TOPICS IN INFORMATION PROCESSING (1997-98)

DATA MINING

1) Introduction
2) KDD and Data Mining
3) Data Mining and Data Querying
4) Data warehousing
5) Machine learning
6) Suggested Seminar Titles

Frans Coenen

Room G33, email `frans`

INFORMATION AS A FACTOR OF PRODUCTION

- Increase in available data
- exacerbated by WWW
- Information overload
- Computer assistance to filter, select and interpret data
- Extend this to allow computers to discover relevant information
- In the future machine assistance will become more and more important

MOTIVATION FOR DATA MINING

- Organisation wish to gain some advantages over competitors
- Organisations have large amounts of data stored in Data Bases (DB) - "can this be used to gain the desired advantage?"
- Hence *Knowledge discovery in databases* (KDD) or data mining
- There is a corollary between the mechanical production of data and the mechanical consumption of data

KDD AND DATA MINING

- KDD - The process of extraction of knowledge (relationships and patterns) from data:

  "The non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data"

- Data mining - The discovery of relations and patterns (knowledge) between data elements
- Data mining is thus an aspect of KDD
- KDD in turn thus encompasses aspects of machine learning, statistics, data base technology, expert systems and data visualisation

CURRENT POPULARITY OF DATA MINING

- All major organisations have well established DB which continue to grow
- All such organisation wish to obtain a "commercial advantage"
- Use of networks raises the possibility of *distributed* data mining (data mining can be implemented across a number of databases)
- Research in machine-learning techniques has made great advances
- In short KDD/data mining is now a viable proposition (this was not the case a few years ago)

KDD AND SQL

- Query tools such as SQL are used to find data under known constraints - i.e. they provide a "shallow" query facility
- Data mining algorithms find "optimal" clusterings of interesting regularities in data - i.e. they automatically "zoom" in on interesting sub-parts of data
- Data mining and query tools are complimentary - one does nor replace the other:

"If you know what you are looking for use a query tool"

"if you have only a vague notion of what you are looking for use a data mining tool"

DATA WAREHOUSING

- A prerequisite to data mining is a data "mine" or *warehouse*. This may be defined as:

"A subject-oriented, integrated, time-variant, non-volatile collection of data in support of management"

1) **subject-oriented**: Built for application decision support (not all data on operational DB will be of use)
2) **integrated**: e.g. Similar and common fields combined
3) **time-variant**: Information collected over time - series of "snap shots" of operational DB
4) **non-volatile**: Data in warehouse is static, i.e. it is historical data which is never updated

META-DATA

- To mine a data warehouse we need information concerning the structure and contents of that warehouse.
- This data is referred to as *meta data*, and is usually stored in a second DB.

DATA MARTS

- It is sometimes convenient to consider a particular part of a data warehouse in isolation with respect to some sub-application
- In this case it is more computationally effective to work with a subset of the data contained in the warehouse and store this in a separate DB - the latter is then referred to as a *data mart*.

WHAT IS LEARNING

- Given:
  1) A task to be carried out (well or badly)
  2) A subject which must be "learnt" in order to carry out this task

Learning can be defined as followed:

"An individual **learns** *how to carry out a certain* **task** *by making a transition from a situation where the task cannot be undertaken to one where the task can be undertaken (under the same circumstances)"*

MACHINE LEARNING

- The capacity to learn is an essential characteristic of life
- The ultimate goal of computer science is to create a machine that can "think", i.e. learn
- The computer science field of Artificial Intelligence (AI) has been focused on machines that can learn since the earliest days of computer technology (c1950s)
- New generation of work (1980s): self organising tress, neural networks, genetic algorithms, Etc.
- Also more powerful computers

MACHINE LEARNING AND EXPERT SYSTEMS

- An expert system is a computer program designed to operate in a manner akin to an "expert" in a given domain
- Originated at the end of 1970s
- Expert "know-how" represented as a set of `if ... then ...` rules gleaned from domain experts
- Simply a problem of finding the right rules - the "knowledge acquisition bottle-neck"
- Machine learning?

SELF-LEARNING COMPUTER SYSTEMS

"A self-learning computer system can generate programs by itself, enabling it to carry out new tasks."

- Computers greatest advantages are (1) speed and (2) accuracy
- Computers greatest disadvantage is its lack of credibility

Can the advantages offered by the computer be harnessed to achieve machine learning?

THE METHODOLOGY OF SCIENCE

- The *empirical cycle*:

  1) Observation

  2) Analyse

  3) Theory

  4) Prediction

"A general law can never be verified by a finite number of observations. It can, however, be falsified by only one observation"
(Karl Popper).

SUGGESTED SEMINAR TITLES

1. The KDD knowledge discovery process - what are the steps involved?

2. Data miming and the data warehouse

3. Data mining and knowledge discovery - what techniques are available?

4. Setting up a KDD environment