

750762, Data Mining and Data Warehousing

3 hours per week, 3 credit hours, prerequisite: **none**

Teaching Method: 37 hours Lectures (2-3 hours per week), 8 hours Seminars (1 per 2 weeks)

Aims: The data warehousing part of the module aims to give students a good overview of the ideas and the techniques, which are behind recent developments in the data warehousing and OnLine Analytical Processing (OLAP) fields, in terms of data models, query languages, conceptual design methodologies, and storage techniques. Laboratory sessions will ground the abstract notions on practical cases and tools.

The data mining part of the module aims to motivate, define and characterize data mining as a process; to motivate, define and characterize data mining applications; to survey, and present in some

detail, a small range of representative data mining techniques and tools. Laboratory sessions will ground the abstract notions on practical cases and tools.

Learning Outcomes:

On completion of this module, the student should be able to:

- Understand the techniques behind the recent development in data warehousing and data mining.
- Understand query languages and conceptual design methodologies.
- Practice on different tools of data warehousing and data mining.
- Design small projects with data mining and data warehousing.

Textbooks and Supporting Materials:

1. M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis (ed.), Fundamentals of Data Warehouses, Springer-Verlag, 1999.
2. Ralph Kimball, The Data Warehouse Toolkit, Wiley 1996.
3. I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufman, 1999. (This is the one that lectures notes are most closely based on.)

4. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufman, 2000. (This is more database-centred, in contrast to Witten and Frank, who takes a machine-learning viewpoint of data mining. It is also useful in covering data warehouses too, to some extent.)

5. D. Hand, H. Mannila and P. Smyth. *Principles of Data Mining*, MIT Press, 2001. (This takes yet another viewpoint on data mining, viz., the statistical one. In this sense, it is the least related to the approach followed in this part of the course.)

6. M. H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2003. (This has yet another slight shift in emphasis, as it more or less favours an algorithmic viewpoint and is, in this sense, a core computer-science view of the issues.)

Research papers

Website(s): <http://www.cs.man.ac.uk/~sattler/teaching/cs636.html>

Synopsis:

26

Part 1. Data Warehousing:

- Introduction to Data Warehousing: Heterogeneous information; the integration problem; the Warehouse Architecture; Data Warehousing; Warehouse DBMS.
- Aggregations: SQL and aggregations; aggregation functions; grouping.
- Data Warehouse Models and OLAP Operations: Decision support; Data Marts; OLAP vs OLTP; the Multi-Dimensional data model; Dimensional Modelling; ROLAP vs MOLAP; Star and snowflake schemas; the MOLAP cube; roll-up, slicing, and pivoting.
- Some Issues in Data Warehouse Design: monitoring; wrappers; integration; data cleaning; data loading; materialised views; warehouse maintenance; OLAP servers; metadata.

Part II. Data mining:

- Introducing Data Mining: Why data mining?; What is data mining?; A View of the KDD Process; Problems and Techniques; Data Mining Applications; Prospects for the Technology.
- The CRISP-DM Methodology: Approach; Objectives; Documents; Structure; Binding to Contexts;

Phases, Task, Outputs.

- Data Mining Inputs and Outputs: Concepts, Instances, Attributes; Kinds of Learning; Providing Examples; Kinds of Attributes; Preparing Inputs. Knowledge Representations; Decision Tables and Decision Trees; Classification Rules; Association Rules; Regression Trees and Model Trees; Instance-Level Representations.

- Data Mining Algorithms: One-R; Naïve Bayes Classifier; Decision Trees; Decision Rules; Association Rules; Regression; K-Nearest Neighbour Classifiers.

- Evaluating Data Mining Results: Issues in Evaluation; Training and Testing Principles; Error Measures, Holdout, Cross Validation; Comparing Algorithms; Taking Costs into Account; Trade-Offs in the Confusion Matrix.

Assessment: Two 1-hour midterm exams (15% each); Assignments (10%); Seminars (10%); 2-hours

Final Exam (50%)