Search Engines

Exercise 2: Crawling

Dustin Lange & Saeedeh Momtazi

28 April 2011

Googlewhacking Evaluation

My favourite terms with perfect scores:

Potsdam	Search	HPI
Archäologiefund	Bartkrätze	genussfreudig
Ericssonmotor	gartenbauwettbewerb	Hinterlegungsverfahren
gartenbaukurs	Gärtnerlohn	hühnerkäfig
Haveldampfschifffahrtsgesellschaft	Sekundärstandarddosimetrielabor	Prozessorerweiterungen
neutronenbeschleunigung	Wegtragsel	räuberhöhlen
Vergrauungsinhibitor		Ziegenweide

Googlewhacking Evaluation

- 28 submissions – thanks to all!
- 10 submissions with perfect results
- First submission = first submission with perfect results
- Winner submission at 10:21 am
- Congratulations to Marika Marszalkowski and Peter Retzlaff!
  - kreuchend, fraternisierendes, druckendes

International Edition
Search Engines
INFORMATION RETRIEVAL IN PRACTICE
W. Bruce Croft • Donald Metzler • Trevor Strohman
PEARSON

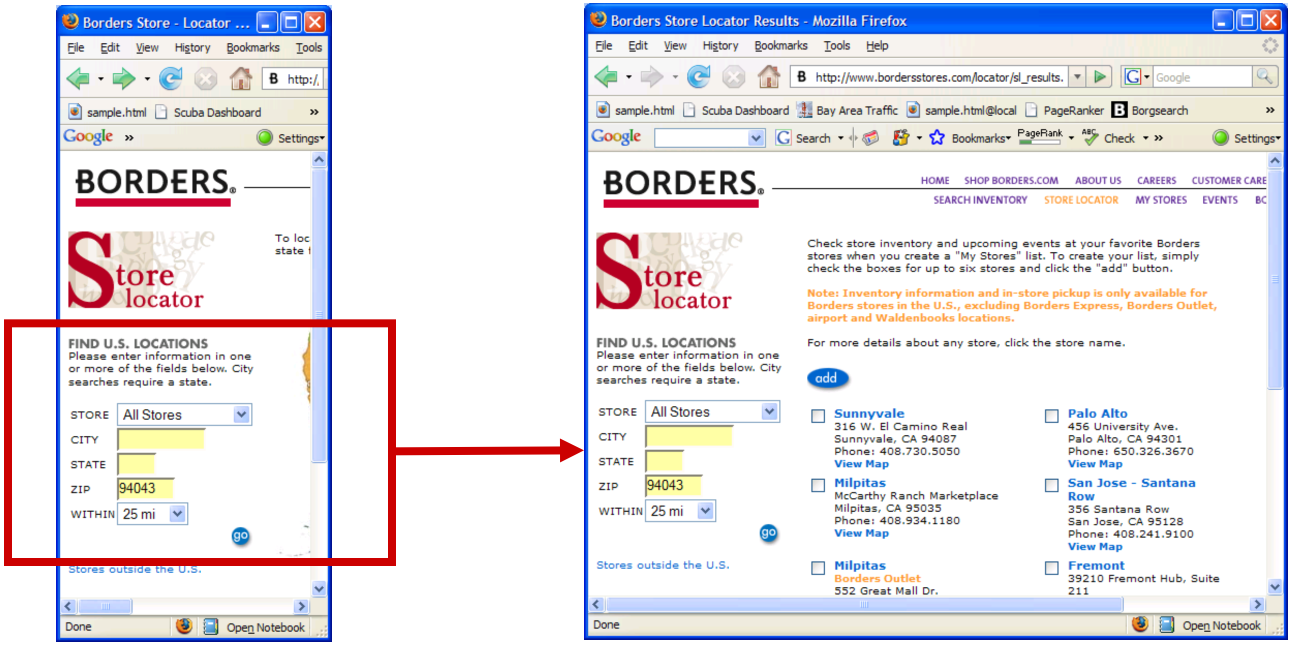
Task: Journal Club

- Read one of the presented papers
- Present the key ideas (not the entire paper)
  - 5-10 minutes
  - Focus on answering the questions
  - Your fellow students should get the idea

Paper 1: Google‘s Deep-Web Crawl

http://www.cs.cornell.edu/~lucja/Publications/i03.pdf

- VLDB 2008
- Cited 87 times


Borders Store - Locator ...
File Edit View History Bookmarks Tools
sample.html Scuba Dashboard
Google Settings
Borders Store Locator Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.bordersstores.com/locator/sl_results.
sample.html Scuba Dashboard Bay Area Traffic sample.html@local PageRanker Borgsearch
Google Search Bookmarks PageRank Check Settings
BORDERS.
HOME SHOP BORDERS.COM ABOUT US CAREERS CUSTOMER CARE
SEARCH INVENTORY STORE LOCATOR MY STORES EVENTS BC
Store locator
Done
Open Notebook

To loc state

FIND U.S. LOCATIONS
Please enter information in one or more of the fields below. City searches require a state.

STORE All Stores
CITY
STATE
ZIP 94043
WITHIN 25 mi
go

Stores outside the U.S.

Check store inventory and upcoming events at your favorite Borders stores when you create a "My Stores" list. To create your list, simply check the boxes for up to six stores and click the "add" button.

Note: Inventory information and in-store pickup is only available for Borders stores in the U.S., excluding Borders Express, Borders Outlet, airport and Waldenbooks locations.

For more details about any store, click the store name.

add

□ Sunnyvale
316 W. El Camino Real
Sunnyvale, CA 94087
Phone: 408.730.5050
View Map

□ Milpitas
McCarthy Ranch Marketplace
Milpitas, CA 95035
Phone: 408.934.1180
View Map

□ Milpitas
Borders Outlet
552 Great Mall Dr.

□ Palo Alto
456 University Ave.
Palo Alto, CA 94301
Phone: 650.326.3670
View Map

□ San Jose - Santana Row
356 Santana Row
San Jose, CA 95128
Phone: 408.241.9100
View Map

□ Fremont
39210 Fremont Hub, Suite 211

- Which attributes to query?
- Which values to use?

Google™

Slide contents by Jayant Madhaven, Google Inc., 2008

Paper 1: Google‘s Deep-Web Crawl

Questions

1. Introduction
   a. What is the **deep web**? Give an example. Which kind of deep web categories (from the lecture) is addressed in the paper?
2. The surfacing problem
   a. Find an **example form** in the web that is not mentioned in the paper. Give examples for the following terms using your example form: web form, inputs, selection inputs, wild card value, presentation inputs, database.
   b. What is a query template? Reuse your example.
3. Selecting query templates
   a. What are characteristics of good query templates?
   b. How is informativeness determined? (brief)
   c. Briefly describe the algorithm for **incremental template search**.
4. Generating input values
   a. Why is it difficult to generate appropriate input values for text boxes?
   b. Briefly describe the **iterative probing algorithm**. (How are seed words determined? What happens during one iteration? How are final keywords selected?)

Paper 2: Do Not Crawl in the DUST

http://www2007.org/papers/paper194.pdf

- DUST – Different URLs with Similar Text
- Examples:
  - “http://domain.name/index.html” → “http://domain.name”
  - “http://news.google.com” → “http://google.com/news”
- How to find URL transformation rules from a list of URLs?

- WWW 2007
- Cited 37 times

Slide contents by Uri Schonfeld, Technion, 2007

Paper 2: Do Not Crawl in the DUST

Questions

1. Introduction
   a. What is **DUST**? Find an example that is not mentioned in the paper.
2. Problem Definition
   a. What is the **definition** of DUST rules?
   b. What is the **definition** of valid DUST rules?
3. Basic heuristics (briefly describe the three **basic heuristics**)
   a. Why are rules with large support sufficient?
   b. Why are small buckets more interesting?
   c. How can the similarity of two pages help?
4. DustBuster
   a. Briefly describe the **main algorithm for discovering likely DUST rules**. (Do not discuss details.)
   b. What are redundant rules? How can they be detected (what is the key idea)?
   c. Why is validation of DUST rules necessary? How can rules be validated?

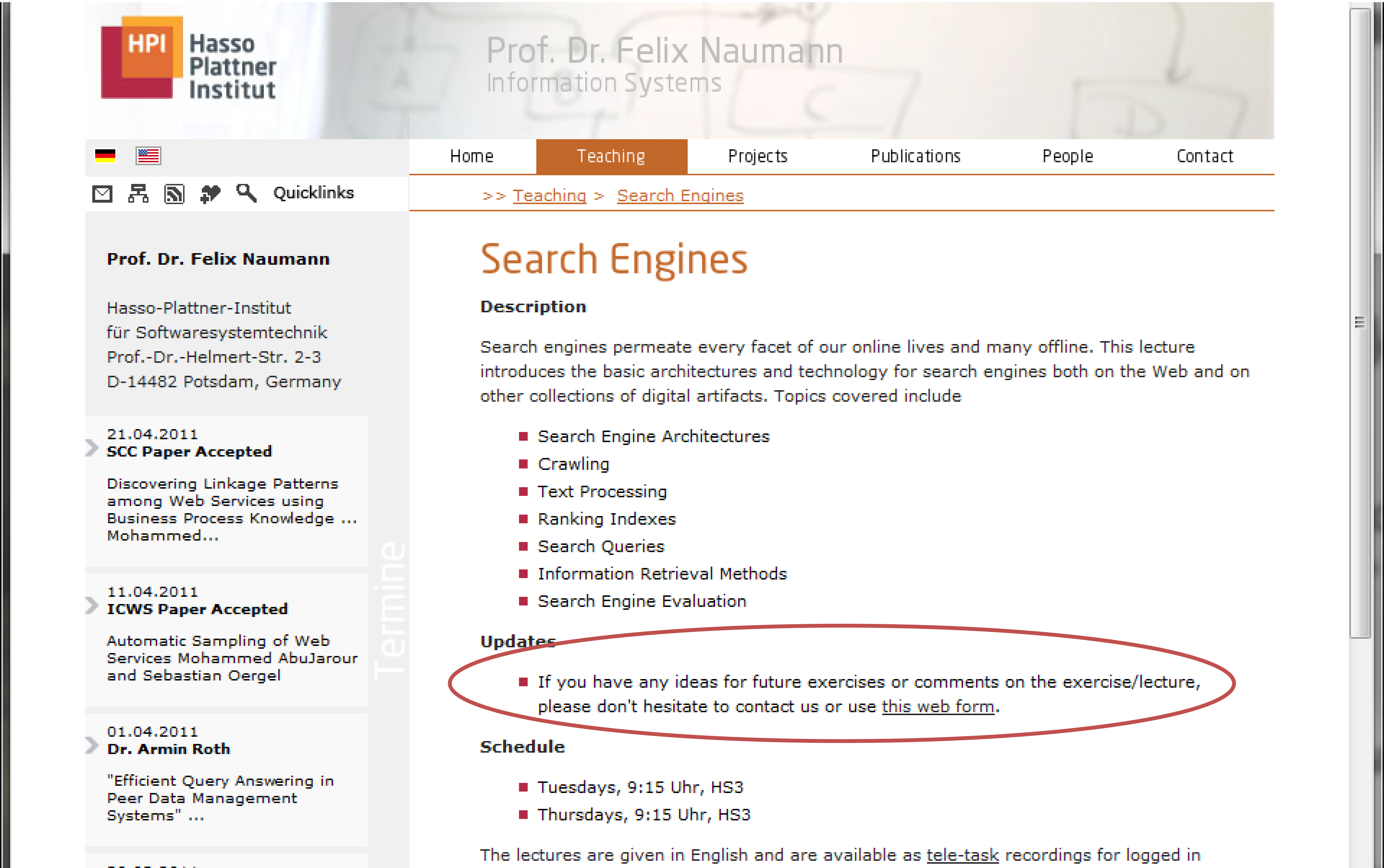
Selection Procedure

- Who wants to read which paper?

Submissions & Next Exercise

- Selection:
  - Select a paper **today**: http://goo.gl/jv2ED
- Submissions:
  - Create slides to present your selected paper.
  - Send us your presentation
    - as PDF or PPT(X) or ODP: *SearchEngines**2**[Name1][Name2].[pdf|ppt|pptx|odp]*
    - via e-mail with subject: *Search Engines 2 [Paper 1|Paper2]*
    - to *dustin (dot) lange (at) hpi (dot) …*
    - until **4 May 2011, 5:00 pm**
- On **5 May 2011**: Be prepared to present the paper
  - English (or German)
  - Absent: Send me an e-mail in advance

Feedback



>> Teaching > Search Engines

Prof. Dr. Felix Naumann

Hasso-Plattner-Institut
für Softwaresystemtechnik
Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam, Germany

21.04.2011
SCC Paper Accepted

Discovering Linkage Patterns among Web Services using Business Process Knowledge ... Mohammed...

11.04.2011
ICWS Paper Accepted

Automatic Sampling of Web Services Mohammed AbuJarour and Sebastian Oergel

01.04.2011
Dr. Armin Roth

"Efficient Query Answering in Peer Data Management Systems" ...

Search Engines

Description

Search engines permeate every facet of our online lives and many offline. This lecture introduces the basic architectures and technology for search engines both on the Web and on other collections of digital artifacts. Topics covered include

- Search Engine Architectures
- Crawling
- Text Processing
- Ranking Indexes
- Search Queries
- Information Retrieval Methods
- Search Engine Evaluation

Updates

- If you have any ideas for future exercises or comments on the exercise/lecture, please don't hesitate to contact us or use this web form.

Schedule

- Tuesdays, 9:15 Uhr, HS3
- Thursdays, 9:15 Uhr, HS3

The lectures are given in English and are available as tele-task recordings for logged in

Thanks for Listening

- Updates
  - – See website
  - – Mailing list: tbd (very soon)
- Questions
  - – Via e-mail:
    - dustin (dot) lange (at) hpi (…)
    - saeedeh (dot) momtazi (at) hpi (…)
  - – Office: A-1.6 / A-1.7