = {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
```

Crawling : BFS virker godt

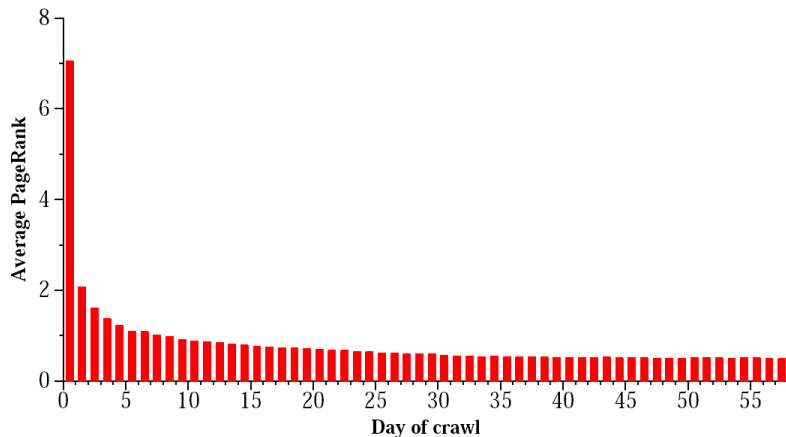


Figure 1: Average PageRank score by day of crawl

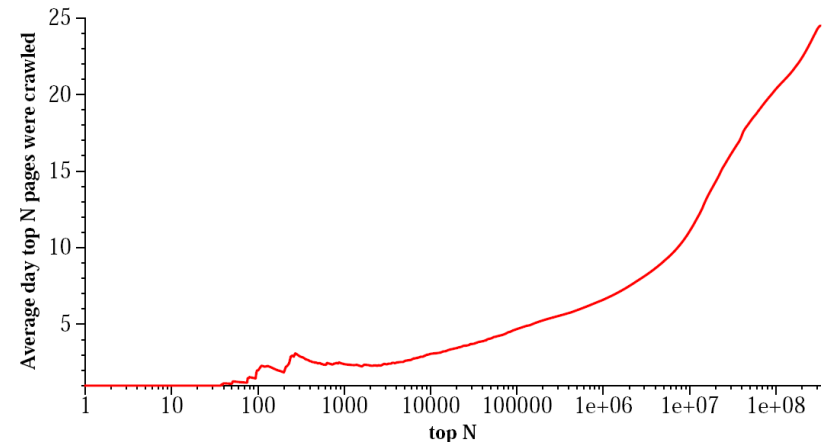
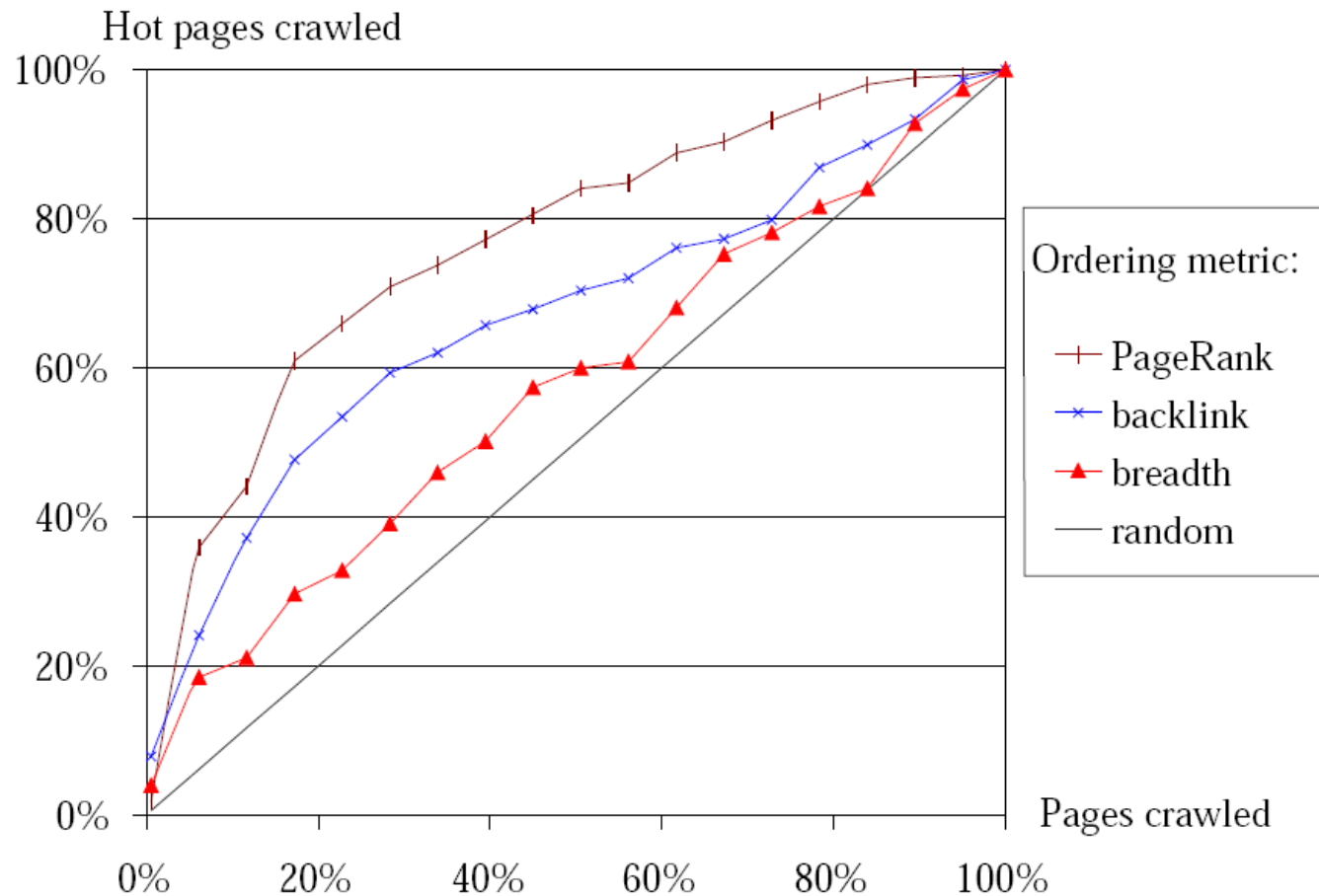


Figure 2: Average day on which the top N pages were crawled

PageRank prioritet er endnu bedre (men mere beregningstung. . .)



Crawl af 225.000 sider på Stanford Universitet [Arasu et al., 2001]

Resourceforbrug

- **Egne resourcer**
 - Båndbredde (global request rate)
 - Lagerplads (brug kompakte representationer)
 - Distribuér på flere maskiner (opdel fx rummet af ULR'er)
- **Andres resourcer (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 filer, senest mødte URL'er)
- **Flaskehals** er ofte disk I/O under tilgang til datastrukturerne
- CPU cykler er ikke flaskehals
- En tunet crawler (på een eller få maskiner) kan crawle 200-400 sider/sek 35 mio sider/dag

Indeksering

Indeksering af dokumenter

- Preprocessér en dokumentsamling så dokumenter med et givet søgeord kan blive returneret hurtigt

Input: dokumentsamling

Output: søgestruktur

Indeksering: Inverteret fil + leksikon

- **Inverteret fil** = for hvert ord w en liste af dokumenter indeholdende w
- **Leksikon** = ordbog over alle forekommende ord (nøgle = ord, værdi = pointer til liste i inverteret fil + evt. ekstra info for ordet, fx længde af listen)

For en milliard dokumenter:

Inverteret fil: totale antal ord 100 mia

DISK

Leksikon: antal forskellige ord 2 mio

RAM

Inverteret Fil

- 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, anker text, . . .)
Kan bruges under ranking

Bygning af index

foreach dokument D i samlingen

Parse D og identificér ord

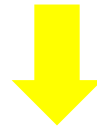
foreach ord w

Udskriv (DocID, w)

if w ikke i leksikon

indsæt w i leksikon

(1, 2), (1, 37), ... , (1, 123) , (2, 34), (2, 37), ... , (2, 101) , (3, 486), ...



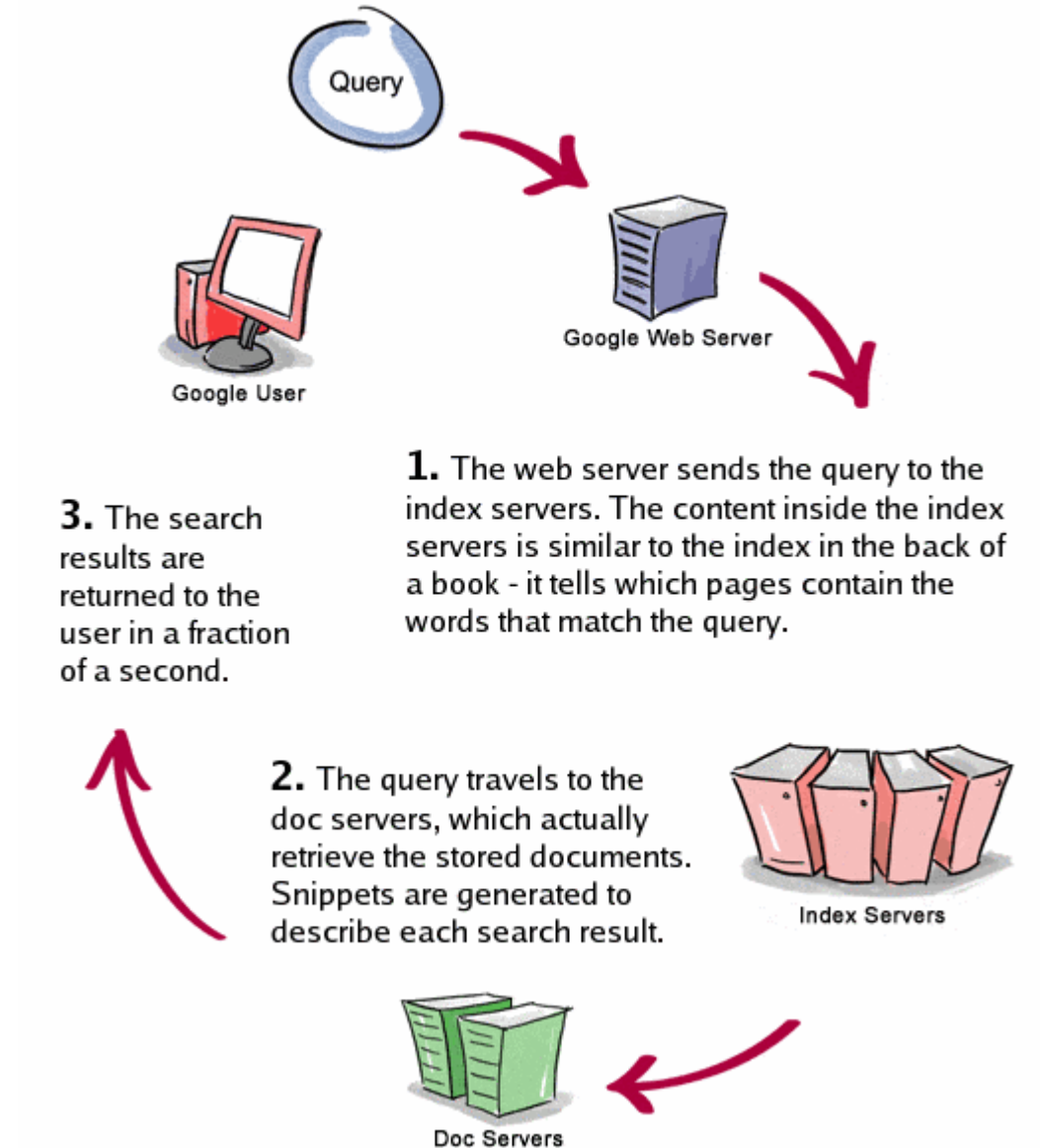
Disk sortering (MapReduce)

(22, 1), (77, 1), ... , (198, 1) , (1, 2), (22, 2), ... , (345, 2) , (67, 3), ...

Inverteret fil

Søgning & Ranking

“Life of a Google Query”



Søgning og Ranking

Søgning: **unf** AND **københavn**

1. Slå **unf** og **københavn** op i leksikon.
Giver adresse på disk hvor deres lister starter.
2. Scan disse lister og “flet” dem (returnér DocID’er som er med i begge lister).

unf: 12, 15, **117**, 155, **256**, ...

københavn: 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.

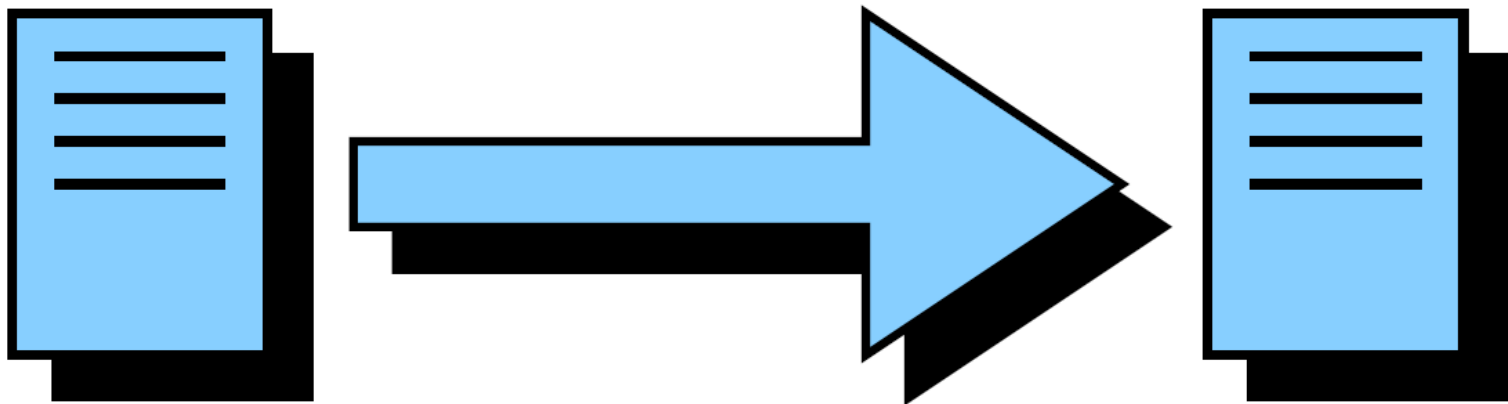
OR og NOT kan laves tilsvarende. Hvis lister har ord-positioner kan frase-søgninger (“**unf københavn**”) og proximity-søgninger (“**unf** tæt på **københavn**”) også laves.

Tekstbaseret Ranking

- Vægt forekomsten af et ord med fx
 - 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 fx 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

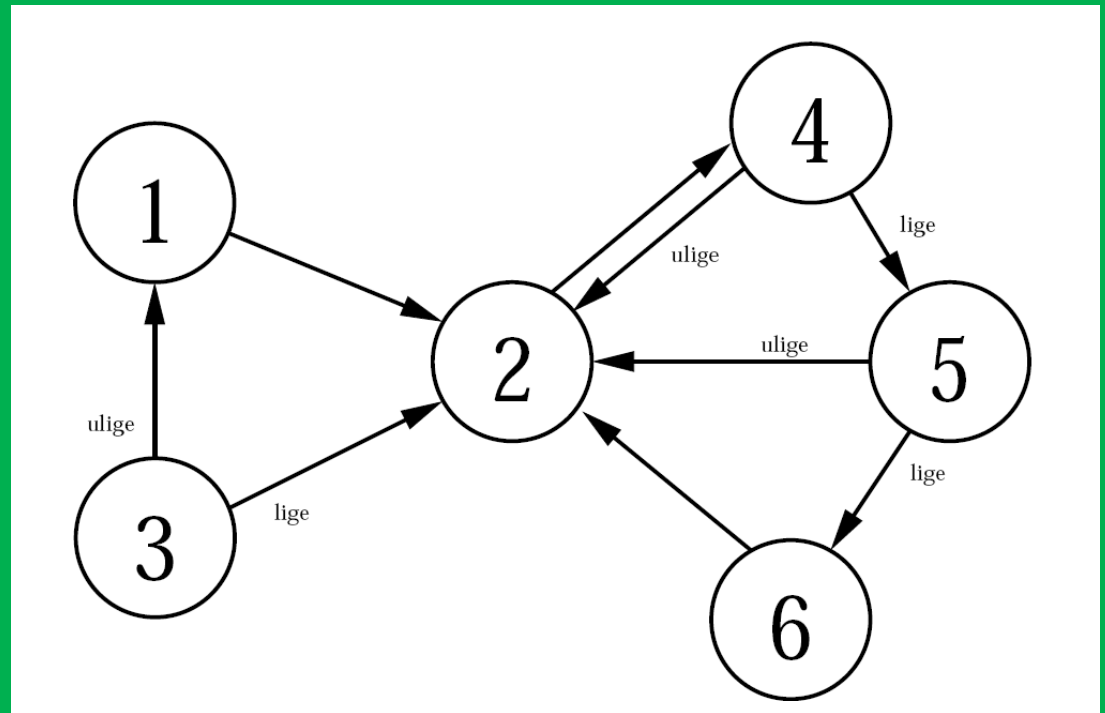
- med 85% sandsynlighed vælger at følge et tilfældigt link fra nuværende side,
- med 15% sandsynlighed vælger at gå til en tilfældig side i hele internettet.

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

Simpel Webgraf

Hvilken knude har størst "rang" ?

- a) 1
- b) 2
- c) 3
- d) 4
- e) 5
- f) 6
- g) Ved ikke

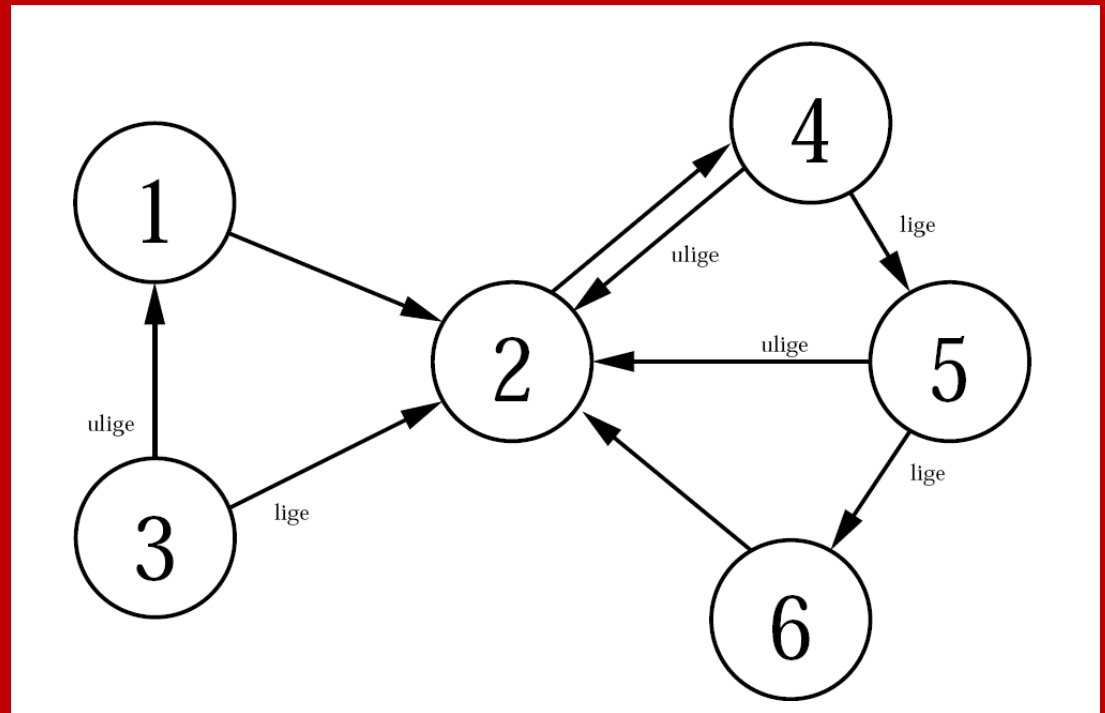


Simpel Webgraf

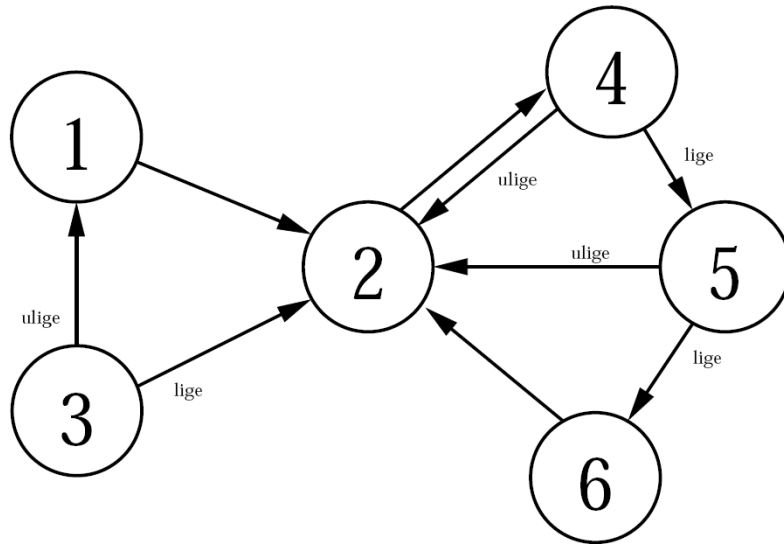
Hvilken knude har størst "rang" ?



- a) 1
- b) 2
- c) 3
- d) 4
- e) 5
- f) 6
- g) Ved ikke



RandomSurfer



Metode RandomSurfer

Start på knude 1

Gentag mange gange:

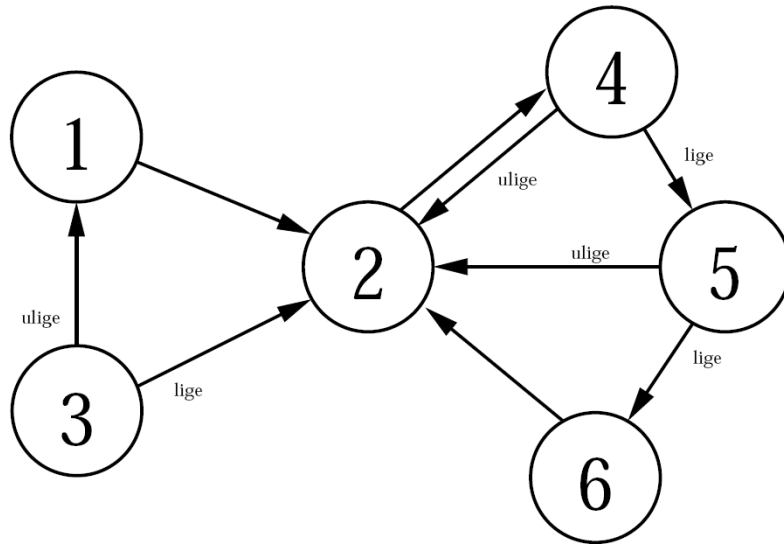
Kast en terning:

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



I starten står man i “1”
med sandsynlighed 1.0
og i “2-6” med
sandsynlighed 0.0

Sandsynligheden for at
stå i “2” efter ét skridt?

Metode RandomSurfer

Start på knude 1

Gentag mange 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

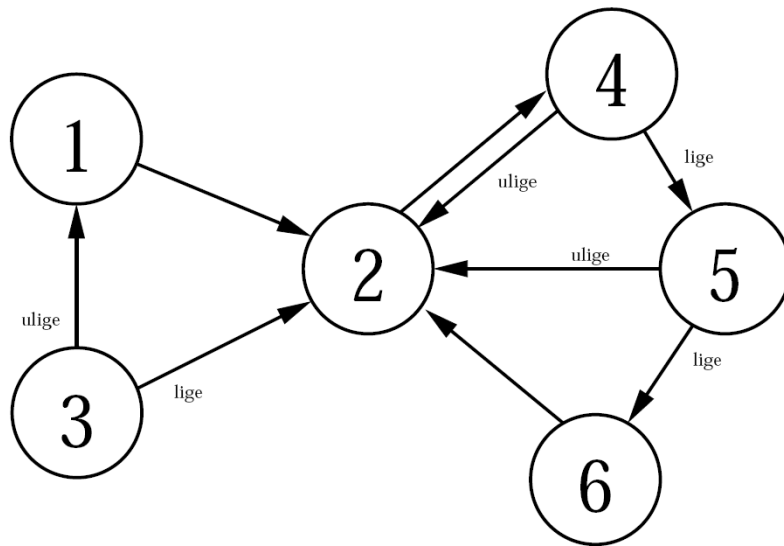
a) $1/6$

b) $31/36$

c) $5/6$

d) 1.0

e) Ved ikke



I starten står man i “1”
med sandsynlighed 1.0
og i “2-6” med
sandsynlighed 0.0

Sandsynligheden for at
stå i “2” efter ét skridt?

Metode RandomSurfer

Start på knude 1

Gentag mange 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



a) $1/6$

b) $31/36$

c) $5/6$

d) 1.0

e) Ved ikke

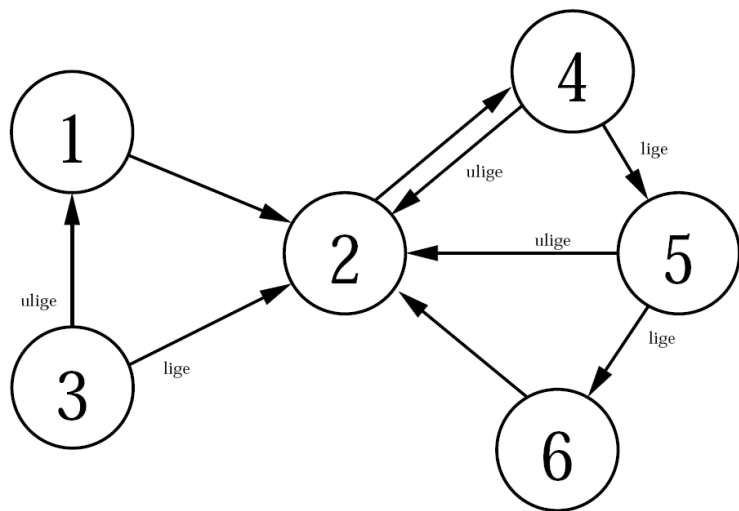
Beregning af Sandsynligheder

Sandsynligheden for at stå i i efter s skridt:

$$p_i^{(s)} = \frac{5}{6} \sum_{j:j \rightarrow i} p_j^{(s-1)} \cdot \frac{1}{\text{udgrad}(j)} + \frac{1}{6} \cdot \frac{1}{6}$$

$$p_1^{(0)} = 1.0 \quad p_2^{(0)} = \dots = p_6^{(0)} = 0.0$$

Simpel Webgraf — 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

Konvergens af PageRank

Hvor mange skridt skal man beregne sandsynlighedsfordelingen for, før den ikke ændrer sig, når der er milliarder af websider?

- a) 10-20
- b) 20-50
- c) 50-100
- d) 100-500
- e) 500-1000
- f) > 1000

Konvergens af PageRank

Hvor mange skridt skal man beregne sandsynlighedsfordelingen for, før den ikke ændrer sig, når der er milliarder af websider?



- a) 10-20
- b) 20-50
- c) 50-100
- d) 100-500
- e) 500-1000
- f) > 1000

Med sandsynlighed 0.9997 har man lavet et tilfældigt spring inden for de seneste 50 skridt

[Avanceret søgning](#)
[Sprogværktøjer](#)Google.dk på: [Føroyskt](#)[Annoncér med Google](#)[Forretningsløsninger](#)[Alt om Google](#)[Google.com in English](#)© 2011 - [Fortrolighed](#)

Britney Spears' Guide to Semiconductor Physics

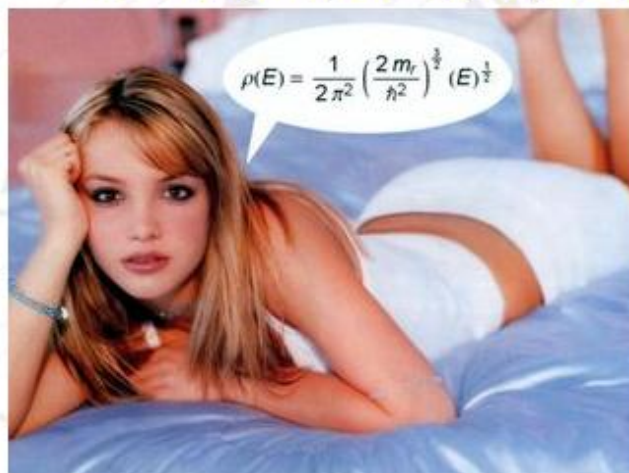
[Spectral Optics](#) Offers custom and standard line of laser and optical components. www.spectraloptics.com

[IGL](#) Red and green positioning lasers. Industrial quality made in germany. www.igl.de.com

[Quattro Titanium](#) Få en glat og behagelig barbering. Modtag en gratis prøve nu! www.efi.dk

Ads by Google

[\[Home \]](#) [\[Picture Galleries \]](#) [\[Britney Spears guide to Semiconductor physics \]](#)
[\[Links \]](#) [\[Lyrics \]](#) [\[Advertise \]](#) [\[Stuff \]](#) [\[Chat \]](#) [\[Link to us \]](#) [\[Awards \]](#) [\[Britney Gossip \]](#)



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 semiconductor laser components that have made it possible to hear her super music in a digital format.



☒ Web ☐ britneyspears.ac

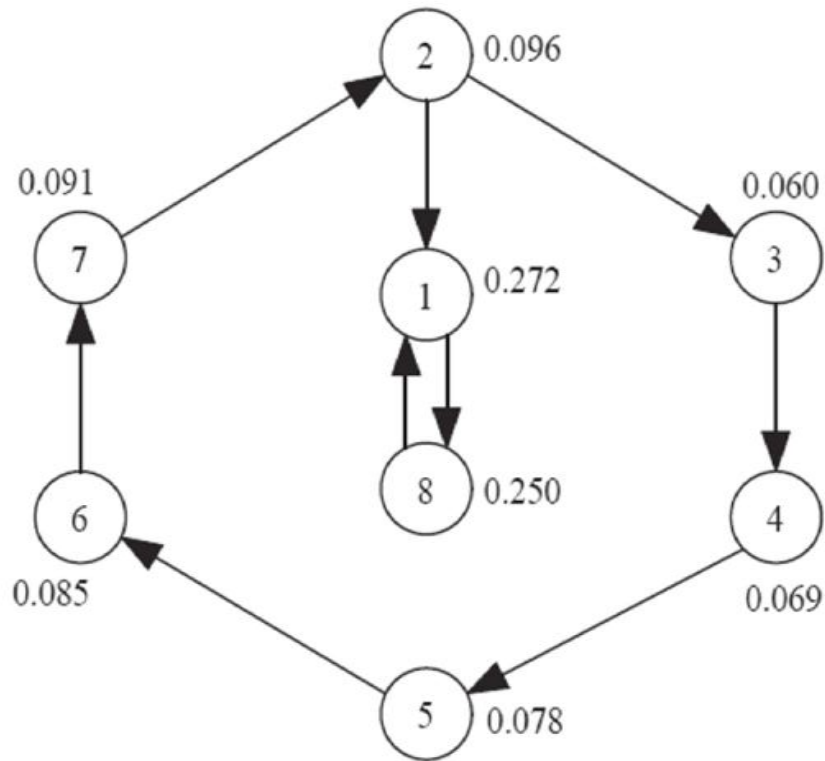
[Scientific Calculator](#)

[Advertise Here](#)

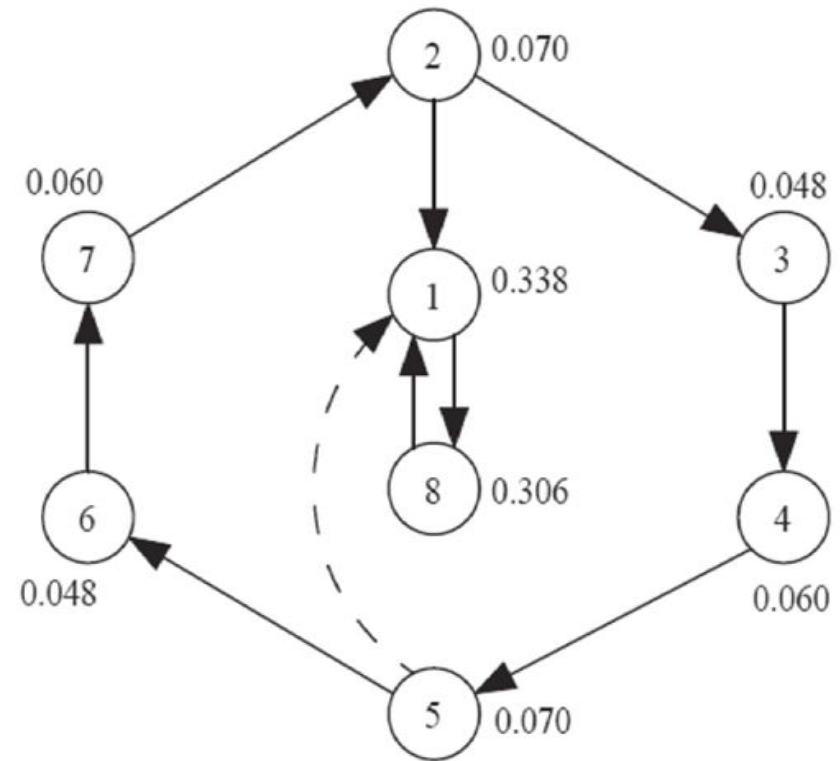
[Click here](#) to donate food to the starving people of the world.



Søgemaskine Optimering

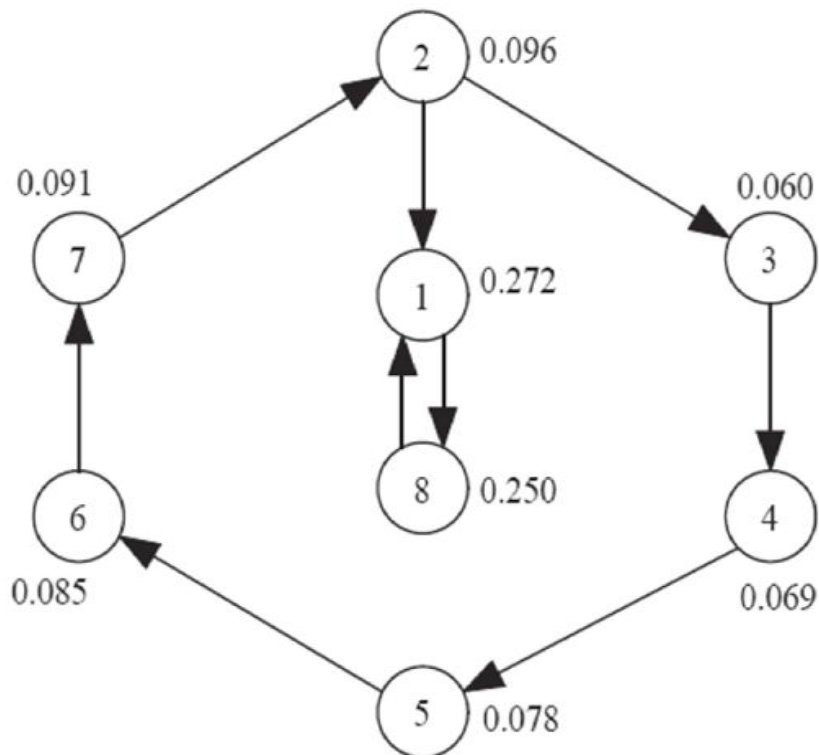


(a) The original graph.



(b) One optimal new link.

Søgemaskine Optimering

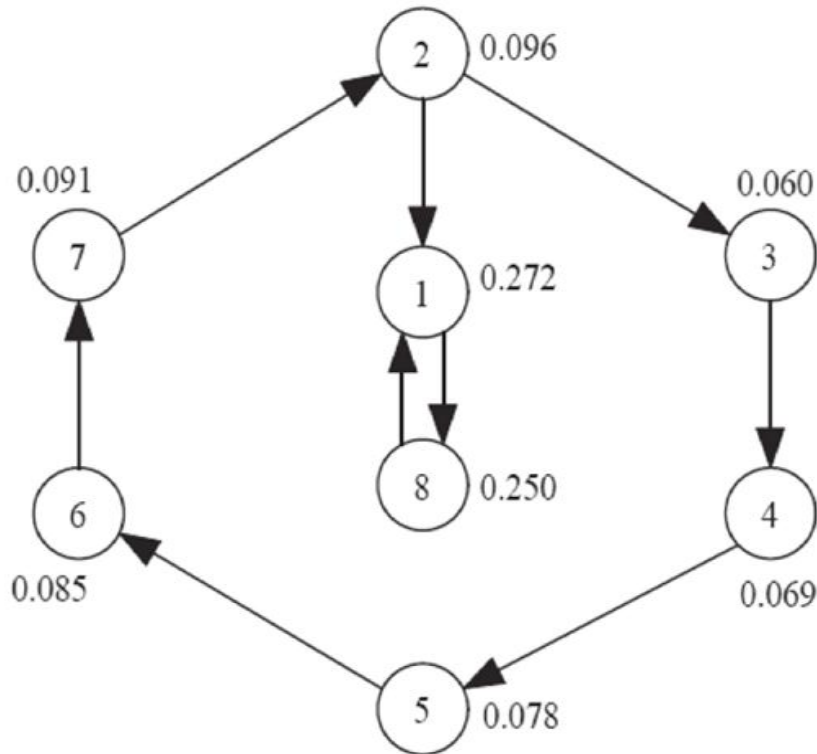


(a) The original graph.

Hvilke to kanter skal man tilføje for at maksimere "1"s PageRank ?

- a) (6,1) og (7,1)
- b) (5,1) og (7,1)
- c) (6,1) og (5,1)
- d) (6,1) og (4,1)
- e) Ved ikke

Søgemaskine Optimering



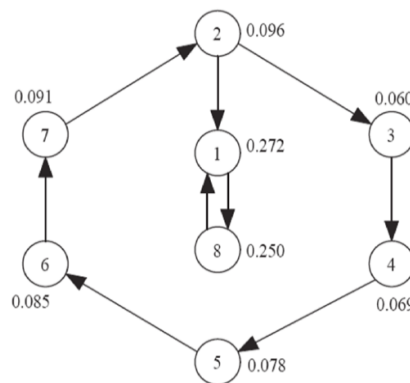
(a) The original graph.

Hvilke to kanter skal man tilføje for at maksimere "1"s PageRank ?

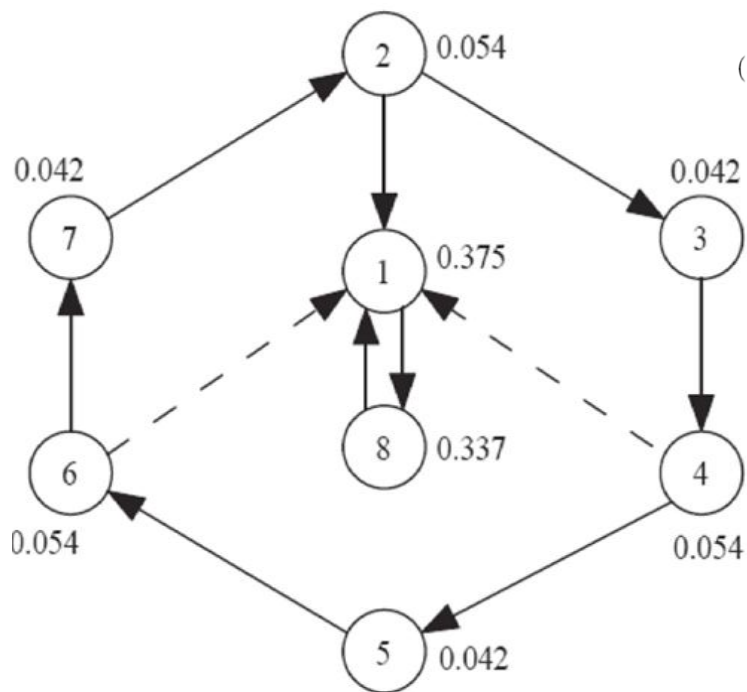
- a) (6,1) og (7,1)
- b) (5,1) og (7,1)
- c) (6,1) og (5,1)
- d) (6,1) og (4,1)
- e) Ved ikke



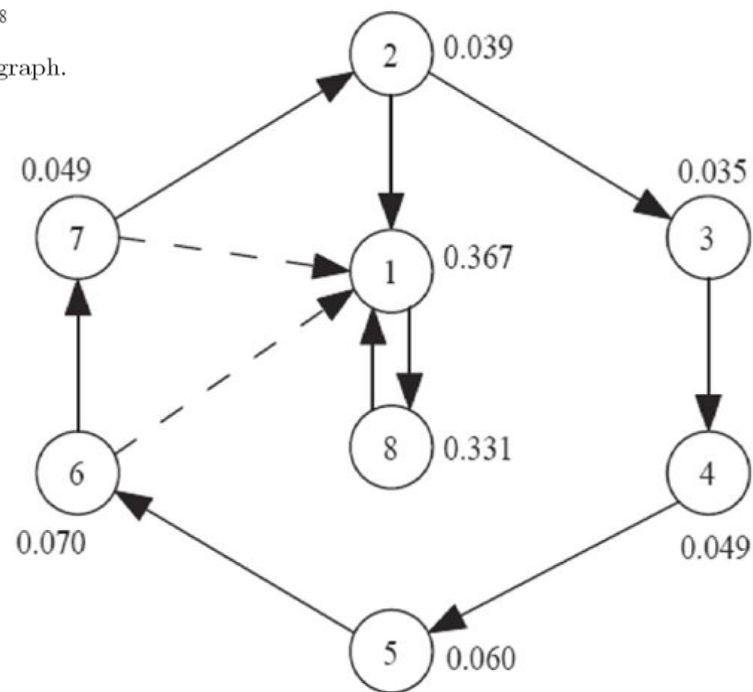
Søgemaskine Optimering



(a) The original graph.



(c) Two optimal new links.



(d) Two new links from the most popular nodes prior to the modification.

Søgemaskine Optimering (SEO)

- **Mål:** Optimer websider til at ligge højt på søgemaskiner
- **Metoder:** Lav nye websider der peger på en side, bytte links, købe links, lave blog indlæg, meta-tags, ...
- **SEO:** Milliard industri
- **Søgemaskiner:** Blacklisting, fjernelse af dubletter, ...



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.

Gør-det-selv

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

- **Projekt:** Lav en søgemaskine for domæne .dk
- 15 studerende
- 4 parallelt arbejdende grupper (crawling, indexing, PageRank, søgning/brugergrænseflade)
- Erfaring: Rimmelig vellykket søgemaskine, hvor rankningen dog kræver **yderligere finjustering...**

Referencer

- 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.
- Martin Olsen, *Link Building. Ph.d. afhandling, Datalogisk Institut, Aarhus Universitet*, august 2009.