Scientific Project: Databases for High-Dimensional Data in conj. with.
Seminar forensische Datenbanken

Martin Schäler, Alexander Grebhahn, Veit Köppen, Prof. Gunter Saake
October 10, 2012
Overview

- Introduction to AG DB (Lectures, Research)
- Short introduction to Digi-Dak
- Concepts
- Overview of project topics & forming project teams
- Course of action (milestones, presentations)
- How to perform literature research?
Overview

▶ Introduction to AG DB (Lectures, Research)
▶ Short introduction to Digi-Dak
▶ Concepts
▶ Overview of project topics & forming project teams
▶ Course of action (milestones, presentations)
▶ How to perform literature research?
▶ Further lectures:
  ▶ Academic writing (2-3 lectures)
Organization
AG Databases: Research Fields

- Database Technologies
  - Integration of Information Systems
  - Tuning/Self-Tuning of DBMS
  - *Multimedia Databases*
  - Tailor-made Data Management
  - *Data Provenance*

- Software Engineering
  - Feature-Oriented Software Development (FOSD)
  - Adaptive Information Systems

- Complex and Embedded Information Systems
  - Embedded Database/Software Systems
  - Interoperability

- Research projects:
  - [http://wwwiti.cs.uni-magdeburg.de/iti_db/forschung/index.php#projekte](http://wwwiti.cs.uni-magdeburg.de/iti_db/forschung/index.php#projekte)
AG Databases: Lectures (I)

Winter Term

Bachelor
- Datenbanken I
- EPMD
- HighDim
- Seminar: Empirical Software Engineering
- Seminar: Forensische Datenbanken

Master
- Transaktionsverwaltung (TV)
- Distributed Data Management (DDM)
- Data Warehouse Technologies (DWT)
- EPMD
- HighDim
- Laborpraktikum: Refactoring HSQL
- Filmseminar
Summer Term

Bachelor
- Spezifikationstechnik
- Databases 2 (implementation techniques)
- Digi-Dak Database Project (in conj. with HighDim)
- Seminare

Master
- Advanced Topics in Databases
- Advanced Database Models
- HighDim
- Student Conference (SCSEDB)

General
- Different Software Projects: http://wwwiti.cs.uni-magdeburg.de/iti_db/study/index.php#praktika
- Bachelor and Master Thesis
Digi-Dak(+) : Introduction and Use Case

- Research project funded by Federal Ministry for Education and Research (BMBF)
- Digi-Dak ⇒ Digital Dactyloscopy
- Research on pattern matching techniques of (digital) fingerprints
- Different fields of research: signal/image processing, IT security/forensics, database systems
- Different scenarios for preventative/forensic processes (e.g. age determination)
- Digi-Dak+ ⇒ Graduate program
- More information: http://omen.cs.uni-magdeburg.de/digi-dak
HighDim: Course Grading

Bachelor

► **Einordnung:** WPF FIN SMK (Schlüssel- und Methodenkompetenzen)

► 5 CP = 150h → 42h presence time (3 SWS) + 108h autonomous work

Master

► **Einordnung:** Wissenschaftliches Teamprojekt

► 6 CP = 180h → 42h presence time (3 SWS) + 138h autonomous work

*Grade at the end of the course for the whole project team*
HighDim: Course Grading II

- Notengebung Wichtung:
  - Vortrag 30%,
  - Implementierung 30%,
  - Ausarbeitung 30%,
  - Weiche Kriterien 10%

- Verbindliche Prüfungsanmeldung: Zweiter Meilenstein
Concepts & Content
Project Topics

1. EDBT Contest: String similarity search (two topics)
2. QuEval Project http://wwwiti.cs.uni-magdeburg.de/iti_db/research/iJudge/index_en.php
   2.1 Integration and Evaluation of real fingerprints of the NIST database
   2.2 Index parallelization in QuEval (CPU)
   2.3 String search extension for QuEval
   2.4 Indexing of encrypted data

More information later in this lecture ...
**Lecture, Meetings & Presentation**

**Lecture & Presentation**

- Time/Place: Wednesday, 3(s.t.) to 5 p.m., G29 - room K059
- Lectures with content of course → all
- Presentation of *main milestones* (see gantt chart) → each project team

**Meetings (Exercise)**

- Individual for each project team
- Time and room to be agreed in project teams!
- Presentation of all intermediate results/milestones (informal)
- Discussion, discussion, discussion . . .
The Idea . . .

Role-playing game... Imagine
You

We
The Idea . . .

Role-playing game . . . Imagine

You

► ...are an upcoming project team
► ...are searching for innovative DB solutions
► Research is your key to success

We
Role-playing game... Imagine

You

- ... are an upcoming project team
- ... are searching for innovative DB solutions
- Research is your key to success

We

- ... are the customer
- ... want the best solution that gives us an unique selling proposition
- ... want you to provide evidence of your scientific capabilities
Acquired skills, specific to research

- Performing literature research
- Understanding and structured review of scientific work
- Autonomous, solution-based reasoning on research task (e.g., finding alternative solutions)
- How to ask? How to adapt a task (extend/reduce)?
- Academic writing
Objectives & Qualification (II)

Acquired skills, always needed

- Team management
- Project and time scheduling
- Presentation of results
- Flexibility regarding changing conditions
- Reasoning about solutions (“Why is this the best/not adequate...”)
Progress of Course

**Deliveries**

- 5 mile stone presentations (*main milestones*)
- Each team member has to present at least once
- Reporting of (sub) milestones in exercises/meetings
- Written paper about literature research (technical report)
- Management report
- Prototypical implementation
Deliveries and Grading (I)

Technical Report

- Delivery of report in sufficient time
- Number of pages
- Quality/Quantity of structure and evaluation
- Quality/Quantity of literature research
- Own contribution
Deliveries and Grading (II)

Management Report

- Description of project realization (timeline, milestones)
- Separation of roles and contributions of single team members
- Meeting protocols
- Self-evaluation of member and group work (strengths, weaknesses)
Deliveries and Grading (III)

Presentation & Discussion

- Quality of scientific presentation (structure, references, time)
- Assessment regarding the content (e.g., results of particular milestones)
- Participation of discussion

Organization

- Strictness
- Communication (just-in-time answers, satisfying time contraints)
- Self-organization (Sharing tasks, internal reporting of current state-of-work, dealing with problems)
- Autonomous working
Topic & Project Teams

- Teams with 3 to 5 students (depends on the task)
- Every task can be chosen once
- **Projects**
  - **Theoretical part**
    - State of the art
    - New ideas
  - **Practical part**
    - Usually JAVA
    - Prototypical implementation
http://www2.informatik.hu-berlin.de/~wandelt/searchjoincompetition2013/

▶ Two tracks - choose one
  ▶ Track I: String Similarity Search
  ▶ Track II: String Similarity Join

▶ Important Dates
  ▶ registration: November 15, 2012
  ▶ release of experimentation data sets: November 16, 2012
  ▶ final submission: January 20, 2013

▶ Data sets
  ▶ DNA sequences
  ▶ Geographical names
Task characteristics

- appr. 3 Students
- Bachelor/Master
- Successful participation means:
  - running submission
  - outperform a simple single threaded seq. scan.
Theoretical part

- Background on String searches/similarity
- Qualitative evaluation
  - Alternative representations
  - Indexing schemes
  - Challenges in contrast to numerical data
  - Domain specific approaches (esp. DNA)
  - Identification of promising approaches
Topic 1 - EDBT Contest

Practical part

- Choose programming language (Java, or C recommended)
- Benchmark your own solution
- Try alternatives ...
QuEval: Evaluation of high-dimensional index structures

- WiSe 2010/11 - First qualitative evaluation and overview
- SoSe 2011 - Quantitative evaluation
- WiSe 2011/12 - Distance metrics
- SoSe 2012 - Impact of data distribution
Task characteristics

- Images and Features (Minutiae)
- 4-5 Students
- Rather Master
- Integration and evaluation
  - Integration (DataTypes, What, etc.)
  - Experiments with different query types
- We have: Uniform (Some) Clustering
Theoretical part

- Digital representations from DB perspective
- Application scenarios (query types)
- What means similar
  - Naive approach
  - Suggestion for improvements
  - Classification/Qualitative analysis of approaches
  - ...
- Identification of important distributions
Practical part

- Integration into framework
- Quantitative evaluation
  - Performance
  - Precision
  - Correlation: Index Structure / Distance metric
  - ...

Topic 2.1 - NIST database
Task characteristics

- 3-5 Students
- Rather Master
- How to benefit from multi-core architectures?
  - Parallelization of well-known indexes
  - Newly developed indexes/own ideas
  - Prototypical implementation
  - Integration into/Interfaces to index framework
  - Evaluation

- We have: Indexes
Topic 2.2 - Index parallelization

Theoretical part

- Discussing state of the art
- Qualitative evaluation
  - Application scenarios
  - Limitations/problems
  - Possible classifications of found indexes
  - Determining evaluation criteria
  - Lesson learned (explorative studies)
Topic 2.2 - Index parallelization

Practical part

- Prototypical implementation
  - Single vs. multi core
  - Determination of sync. overhead (different strategies)
- Evaluation (examples)
  - In comparison to the original index
  - Criteria from theoretical part
  - ...
- Statistical analysis → Finding possible explanations for evaluation results
Task characteristics

▶ 3 Students
▶ Bachelor / Master
▶ What to do with non-numerical data?
  ▶ String indexing
  ▶ Query types and application scenarios
  ▶ Prototypical implementation of (two) selected approaches
  ▶ Evaluation
▶ We have: Framework
▶ Difference to Topic 1: General (not domain-specific) approach.
Topic 2.3 - String search extension for QuEval

Theoretical part

- Discussing state of the art
  - What approaches are known (in databases)?
  - Application field & limitations
  - Possible classifications
  - Discuss evaluation criteria
  - Identification of (possibly) useful approaches for practical part

- Possible definitions of unimportant/dependent
  - Literature
  - Own idea
Practical part

- Prototypical implementation
  - Concept: How to integrate?
  - Reimplementation/existing implementations
  - Interface/integration in framework

- Evaluation (examples)
  - Strings in Java are evil ... so we need alternative representation
  - Criteria from theoretical part
  - ...

- Explain evaluation results
Topic 2.4 - Indexing of encrypted data

Task characteristics

- 3-5 Students
- Rather Master
- Privacy concerns: → Crypted data, but how to index/search
  - What approaches are known?
  - State of the art and if possible classification
  - What approaches are useful for indexing?
  - What happens to the space after encryption (geometric properties)
  - Prototypical implementation/evaluation
- We have: Indexes
Topic 2.4 - Indexing of encrypted data

**Theoretical part**

- Discussing state of the art
  - Literature research → Collect approaches
  - Is there any work at all w.r.t. indexing?
  - Identification of solutions for practical part
- What would be nice to have?
**Practical part**

- Prototypical implementation
  - Highly depends on theoretical part.
  - Convince us that what you do is cool (and important).

- Evaluation (examples)
  - Trivial approach
  - Overhead compared to un-encrypted data
  - ...

- Statistical analysis → Effort estimation.
Task & Time Management

Task Management

- *Main milestones* have to be finished in time
- *(Sub) milestones* are less strict (but don’t be sloppy)
- Pre-defined work packages \(\implies\) each project team
  - ...defines sub work packages
  - ...determines responsibilities for these packages
    (divide&conquer)

Time Management

- Planning of periods
- Regarding capacities and resources
- Considering other tasks and activities
- Reporting of delays immediately to project members!
Role Management

- Possible roles: team leader, design, implementation, testing, writing, ...
- Delegate for important roles/work packages
- Assignment of (sub) tasks to role for each milestone
Individual Criteria: Examples

- Is open and forthcoming in discussions
- Is democratic (i.e. willingly accepts team decisions)
- Is punctual for team meetings
- Is reliable in finishing assigned work
- Is creative and full of ideas in respect to (a) team organization and b) technical approaches
- Communicates clearly
- Contributes significantly to the project (not only in spending time)
- Takes the initiative when appropriate
- Is cooperative
- Is prepared to spend the time necessary to complete his/her tasks
- Is prepared to take responsibility
Literature Research
How to Perform Literature Research

- Efficient literature research requires
  - Knowledge of *Where* to search
  - Knowledge of *How* to search
  - Finding adequate search terms
  - Structured review of papers
  - Knowledge of how to find information in papers
Where to Search (I)

- Different websites available that provide large literature databases

1. Google Scholar: http://scholar.google.de/
   - Key word and concrete paper search
   - Often, PDFs are provided

2. DBLP: http://www.informatik.uni-trier.de/~ley/db/
   - Search for keyword, conferences, journals, author(s)
   - BibTex and references to other websites

3. Citeseer: http://citeseerx.ist.psu.edu/about/site
   - keyword, fulltext, author, and title search
   - BibTex and (partially) PDFs are provided
Where to Search (II)

- Publisher sites are also a suitable target
- **ACM Digital Library**: [http://portal.acm.org/dl.cfm](http://portal.acm.org/dl.cfm)
  - Keyword, author, conference/literature (proceedings), and title search
  - Bibtex, mostly PDFs and other information are provided
  - Similar to ACM, but only few PDFs
  - Extended access within university network
- **Springer**: [http://www.springerlink.de/](http://www.springerlink.de/)
  - Similar to previous
  - Extended access within university Network
- Further search possibilities: on author, research group or university sites
How to Search

*Some hints to not get lost in the jungle*

➤ Use distinct keywords (*fingerprint* vs. *fingerprint data*)
➤ Keep keywords simple (at most three words)
➤ Otherwise, search for *whole* title
➤ Read abstract (and maybe introduction) ⇒ decision for relevance

*First insights*

➤ Read abstract, introduction and background/related work (coarse-grained) to
  ➤ ... get a first idea of the approach
  ➤ ... find other relevant papers
Finding the required information

- Read the paper carefully
- Omit formal parts/sections
- Try to classify (core idea, main characteristics) ⇒ develop classification/evaluation in mind
- Understand the big picture
- Make notes
- Do NOT translate each sentence