Digi-Dak Database Project (DDDP)
Mittwochs, 15-17 Uhr, Raum 29/335

Prof. Gunter Saake, Sandro Schulze, and Martin Schäler

April 5, 2011
Overview

- Introduction to AG DB (Lectures, Research)
- Short introduction to Digi-Dak
- Concepts of DDDP
- 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 of DDDP
- 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

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

- Complex and Embedded Information Systems
  - Embedded and Automotive Systems
  - Interoperability

- Research projects:
  http://wwwiti.cs.uni-magdeburg.de/iti_db/forschung/index.php#projekte
<table>
<thead>
<tr>
<th>Bachelor</th>
<th>Master</th>
</tr>
</thead>
<tbody>
<tr>
<td>◀ Datenbanken I</td>
<td>◀ Transaktionsverwaltung</td>
</tr>
<tr>
<td>◀ EPMD</td>
<td>◀ Distributed Data Management (DDM)</td>
</tr>
<tr>
<td>◀ Digi-Dak Database Project</td>
<td>◀ Data Warehouse Technologies (DWT)</td>
</tr>
<tr>
<td>◀ Seminar: Empirical Software Engineering</td>
<td>◀ Erweiterte Programmierkonzepte für maßgeschneiderte Datenhaltung (EPMD)</td>
</tr>
<tr>
<td>◀ Seminar: Datenqualität</td>
<td>◀ Digi-Dak Database Project</td>
</tr>
<tr>
<td></td>
<td>◀ Filmseminar</td>
</tr>
</tbody>
</table>
AG DB: Lectures (II)

Summer Term

Bachelor

- Spezifikationstechnik
- Databases 2 (implementation techniques)
- Digi-Dak Database Project
- Seminar: DM in eingebetteten Systemen

Master

- Advanced Topics in Databases
- Advanced Database Models
- Digi-Dak Database Project
- Student Conference (SCSEDB)

General

- Different Software Projects:
  http://wwwiti.cs.uni-magdeburg.de/iti_db/study/index.php#praktika
- Bachelor and Master Thesis

Prof. Gunter Saake
Digi-Dak: Introduction

- 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):
  - Aging detection
  - Fingerprint overlapping
- More information:
  http://omen.cs.uni-magdeburg.de/digi-dak
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*
Concepts & Content
1. Kickoff project: Secure database infrastructure
2. Follow-up project: High-dimensional index structures
   2.1 Cost models
   2.2 Distance metrics
   2.3 Implementation & evaluation

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

*Lecture & Presentation*

- Time/Place: Wednesdays, 3 to 5 p.m., G29 - room 335
- 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
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
Objectives & Qualification (I)

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
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
Grading Criterias (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
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)
**Grading Criteria (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 constraints)
- Self-organization (Sharing tasks, internal reporting of current state-of-work, dealing with problems)
- Autonomous working
Teams with 3 to 6 students

Every task can be chosen once

Projects

- Theoretical part
  - State of the art
  - New ideas

- Practical part
  - DBMS PostgreSQL 8.4 / JAVA
  - Prototypical implementation
Topic 1 - Secure Data Base Infrastructure

Prevent unwanted Access to the DB

- Focus: theoretical part
- Theoretical part
  - State of the art
  - New ideas
- Practical part
  - DBMS PostgreSQL 8.4
  - Show basic idea
  - Security test framework
Theoretical part

▶ Which security problems do exist? (Classification: Attacks,Leaks, ...)
▶ Threat model
  ▶ Where?
  ▶ Who?
▶ Making the DB secure
  ▶ Generally
  ▶ Especially in PostgreSQL 8.4
▶ Evaluation of Postgre’s security mechanism
▶ Literature
  ▶ Handbook of DB Security
  ▶ PostgreSQL Documentation
Topic 1 - Secure Data Base Infrastructure

Practical part

- Demonstrator
  - PostgreSQL 8.4
  - Client server communication
- Security test framework
  - Collection of Attacks
- Intra team competition
  - E.g. sub-teams with two students
  - Team A(ttacker) tries to get access to the DB controlled by team B
  - Switch the roles after each successful attack
  - be creative ...
Scenario

- DB with large number of images
- Task: Efficient search
- Simplification
  - All images: 8bit gray values, same resolution, size, ...
  - Image = Feature vector
  - Each pixel is a dimension
  - No dimension reduction
  - No normalization (8bit gray values!)
- Query profile
  - Exact match
  - Nearest neighbors
Follow-up project: DB Tuning in High-dimensional data sets
▶ Last terms task: How to find all “similar” Data Sets as fast as possible?
  ▶ Problem: Dimensional curse
  ▶ Exact match
  ▶ K-nearest neighbor search
▶ Last group’s contributions
  ▶ Overview & evaluation of promising structures
  ▶ Test framework for evaluation of index structures (JAVA) & Data generation
  ▶ Implementation & evaluation of
    ▶ Sequential search
    ▶ Kd-tree
    ▶ VA-File
    ▶ Variant of locality sensitive hashing
Current limitations of the framework

- Measuring time is imprecise!
  - JVM, Op. System, Hardware, etc.
  - → Topic 2.1: Cost Models

- What is equal?
  - Different distance metrics might produce different results
  - → Topic 2.2: Distance metrics

- We need more index structures!
  - There are tons of structures (e.g., Trees)
  - → Topic 2.3: Integration & evaluation of new structures
How to compare different index structure (implementations)?

- Focus theoretical part: What exists and implementation concept
- Theoretical part
  - State of the art
  - New ideas
- Practical part
  - Java Framework
  - Integration concept
  - Implementation for selected index structures
  - Testing and evaluation
Topic 2.1 - Cost models

Theoretical part

- What cost models exist?
- Categorization and evaluation
  - General purpose?
  - Are there any frameworks for this purpose?
  - What factors taken into account (e.g., Main Memory Access vs. Hard disk access)?
- Quality criteria
- Implementation of cost models → ideas for implementation part
- Identification of promising models (implementation effort vs. speed vs. quality)
Practical part

▶ Integration of cost models for selected structures and selected or own models
▶ Choose at least three structures (sequential scan recommended)
▶ Integration
▶ Evaluation of cost model vs. time measurement
Topic 2.2 - Distance metrics

What is similar?

- Focus practical part: Testing
- We have: Euclidean and Manhattan
- Theoretical part
  - State of the art
  - Own ideas
  - Evaluation
- Practical part
  - Java Framework
  - Implementation
  - Testing and evaluation
Topic 2.1 - Cost models

Theoretical part

- What metrics exist?
- Categorization and evaluation
  - Scope of a metric
  - Visualization of the metric
  - Is there anything like quality criteria?
  - Categorization & evaluations
  - Identification of promising metrics
Topic 2.2 - Cost models

Practical part

- Integration into the frame work
- Testing
  - Comparing of results for same structures (Precision Recall)
  - Performance
  - Correlation: Index Structure / Distance metric
Intra and inter structure differences

- Focus: practical part - Implementation/Experiments
- Theoretical part
  - Short general information: Dimensional curse
  - Intra structure differences (implementation efficiency esp. trees)
  - Locality sensitive hashing (esp. How to find an LSH-function)
- Implementation part
  - Extending the framework
  - Experiments
**Theoretical part**

- Basics (e.g., dimensional curse)
- Intra index differences (esp. trees)
  - Implementation details
  - Tailor-made indexes (e.g., data types, domain specific)
- Locality sensitive hashing approaches
  - How to find a suitable LSH function
  - Circumvention of LSH-function problem
  - Combination of hash and tree approaches
Topic 2.3 - Implementation & evaluation of indexes

Practical part

► Integration into the frame work
► Testing
  ► Evaluation of different implementations
  ► Experiments (different data size and dimensions)
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 $\Rightarrow$ 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