Algorithms for Web Indexing and Searching

Gerth Stølting Brodal and Rolf Fagerberg

Fall 2002

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

Course Outline

1. Introduction to Course
2. General Anatomy of Web Search Engines
3. Building blocks of Search Engines
   (a) Web Crawlers
      - Anatomy of crawlers
      - Crawling strategy
   (b) Index
      - Inverted files
      - Suffix trees
      - Signature files
      - Compression
      - Issues of efficient construction
      - Duplicate removal

Course Outline

(c) Types of Queries

(d) Ranking

- Textbased methods
  - Vector based methods
  - Latent semantic indexing
- Link based methods
  - PageRank
  - HITS
  - SALSA
  - Others

Course Outline

4. Further topics
   (a) Clustering
   (b) Automatic Categorization/Hierarchy Building
   (c) Evaluation of search engines
   (d) Structure of and Models for the Web Graph
   (e) Data Mining

Formal Course Description

Prerequisites:	dADS
Literature:	Handouts
Course language:	Danish or English
Credits:	2 points/10 ECTS
Evaluation:	Programming project

Course page:

`http://www.daimi.au.dk/~gerth/webalg02/index.html`

Programming Project

Implement a Web Search Engine

Programming Project

Implement a Web Search Engine

Distributed project

Groups (2–4 persons) doing:

Web crawling

Index building

Ranking

Query interface

Start: index Aarhus University website

Goal: index domain `.dk`