Algorithms for Web Indexing and Searching

Rolf Fagerberg

Fall 2004

The Internet

- Very large amount of information.
- Unstructured.

How do we find relevant info?

The Internet

- Very large amount of information.
- Unstructured.

How do we find relevant info?

Search Engines!

The Internet

- Very large amount of information.
- Unstructured.

How do we find relevant info?

Search Engines!

History:

94: Lycos,. . . : First search engines

96: Alta Vista: many pages indexed .

99: Google: many pages indexed *and* good ranking.

Modern Search Engines

Impressive performance. E.g. Google:

- Searches $4.3 \cdot 10^9$ pages (Sept 04).
- Response time $\approx$ 0,1 seconds.
- 1000 queries per second.
- Finds relevant pages (*Do you feel lucky. . . ?*)

Modern Search Engines

Impressive performance. E.g. Google™:

- Searches $4.3 \cdot 10^9$ pages (Sept 04).
- Response time $\approx$ 0,1 seconds.
- 1000 queries per second.
- Finds relevant pages (*Do you feel lucky. . . ?*)

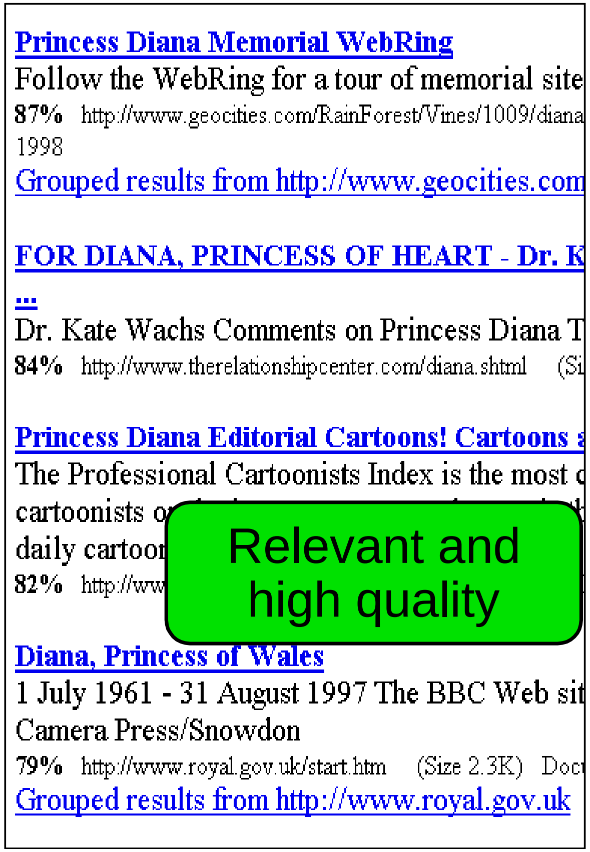
Who uses bookmarks any more?

Not So Modern Search Engines

Advanced methods do make a difference (example, circa 1998):

`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

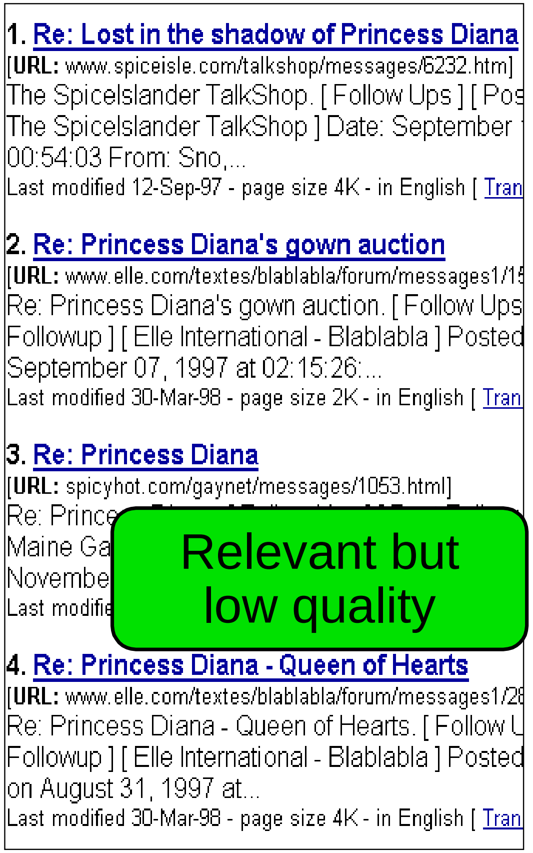
FOR DIANA, PRINCESS OF HEART - Dr. K
...
Dr. Kate Wachs Comments on Princess Diana T
84% http://www.therelationshipcenter.com/diana.shtml (Si

Princess Diana Editorial Cartoons! Cartoons a
The Professional Cartoonists Index is the most c
cartoonists o
daily cartoon
82% http://ww

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 SpiceIslander TalkShop. [Follow Ups] [Pos
The SpiceIslander TalkShop] Date: September
00:54:03 From: Sno,...
Last modified 12-Sep-97 - page size 4K - in English [Tran

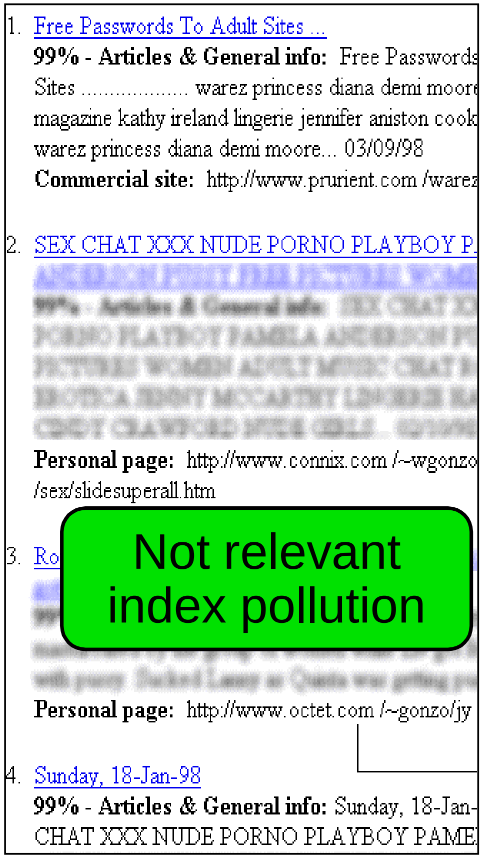
2. Re: Princess Diana's gown auction
[URL: www.elle.com/textes/blablabla/forum/messages1/15
Re: Princess Diana's gown auction. [Follow Ups
Followup] [Elle International - Blablabla] Posted
September 07, 1997 at 02:15:26:...
Last modified 30-Mar-98 - page size 2K - in English [Tran

3. Re: Princess Diana
[URL: spicyhot.com/gaynet/messages/1053.html]
Re: Prince
Maine Ga
Novembe
Last modifie

4. Re: Princess Diana - Queen of Hearts
[URL: www.elle.com/textes/blablabla/forum/messages1/28
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.prurient.com /warez

2. SEX CHAT XXX NUDE PORNO PLAYBOY P
[illegible]
Personal page: http://www.connix.com /~wgonzo
/sex/slidesuperall.htm

3. Ro
[illegible]
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

Course Motivation

How does Google™ work?

Course Motivation

How does Google™ work?

⇓

How do search engines work?

Course Motivation

How does Google™ work?

⇓

How do search engines work?

⇓

Algorithms for web indexing and searching

Subjects – Search Engines

Aquiring data

- Web crawling

Processing data

- Parsing
- Indexing
- Sorting
- Duplicate removal

Storing data

- Data structures storing:
  - Keywords
  - URLs
  - links
  - full pages
- Distribution of data storage
- Compression of data.

Subjects – Search Engines

Searching in data

- Query types
- Algorithms

Ranking results

- Word based (number and position of occurences)
- Link based (PageRank, others)
- Query dependent
- Query independent
- Other heuristics (e.g. recognition of home pages, news, . . .)

Related Subjects

- String algorithms and data structures.
- Techniques for massive data sets.
- Internet protocols
- Classical IR (high-dimensional vector spaces).
- Search engine evaluation.
- Graph models of the web.
- Data mining.
- Similarity measures (nearest neighbor, clustering, latent semantic indexing).
- Web caching.
- Web applications of game theory (auctions, mechanism design).

Formal Course Description

Prerequisites:	DM02
Literature:	Research papers
Evaluation:	Implementation project, oral exam
Credits:	7.5 ECTS
Course language:	Danish or English

Project

Implement a search engine

Goal: Search engine for domain `.dk`

- Large scale project.
- Programming groups (crawling, indexing, ranking, query interface).
- Cooperation.

Informal Course Description

In the course you will meet:

- Real life search engines
- Algorithms and data structures
- Mathematical models
- Hands-on experience
- Teamwork