

Ontologiestützige Suche in unstrukturierten Daten

Veranstalter: Prof. Dr. Lausen

Betreuer: Kai Simon, Thomas Hornung

(Team) – Projekt Anforderungen

- 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**

- 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.

Zeit & Ort:

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

Information Retrieval comprises three activities:

- Crawling and Indexing
- Searching
- Ranking

Crawling and Indexing

Seed Pages

Top: Computers: Software: Databases: Data Mining (213)

- Consultants (40)
- Events (23)
- Online Publications (14)
- Projects and Research (22)
- Public Domain Software (29)
- Tool Vendors (72)

See also:

- Computers: Artificial Intelligence: Machine Learning (235)
- Computers: Software: Databases: Data Warehousing (321)
- Computers: Software: Databases: Data Warehousing: Data Integrity and Cleaning Tools (45)
- Computers: Software: Databases: OLAP (77)
- Computers: Software: Internet: Site Management: Log Analysis (274)
- Computers: Software: Marketing: Market Analysis (14)
- Reference: Knowledge Management: Knowledge Discovery (183)
- Reference: Knowledge Management: Knowledge Discovery: Information Visualization (119)
- Reference: Knowledge Management: Knowledge Discovery: Text Mining (23)
- Science: Math: Statistics (232)

• About.com on Data Mining - About.com presents a collection of original feature articles, net links, forum discussions, and more.

• The Data Mine - Launched in April 1994 to provide information about Data Mining (AKA Knowledge Discovery).

• Data Mining and Knowledge Discovery - A peer-reviewed journal publishing articles on all aspects of Knowledge Discovery (patterns and models) from data. Accepts submissions of original research or technical surveys.

• Data Mining Resources - A collection of Data Mining links edited by the Central Connecticut State University.

• Distribution Analysis module for PostgreSQL - Graphical parameter distribution and function relations analysis.

• Estimating Consumer Benefits and Modeling Left (Overheads) - In assessing the potential of data mining based on predicting behavior of some target population. We present a methodology for initial cost/benefit analysis on a scalable model accuracy.

• Tool Ratings - Describes personal, professional and research capabilities on Intelligent Systems, Data Mining, and CRM.

• Kurt Thearling Data Mining and CRM - Information on data mining and CRM technology. Includes a list of software products.

• Tyson Software - Flagship product, The Query Tool, a data mining application that performs data analysis from a variety of data sources.

• Web

• Das

• Bild

• Die

• Die

Google Web Bilder News Maps Neu! Produkte Groups Mehr »
Data mining [Suche] Erweiterte Suche
Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web

Data Mining
www.komdat.com/Analysen Steigern Sie Ihre Marketing-Effizienz: e-business delivered!

Data-Mining Software
www.acicos.ch/cart-data-mining Erzeugen von Vorhersagemodellen durch Analyse komplexer Daten

Data Mining - Wikipedia
Unter **Data Mining** versteht man die Anwendung von (statistisch-mathematischen) Methoden auf einen Datenbestand mit dem Ziel der Mustereerkennung. ...
de.wikipedia.org/wiki/Data_Mining - 97k - im Cache - Ähnliche Seiten

Forum Database Marketing & Data Mining
Forum Database Marketing + Data Mining database marketing database marketing.
data mining ... data mining expertenrunde expertenrunde ...
www.data-mining.de/ - 3k - im Cache - Ähnliche Seiten

DATA-MINING-CUP
Der DATA-MINING-CUP, jährlicher studentischer Wettbewerb und Anwenderkonferenz.
www.data-mining-cup.de/ - 24k - im Cache - Ähnliche Seiten

Data Mining - Vorlesung, Wintersemester 2005/2006
Data Mining bedeutet Extrahieren von impliziten, noch unbekannt Informationen aus Rohdaten. Dazu sollten Computer in die Lage versetzt werden, ...
www.is.informatik.uni-duisburg.de/courses/dm_ws05/index.html - 12k - im Cache - Ähnliche Seiten

Data Mining
Deshalb finden Techniken aus dem Gebiet des **Data Mining** (auch Wissensentdeckung, Knowledge Discovery) immer stärkere Anwendung in Business, ...
kd.cs.uni-magdeburg.de/wm2002/ws.html - 8k - im Cache - Ähnliche Seiten

Data Mining - 500.000 E-Mails erzählen Geschichte der Enron-Platte
Informatiker haben den spektakulären Zusammenbruch des US-Energiehändlers Enron im Jahr 2001 untersucht. Aus 500.000 E-Mails, die vom Management verschickt ...
www.spiegel.de/netzwelt/technologie/0,1518,360745,00.html - Ähnliche Seiten

WebSeite des DaMIT-Projekts
Data Mining Tutor: Ein generisches Konzept für das Lehren und Lernen im ... Mit DaMIT lernt man die Grundlagen und Anwendungen des **Data Mining** kennen. ...
damit.dtu.dk - 3k - im Cache - Ähnliche Seiten

Data Mining | Digital District

Fetch

Parse

Extract Links & Descriptions
& Normalize

Filter

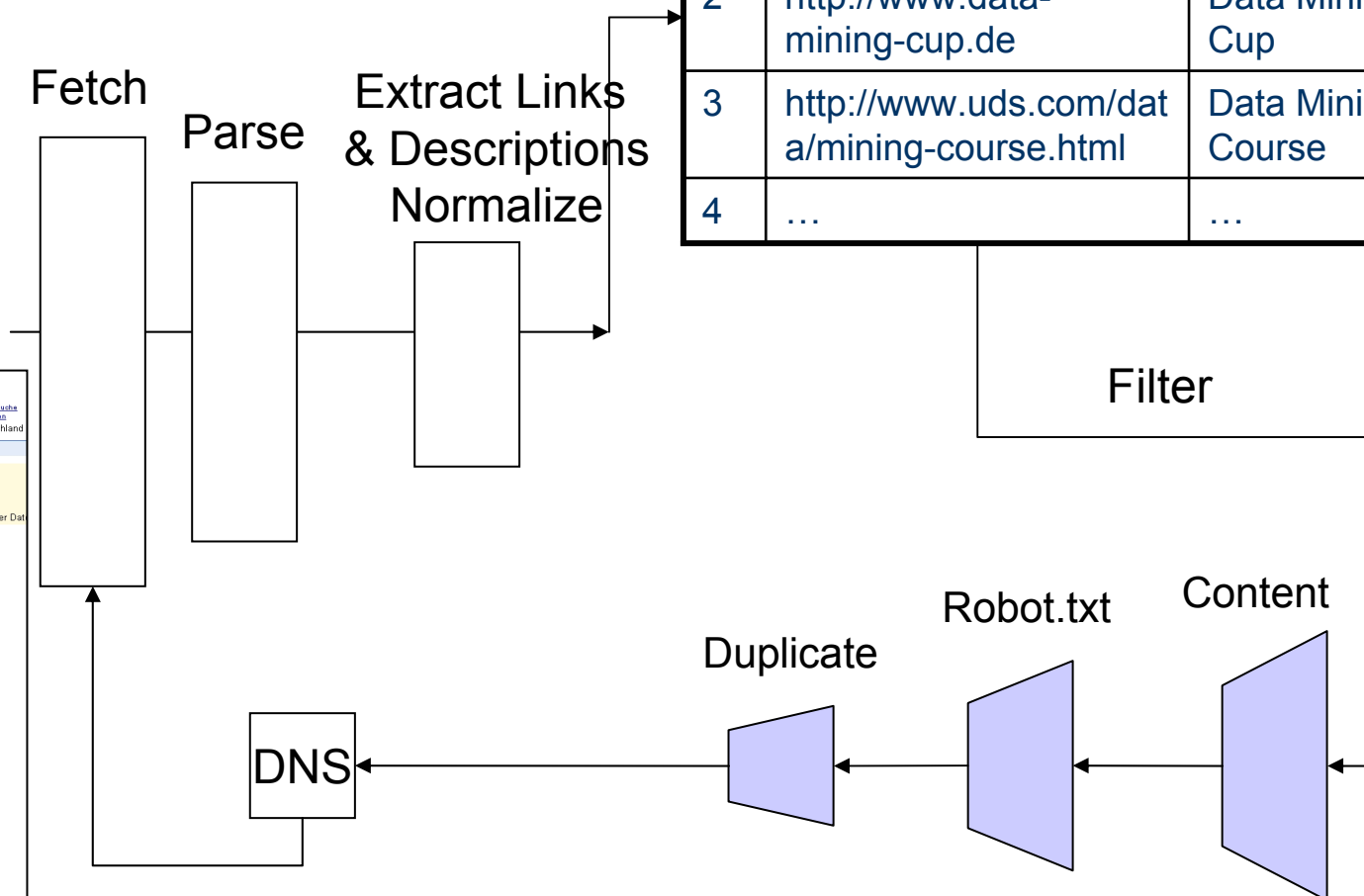
ID	URL	Description
1	http://www.data-mining.de/	Data Mining
2	http://www.data-mining-cup.de	Data Mining Cup
3	http://www.uds.com/data/mining-course.html	Data Mining Course
4

Duplicate

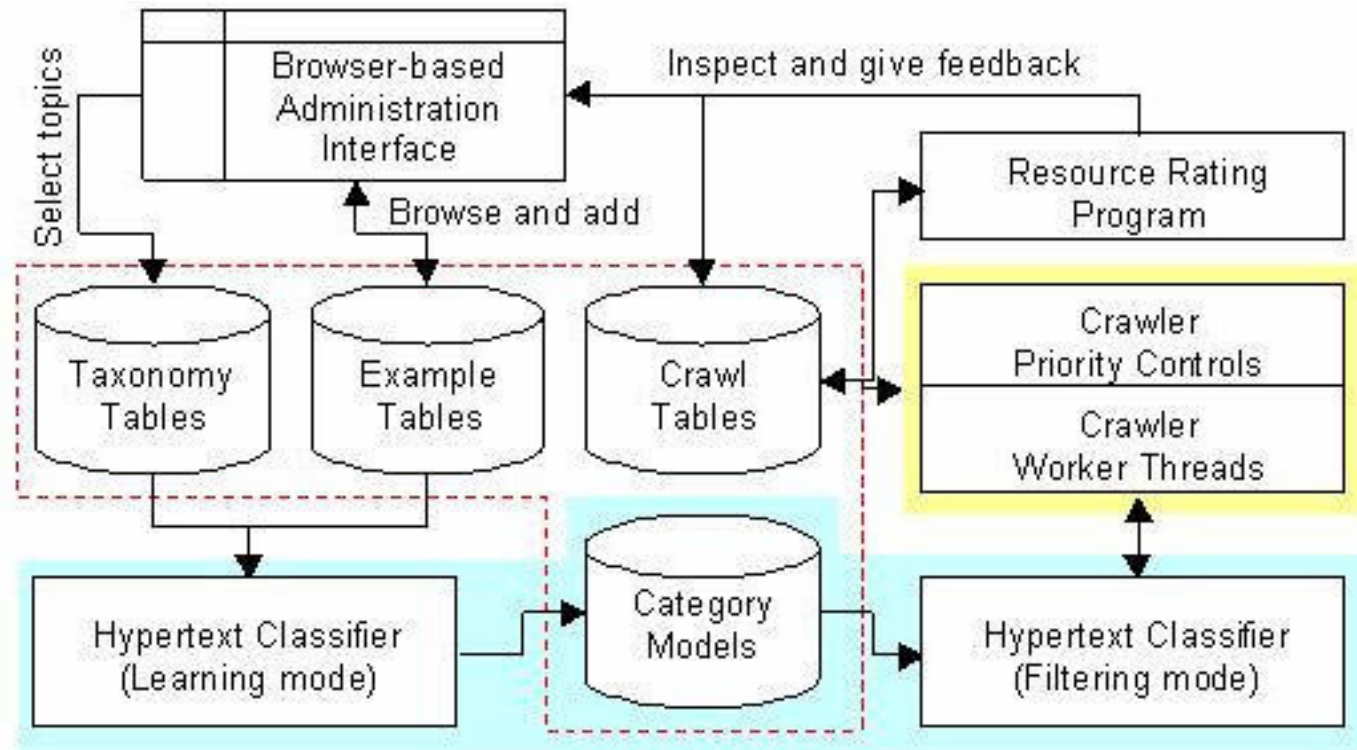
Robot.txt

Content

DNS



Focus Web Crawling



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

- 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 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]

- [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**. Morgan-Kaufmann, 2002, 1-55860-754-4.
- [Kuro04] Dominik Kuroepka: **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

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

"com.cnn.www" →

Contents:	anchor:cnnsi. com	anchor:my.look. ca	Langua ge
<html>...</html> ← t ₁ <html>...</html> ← t ₂	"CNN" ← t ₉	"CNN.com" ← t ₈	en

- Hadoop Distributed File System (HDFS)