Summary of data mining course, Autumn 2002

- Project reports due: Fri 25 Oct 2002
- Course exam: Fri 1 Nov 2002 at 14.00, Auditorium
- Required:
  - project work (at least 10/20 points)
  - course exam (at least 20/40 points)
  - (weekly exercises: up to 10 extra points)

Course exam 1 Nov 2002

Examined material:

- Course text “Knowledge discovery in databases: the search for frequent patterns”
- The 4 original articles used as course text (about closed sets and FP-tree)
- Course slides (slides to the main text probably do not carry additional information if compared with the text; slides to closed sets and FP-tree probably do)

Later exams

- Exams planned for January and April 2003
- Either as an extra course exam
  - requirements exactly as for the course (previous slides)
  - note: possible only if the project work was completed in time!
- Or as a separate exam
  - the examined material will be about 50% larger than in the course exam (all the course material plus several extra articles, TBA)
  - the grade is based on the exam only
  - project work not needed, no extra points available for project work or exercises

Overview of course contents

Pattern types

- association rules
- frequent sets
- episodes, parallel and serial
- the generic “frequent pattern” discovery task
- (+ closed frequent sets, frequent generators)

definitions, uses, properties...

Overview of course contents

Various concepts

- candidate collection
- selection criterion + monotone specialization relation
- border (positive and negative one)
- closure, closed set, generator
- episodes: windows, minimal occurrences
- hypergraph transversals

definitions, uses, properties...

Overview of course contents

Algorithms

- generic levelwise search
- Apriori
- episode discovery
- Apriori modified for frequent generators (A-Close)
- candidate generation
- guess-and-correct
- sampling for frequent sets
- FP-tree construction, FP-growth, CLOSET
- association rule generation

key ideas and properties, intuition

Overview of course contents

Results, e.g.:

- complexity of finding frequent patterns
- infrequent candidates = negative border
- negative border has to be checked
- properties of closed sets, generators, ...
- ...

Misc

- knowledge discovery process