presenting data mining

Aris Gionis
gionis@cs.helsinki.fi
Exactum A340

Helsinki Institute for Information Technology
Basic Research Unit
www.cs.helsinki.fi/hiit_bru

first year computer science students?

that's awesome!

…starting a journey in one of the most fascinating sciences

why fascinating?

- huge impact
- extremely fast evolving
- diverse areas, diverse tools (many still unshaped)

1. impact

revolutionize the world over last two decades

- any-time any-place communication
- information on the tips of our fingers (vast amounts)
- intelligent systems in our service
- impact in all other sciences
  (data collection, data analysis, computational power)
- experience inapproachable environments, entertainment, etc.

2. computer science is evolving very fast

- looking 20 years back seems like "prehistory"
- students' ideas 5 years ago are today's standards
- nobody know what they will work 10 years from now
- extremely active communities
  conferences, workshops
  mobility of people
  cross fertilization of ideas
- a lot of energy and feel of discovering new things

The best way to predict the future is to invent it

Alan Kay
2003 Turing award recipient

3. diverse areas, diverse visions, diverse abilities, diverse tools

- systems
  emphasis on how computer systems work
- theory
  emphasis on studying in-depth limits of computing
- information processing
  emphasis on how to model knowledge and analyse information

3. diverse areas, diverse visions, diverse abilities, diverse tools

systems:

vision: improve the way that computer systems work

specializations: netwoks, distributed systems, software engineering, reliability

abilities: programming skills, creativity,
good engineering design

3. diverse areas, diverse visions, diverse abilities, diverse tools

theory:

vision: understanding in-depth limits of computing

specializations: algorithms, complexity, security, cryptography, quantum computing

abilities: mathematical skills,
discrete and combinatorial math

3. diverse areas, diverse visions, diverse abilities, diverse tools

information processing:

vision: make computers look intelligent
modeling of physical world
representation of knowledge
inference

specializations: data mining, machine learning, intelligent systems

abilities/tools: probability, statistics, algorithms

data mining

vision: find patterns in large collections of data

(also replace patterns with: knowledge, structure, rules, etc)

Data often in too large amounts

- data collected in sciences
- biology (human genome has 3 billion base pairs)
- web (more than 4 billion pages)
- other large text collections
- stock market, customer transactions, industry

…

so, why is it difficult?

efficiency:

searching for patterns can slow down the computer a lot
(too many possible patterns to search for all)

semantics:

what are the right patterns to search for?

example 1

Course/student data set

	C++	Java	Boolean logic	Data-bases	Data-mining	…
Anne P.	1	1	0	1	1	
Heikki M.	0	0	1	1	1	
Jouni S.	1	0	1	0	0	
Kari L.	1	1	1	1	0	
Taneli M.	0	0	0	1	1	
…						

Simple rules: DB => DM (80%)
BL => not DM (80%)

discovering rules Course1 => Course2

Idea!

generate all rules X=>Y and verify them

Unfortunately too many …..

C++ => Java	Java => C++	C++ => BL
BL => C++	C++ => DB	DB => C++
…..		

For n courses, n^2 possible pairs

If we want (X,Y)=>Z we have n^3 possible triples, etc.

example 2

Course/student data set (again)

	C++	Java	Boolean logic	Data-bases	Data-mining	…
Anne P.	1	1	0	1	1	
Heikki M.	0	0	1	1	1	
Jouni S.	1	0	1	0	0	
Kari L.	1	1	1	1	0	
Taneli M.	0	0	0	1	1	
…						

Question: what are the ”core” courses and the ”specializations”

core courses and specializations

	C++	Java	Boolean logic	Data-bases	Data-mining	…
Anne P.	1	1	0	1	1	
Heikki M.	0	0	1	1	1	
Jouni S.	1	0	1	0	0	
Kari L.	1	1	1	1	0	
Taneli M.	0	0	0	1	1	
…						

generate all possible groupings of courses and try each one how well explain the data

For n courses and 2 groups: 2^n possible groupings

example 3

Paleontological data

	Species 1	Species 2	…	Species m
Site 1	1	0		1
Site 2	0	1		1
Site 3	1	0		0
…				
Site n	1	0		0

hidden structure: relative age of each site (an ordering of rows)

Bad news: n! = 1*2*3*…*(n-1)*n possible orderings

Good news: we can still do it

do we always know what we are looking for?

web search:

give a few keywords

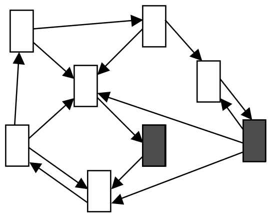
get the most relevant website

Google monopoly

idea of important websites

(a website is important if other many other important websites point to it)

importance of websites



many other ideas, but didn’t work so well

has everything been solved in web searching?

Never try: ”best basketball player after Jordan”

instead: ”top-ten basketball players”

Need more intelligent engines
- better language processing
- representation of the available information
- personalization
- …

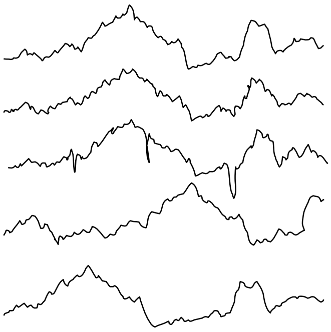
how should we analyse customer behavior?

collaborative filtering:

recommend a product to a customer based on her purchases

what is the right model?

how should we compare time-series?



current themes in our group (BRU)

(group leader: prof. Heikki Mannila)

analysis of scientific data
- data with geographic information
- biology, physics, paleontology

analysis of genomic sequences
- finding structure in the genome

analysis of matrices of 0-1 data

data clustering

summary

computer science is a really exciting science to study with endless possibilities

data mining and data analysis are very important fields

some of the world experts in the field are in the U of H

you should definitely consider taking some courses