Distributed Information Systems
Lecture WS 2003/04

SSC, mandatory, orientation IS, Sem 9
SI optional

Time and Place
Lecture: Tuesday 8-10 Room INM 200
Exercise: Tuesday 10-11 Room INM 200

Karl Aberer
Distributed Information Systems Laboratory

Organizational Info

- Lecture Team
  - Karl Aberer (Lecture),
    - karl.aberer@epfl.ch, 693 4679, PSE 1.32
  - Philippe Cudre-Mauroux (Exercises)
    - philippe.cudre-mauroux@epfl.ch, 693 6787, PSE A 1.51
  - Anwitaman Datta (Exercises)
    - anwitaman.datta@epfl.ch, 693 6615, PSE A 1.62

- Web site
  - Found at http://lsirwww.epfl.ch (menu item Courses)
  - http://lsirwww.epfl.ch/courses/dis/2003ws/index.html

Goal

- Four levels of understanding: be able to
  1. *understand what* a distributed information system is
     - e.g. Web search engines, Web data management, mobile data management etc.
  2. *identify which key problems* are to be solved for realizing a distributed information system
     - e.g. indexing structures, optimization algorithms, abstraction algorithms etc.
  3. *describe selected key techniques* use to solve these problems
     - e.g. XML storage and querying, vector space retrieval, association rule mining etc.
  4. *apply* these techniques in simple cases

- Focus
  - Models and Algorithms for representing, storing and processing information
    - system aspects covered in "Conception of Information Systems"
  - Models for the Web
    - relational model covered in "Relational Databases"

Approach

- Lecture
  - introduce framework in week 1
  - specialized themes in the following weeks
  - includes conceptual questions to be answered on-line (feedback)
- Exercises
  - Practical examples to apply methods (learning by doing)
  - Exercises will be corrected
  - Total exercise grade is up to 20% of final grade
    (each of the 10 exercises contributes 2%)
    and only used to lift the exam grade
  - Solutions presented and discussed in exercise hour
- Exam
  - Mix of conceptual questions from lecture and practical examples from exercises
  - Written
  - Support: Lecture Slides + Exercises + Handwritten Notes

Time Schedule (indicative)

Date	Lecture
Introduction	
21.10.2002	Introduction to DIS / XML Intro
Semi-structured Data	
28.10.2002	XML Storage and Filtering
04.11.2002	Graph databases
11.11.2002	RDF and Semantic Web/OIL
Distributed Data Management	
18.11.2002	Schema Fragmentation
25.11.2002	Mobile Data Management
02.12.2002	P2P Systems
09.12.2002	P2P Systems
Information Retrieval and Data Mining	
16.12.2002	Vector Space Retrieval
06.01.2003	Special Session (announcement follows)
13.01.2003	Advanced Retrieval
20.01.2003	Text Indexing
27.01.2003	Association Rule Mining
03.02.2003	Classification

References

- Parts of the course are based on the following text books
  - M. Tamer Özsu, Patrick Valduriez: Principles of Distributed Database Systems, Second Edition, Prentice Hall, 1999.
  - S. Abiteboul, P. Bunemann, D. Suciu: Data on the Web: From Relations to Semistructured Data and XML, Morgan Kaufman, 2000.
  - Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
  - Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
  - P. Baldi, P. Frasconi, P. Smyth: Modeling the Internet and the Web, Wiley 2003.

References

- Some parts on more recent issues are based on research literature
  - Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)
  - Ling Liu, L. Yan, and M.T. Ozsu. "Interoperability in Large-Scale Distributed Information Delivery Systems," In Advances in Workflow Systems and Interoperability, A. Dogac, L. Kalinichenko, M.T. Özsu and A. Sheth (eds.), Springer-Verlag, 1998.
  - Daniel Barbará: Mobile Computing and Databases - A Survey. TKDE 11(1): 108-117 (1999)
  - Swarup Acharya, Rafael Alonso, Michael J. Franklin, Stanley B. Zdonik: Broadcast Disks: Data Management for Asymmetric Communications Environments. SIGMOD Conference 1995: 199-210
  - Sohail Hameed, Nitin H. Vaidya: Log-Time Algorithms for Scheduling Single and Multiple Channel Data Broadcast. MOBICOM 1997: 90-99
  - Tomasz Imielinski, S. Viswanathan, B. R. Badrinath: Data on Air: Organization and Access. TKDE 9(3): 353-372 (1997)
  - Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, Hari Balakrishnan. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications. Proceedings of the ACM SIGCOMM, 2001.
  - Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, Scott Shenker. A Scalable Content-Addressable Network. Proceedings of the ACM SIGCOMM, 2001.
  - M.A. Jovanovic, F.S. Annexstein, and K.A.Berman. Scalability Issues in Large Peer-to-Peer Networks - A Case Study of Gnutella. University of Cincinnati, Laboratory for Networks and Applied Graph Theory, 2001. http://www.ececs.uc.edu/~mjovanov/Research/paper.ps

References

- Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: A Distributed Anonymous Information Storage and Retrieval System. Designing Privacy Enhancing Technologies: International Workshop on Design Issues in Anonymity and Unobservability. LLNCS 2009. Springer Verlag 2001. http://www.freenetproject.org/index.php?page=icsi-revised
- Karl Aberer. P-Grid:A self-organizing access structure for P2P information systems. Proceedings of the Sixth International Conference on Cooperative Information Systems (CoopIS 2001), 2001.