This course is designed to cover the mathematical foundations of databases. The main focus is to study the design and semantics of search and query languages across a spectrum of database models. This spectrum includes the relational db model, complex and object-oriented db models, and semi-structured dbs models. The latter models include XML and RDF, which respectively model databases as labelled trees and labelled graphs. Time permitting, we will cover some topics in data mining, machine learning, information retrieval, cloud computing, etc. as they relate to search in large data sets. (Note: this course is not a systems course.)


Online version

http://webdam.inria.fr/Alice/

In addition to the textbook, papers will be read. References to these papers will be put under Resources. Lecture notes will also be added under Resources.

Topics

1. Database models

   (a) The relational database model (Chapter 3).

   (b) The complex object database model (Chapter 20).

   (c) The object-oriented database model (Chapter 21).

   (d) Semi-structured databases models:

      i. XML

      ii. The semantic web
iii. Probabilistic database models.

2. **Constraints** (Chapters 8 and 9).

3. **Queries and query languages**
   
   (a) Definition of query (Chapter 16).
   
   i. Generic queries (Chapter 16).
   
   ii. Computable queries (Chapter 16).
   
   iii. Query complexity (Chapter 16).
   
   (b) Semantic restrictions on queries.
   
   i. Safety (Chapter 5)
   
   ii. Domain independence (Chapter 5).
   
   iii. Monotonicity (Chapter 4).
   
   (c) Query languages.
   
   i. First-order query languages.
   
   A. The relational algebra (Chapters 4 & 5).
   
   B. The relational calculus (Chapters 4 & 5).
   
   C. Equivalence of the relational algebra and the safe relational calculus (Chapter 5).
   
   D. Expressiveness of first-order query languages (Chapter 17).
   
   ii. Extensions of first-order query languages.
   
   A. Query languages with generalized quantifiers and aggregate functions.
   
   B. Datalog (Chapters 12 and 13).
   
   C. The fixpoint query languages (Chapters 14 and 17).
   
   iii. Complex object query languages (Chapter 20).
   
   iv. Query languages for object-oriented databases (Chapter 21).
   
   v. Probabilistic query languages.

4. **Current topics**
   
   (a) Meta-data, meta-data queries and query languages.
   
   (b) Introduction to Information Retrieval.
(c) Some topics in data mining and machine learning.
(d) Cloud computing.
(e) Entity and relationship extraction from text and web pages.
(f) Data and workflow provenance.
**Evaluation criteria**

There will be a midterm exam and a final exam, each worth 25% of your grade.

There will be a 10-15 pages midterm paper worth 15% of your grade.

There will be a 15-20 final paper worth 25% of your grade.

I will assign homework problems which you are strongly encouraged to solve. These homeworks will be graded and are essential for understanding the material and for the preparation of the exams. The combined total of the homeworks is worth 10% of your grade.