DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

Ontologiegestützte Suche in unstrukturierten Daten

Veranstalter: Prof. Dr. Lausen

Betreuer: Kai Simon, Thomas Hornung

(Team) – Projekt Anforderungen

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

- Bachelor (6 ECTS) [entsprechen 180 Stunden]
  - Softwareentwicklung (Lasten-,Pflichtenheft,..)
  - Gruppengröße 2-5
  - Protokoll, Referat, Rechnerpräsentation

- Master/Diplom (16 ECTS/15 KP) [entsprechen 480 Stunden]
  - wissenschaftliche Methoden zur Problemlösung
  - eigenständige Themeneinarbeitung anhand von Fachliteratur
  - Gruppengröße 2-5
  - Schriftlich je Gruppenmitglied 15-25 Seiten
  - Mündliche Präsentation je Mitglied 15 Minuten

ANMELDUNG DER PROJEKTTEILNAHME
BEIM PRÜFUNGSAMT IST
ERFORDERLICH UND VERBINDLICH

Voraussetzungen

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

- Gute
  - Programmierkenntnisse in Java
  - Vertrautheit mit dem Umgang von SW Entwicklungsumgebungen, Programmbibliotheken und Dokumentationssystemen

- Grundkenntnisse in den Gebieten
  - Text Mining
  - Information Retrieval
  - Relationen Datenbanken

  sind von Vorteil, können aber in einer Einarbeitungsphase erworben werden.

- Interesse, sich auf ein neues Forschungsgebiet einzulassen.
- Angestrebte Vertiefung im Gebiet Kommunikation und Datenhaltung mit entsprechender Abschlussarbeit.

Organisation

Zeit & Ort:

- Donnerstag 16-19 Uhr (s.t.)
- Raum 00-006, Geb. 051
- Termin für Besprechungen, Fragen & Abstimmung über weiteres Vorgehen,
- ansonsten selbständige Arbeit.

Project Topic

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

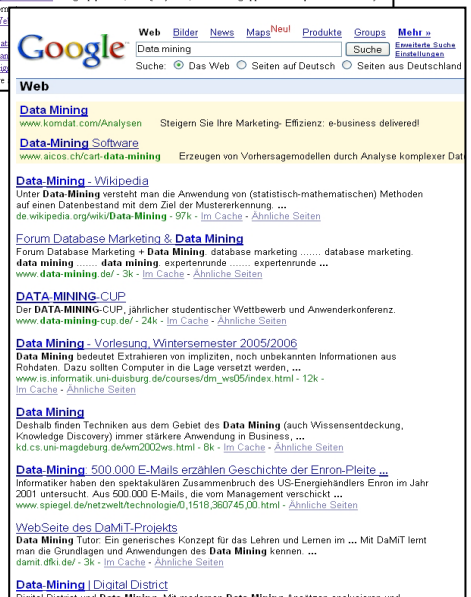
Information Retrieval comprises three activities:

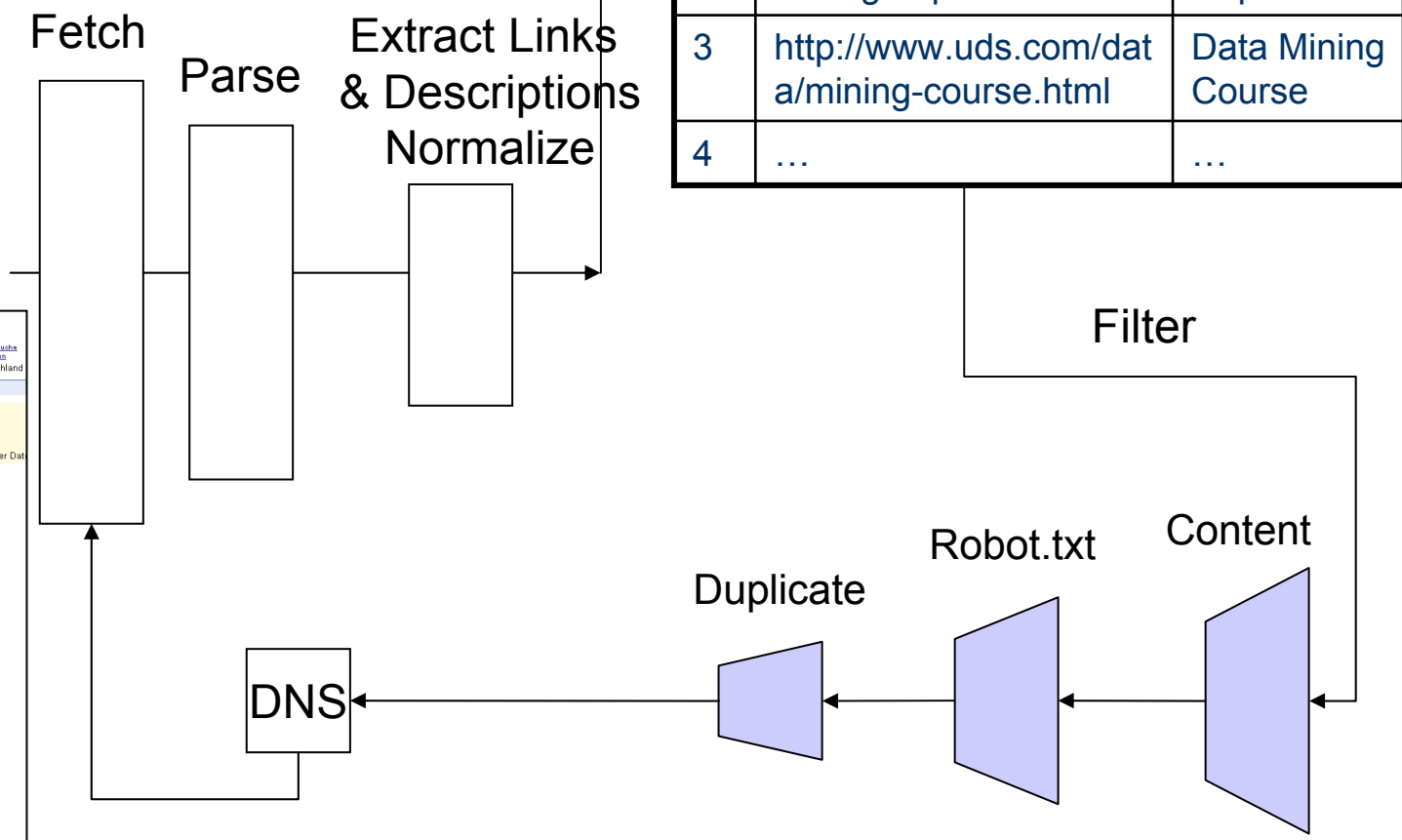
- Crawling and Indexing
- Searching
- Ranking

Crawling and Indexing

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

Seed Pages

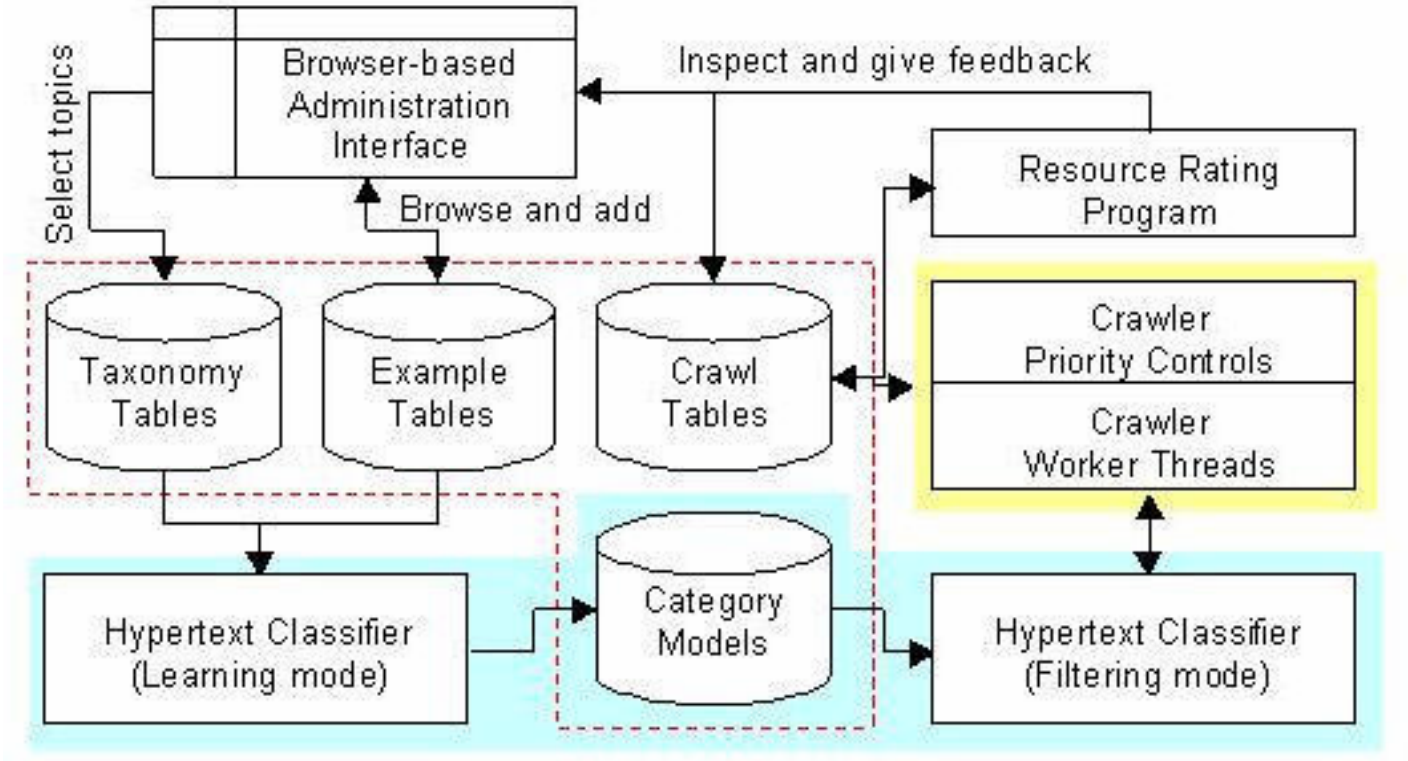




ID	URL	Description
1	http://www.data-mining.de/	Data Mining
2	http://www.data-mining-cup.de	Data Mining Cup
3	http://www.uds.com/data/mining-course.html	Data Mining Course
4	…	…

Focus Web Crawling

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg



source: http://www.cse.iitb.ac.in/~soumen/focus/

Assignments

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

- Project [Bachelors]
  - Search and read literature describing Web crawler systems
  - Give a small presentation of the crawler architecture your team deploys. [max. 10 minutes]
    - Relational DB schema
    - Software components you will use [Parser, Indexer, …]
    - Components to implement + functionality
    - Distribution crawling
- Team project [Master/Diplom]
  - Decide the focus of your crawler
  - Compare different focus crawler approaches described in the literature.
  - Give a small presentation with a summery and techniques you like to deploy. [max. 10 minutes]

Tools

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

Tools you will need in the project:

- Eclipse + SVN (Subclipse)
- Software Engineering Tools (Bachelor)
  ArgoUML http://argouml.tigris.org/
  Microsoft VISIO

  [Objects in class diagrams]
  [association with up to 8 classes provides]

Literatur zum Proseminar

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

- [Baez99] Baeza-Yates, Richardo; Ribeiro-Neto, Berthier: **Modern Information Retrieval.** ACM Press, New York, 1999, 0-21-39829-X.
- [Ferb03] Ferber, Reginald: **Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web**. dpunkt.verlag, Heidelberg, 2003, 3-89864-213-5.
- [Chak02] Soumen Chakrabarti: **Mining the Web Discovering Knowledge from Hypertext Data.** Morgen-Kaufmann, 2002,1-55860-754-4.
- [Kuro04] Dominik Kuropka: **Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und –Retrieval mit relationalen Datenbanken**. In Advances in Information Systems and Management Science, Bd.10, 2004, 3-8325-0514-8.
- [Chan06] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, Robert Gruber: **Bigtable: A Distributed Storage System for Structured Data p. 205-218 OSDI 06**

Free Online Sources

- **http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html**

Relevant Web resources

- Apache Lucene [http://lucene.apache.org/]
  - full-featured open-source text search engine in Java
  - Indexing + Search technology
- Nutch [http://lucene.apache.org/nutch/]
  - Open-source search engine with all functionalities
  - Crawling + HTML parsing functionalities
- Solr [http://lucene.apache.org/solr/]
  - full-featured search server based on Lucene
- Hadoop [http://lucene.apache.org/hadoop/]
  - distributed computing platform

Distributed File System

DBIS Research Group
Computer Science Department
Albert-Ludwigs-University Freiburg

- Bigtables [Chan06] Googles' distributed file system (GFS)

	Contents:	anchor:cnnsi.com	anchor:my.look.ca	Language
"com.cnn.www" →	<html>…</html> ← t_1 <html>…</html> ← t_2	"CNN" ← t_9	"CNN.com" ← t_8	en

- Hadoop Distributed File System (HDFS)