Distributed Information Systems
Lecture WS 2006/2007

SSC and SI Masters
Orientation Internet Computing

Tuesdays, 8am-11am, INM200
Lecture: Tuesday 8-10 Room INM 200
Exercise: Tuesday 10-11 Room INM 200
Karl Aberer & Philippe Cudré-Mauroux
Distributed Information Systems Laboratory
Goals of the Course

• What is a "Distributed Information System"?
  – e.g. Web search engines, Web data management, mobile data management etc.

• Which are key problems studied for DIS?
  – e.g. efficient search, abstraction and modeling, optimization of resource usage etc.

• What are typical techniques used to solve these problems?
  – e.g. XML storage and querying, vector space retrieval, association rule mining etc.

• How to apply these techniques?
Focus of the Course

• **Models and Algorithms** for representing, storing and processing information
  – systems aspect covered in “Conception of Information Systems” resp. in the future course "Middleware"

• **Models and algorithms** for representing, storing and processing information on the **Web**
  – relational model covered in “Relational Databases” and "Introduction to Information Systems"
Have you ever heard of...

• PageRank?
  – Google?

• OWL?
  – RDF?
    • Creative Commons?
      – XML?

• Association Rule Mining?
  – Amazon recommendations?

• DHT?
  – Azureus?
Two Recent Examples...

• **March 2005: Yahoo buys Flickr**
  – Information Retrieval
  – Information Ranking (Flickr Interestingness)
  – XML+RDF
    • 22 million photos legally shared under flexible copyrights licensed through Creative Commons

• **September 2005: Ebay buys Skype for $2.6 billion**
  – Distributed Storage
  – Peer-to-Peer Overlay Network
The Course - Lecture

• Lecture
  – standard ex cathedra lecture
  – but feel free to interrupt, ask questions ...
    ... even if it is early in the morning and cold outside
  – I also will ask questions once in a while ...

• Question
  – paper copies of course notes?
The Course - Exercises

• Exercises
  – applying the algorithms covered in the lectures
  – estimated effort: 4 hours per exercise, 9 exercises in total
  – we will provide
    • Questions + datasets (input/output) for selected algorithmic problems
    • The following week: Solutions + implementation in Matlab and Java
  – solutions presented and discussed during exercise hour
  – your exercises will not be corrected nor graded

• Questions
  – availability of notebook/computer?
  – previous experience in programming: Matlab? Java?
The Exam

- Two midterm exams and one final exam (written)
  - midterms contribute 25% each to final grade, if improvement

- Conceptual questions and practical problems
  - will assume you attended the lecture
  - will assume you did the programming exercises
  - examples from earlier years (exercises, exams) provided for preparation

- Support: Lecture Slides + Exercises + Handwritten Notes
## Time Schedule (indicative)

### Lecture Schedule

<table>
<thead>
<tr>
<th></th>
<th>Lectures</th>
<th>Programming exercises and exams</th>
<th>Assistant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td>- Organizational Info</td>
<td>- Recap XML</td>
<td></td>
</tr>
<tr>
<td>2006.10.24</td>
<td>- Overview of the lecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Semi-structured Data Management</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006.10.31</td>
<td>- XML Storage and Filtering</td>
<td>- XML Document Filtering</td>
<td>Adriana</td>
</tr>
<tr>
<td>2006.11.07</td>
<td>- Graph Databases</td>
<td>- Data guide construction algorithm</td>
<td>Adriana</td>
</tr>
<tr>
<td>2006.11.14</td>
<td>- Semantic Web</td>
<td>- Solution of previous exercise only</td>
<td></td>
</tr>
<tr>
<td>2006.11.21</td>
<td>- Resource Description Framework (RDF)</td>
<td>- RDF graphs, RDF schemas</td>
<td>Adriana</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Distributed Data Management</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006.11.28</td>
<td>- Schema Fragmentation</td>
<td>- 1st intermediate exam (Semi-structured Data Management)</td>
<td>Ali</td>
</tr>
<tr>
<td>2006.12.05</td>
<td>- Data Broadcasting in Mobile Networks</td>
<td>- Implementation of the bond energy algorithm</td>
<td></td>
</tr>
<tr>
<td>2006.12.12</td>
<td>- Peer-2-Peer Systems</td>
<td>- Broadcast disks</td>
<td>Ali</td>
</tr>
<tr>
<td>2006.12.19</td>
<td>- Small World Graphs</td>
<td>- Simulating structured and unstructured P2P networks</td>
<td>Roman</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Information Retrieval and Data Mining</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007.01.09</td>
<td>- Information Retrieval (part 1)</td>
<td>- 2nd intermediate exam (Distributed data management)</td>
<td>Wojciech</td>
</tr>
<tr>
<td>2007.01.16</td>
<td>- Information Retrieval (Vector Space Model, LSI)</td>
<td>- TF/IDF ranking</td>
<td>Wojciech</td>
</tr>
<tr>
<td>2007.01.23</td>
<td>- Information Retrieval (Link-based Ranking)</td>
<td>- Implementation of LSI</td>
<td>Wojciech</td>
</tr>
<tr>
<td>2007.01.30</td>
<td>- Data Mining (Association rules)</td>
<td>- Apriori algorithm</td>
<td>Wojciech</td>
</tr>
<tr>
<td>2007.02.06</td>
<td>- Data Mining (Clustering, Classification)</td>
<td>- Solution of previous exercise only</td>
<td>Wojciech</td>
</tr>
<tr>
<td>2007.03.??</td>
<td>- Final exam (covers the entire lecture)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

©2006/7, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis
Organizational Info

• Lecture Team
  – Lectures
    • Karl Aberer, karl.aberer@epfl.ch, BC 108
    • Philippe Cudré-Mauroux, philippe.cudre-mauroux@epfl.ch, BC 114
  – Exercises
    • Adriana Budura, adriana.budura@epfl.ch, BC 130, Thu. 14-15h
    • Ali Salehi, ali.salehi@epfl.ch, 693 6656, BC 143, Fri. 14-15h
    • Roman Schmidt, roman.schmidt@epfl.ch, BC 118, Wed. 14-15h
    • Wojciech Galuba, wojciech.galuba@epfl.ch, BC 143, Fri. 14-15h

• Web site
  – http://lsirwww.epfl.ch (menu item Students)
References

• Parts of the course are based on the following text books
  – Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.

• References to the research literature will be given during the lecture