CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #25: Data Mining - DB concepts

C. Faloutsos

CMU SCS

Must-Read Material

- Han + Kamber,
  - Chapter 2.1-2.4 (1st edition, 2000) or
  - Chapter 3.1-3.4 (2nd edition, 2006)

CMU SCS

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
- Data Mining

CMU SCS

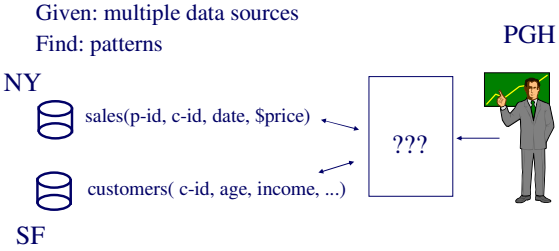
Data Mining - Detailed outline

- data warehouses; data cubes; OLAP
- classifiers
- association rules

CMU SCS

Data Ware-housing + OLAP

Problem:
Given: multiple data sources
Find: patterns



CMU SCS

Data Ware-housing

Problem:
Given: multiple data sources
Find: patterns (such as?)

CMU SCS

Data Ware-housing

Problem:
Given: multiple data sources
Find: patterns (such as?)
- classifiers ('supervised learning')
- 'association rules'; clusters ('unsup. learning')

↘

bread, milk -> butter

CMU SCS

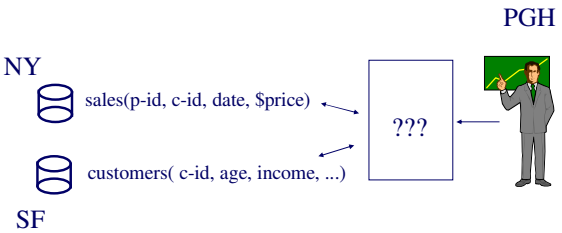
Data Ware-housing

Sub-problems:
P1: how to collect the data (-> Data Warehousing)
P1.1: how to collect counts (-> OLAP; datacubes)
P2: Decision trees
P3: Association rules
P4: Clustering

CMU SCS

Data Ware-housing

P1: how to collect the data ?

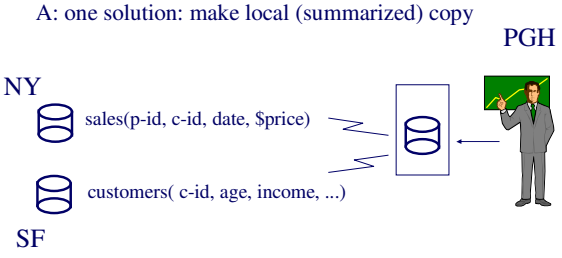


CMU SCS

Data Ware-housing

P1: how to collect the data ?

A: one solution: make local (summarized) copy



CMU SCS

Data Ware-housing

P1: how to collect the data ?

A: one solution: make local (summarized) copy

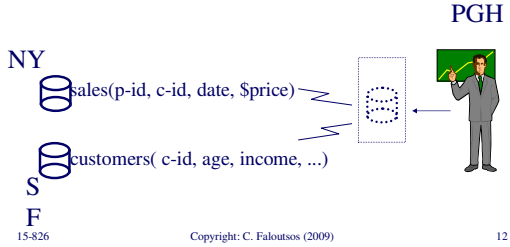
- how often to update?
- what/how to summarize?
- 'wrappers' and 'mediators': s/w modules to automate conversions and smooth discrepancies
- Q: how about a 'virtual' D/W?

CMU SCS

Data Ware-housing

Q: how about a 'virtual' D/W? (ie., 'views')

A: may delay OLTP machines (but: 'Cerebellum')



CMU SCS

D/W - OLAP

(OLAP= On Line Analytical Processing)

Sub-problems:
P1: how to collect the data (-> Data Warehousing)
P1.1: how to collect counts (-> OLAP; datacubes)

Problem: "is it true that shirts in large sizes sell better in dark colors?"

CMU SCS

D/W - OLAP

Problem: "is it true that shirts in large sizes sell better in dark colors?"

sales

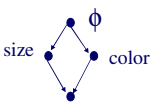
ci-d	p-id	Size	Color	$
C10	Shirt	L	Blue	30
~~C10~~	~~Pants~~	~~XL~~	~~Red~~	~~50~~
C20	Shirt	XL	White	20
…				

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

'color', 'size': DIMENSIONS
'count': MEASURE



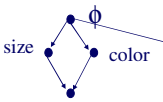
color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

'color', 'size': DIMENSIONS
'count': MEASURE



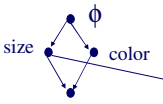
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

color; size

CMU SCS

DataCubes

'color', 'size': DIMENSIONS
'count': MEASURE



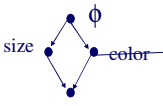
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

color; size

CMU SCS

DataCubes

'color', 'size': DIMENSIONS
'count': MEASURE



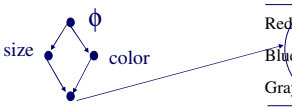
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

color; size

CMU SCS

DataCubes

'color', 'size': DIMENSIONS
'count': MEASURE



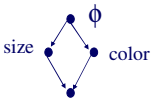
color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

'color', 'size': DIMENSIONS
'count': MEASURE



color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

DataCube

CMU SCS

DataCubes

SQL query to generate DataCube:

- Naively (and painfully:)

```
select size, color, count(*)
from sales where p-id = 'shirt'
group by size, color

select size, count(*)
from sales where p-id = 'shirt'
group by size
...
```

CMU SCS

DataCubes

SQL query to generate DataCube:

- with 'cube by' keyword:

```
select size, color, count(*)
from sales
where p-id = 'shirt'
cube by size, color
```


CMU SCS

DataCubes

(some additional concepts:

- concept hierarchy: eg., time: hour -> day-> month -> year
  (Q: other concept hierarchies?)
- 'star' schema ('snow-flake', 'constellation' etc)

)

CMU SCS

DataCubes

Q1: How to store a dataCube
Q2: What operations should we support?
Q3: How to index a dataCube?

CMU SCS

DataCubes

Q1: How to store a dataCube?

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q1: How to store a dataCube?

A1: Relational (R-OLAP)

Color	Size	count
'all'	'all'	47
Blue	'all'	14
Blue	M	3
…		

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q1: How to store a dataCube?

A2: Multi-dimensional (M-OLAP)

A3: Hybrid (H-OLAP)

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Pros/Cons:
ROLAP strong points: (DSS, Metacube)

CMU SCS

DataCubes

Pros/Cons:
ROLAP strong points: (DSS, Metacube)
- use existing RDBMS technology
- scale up better with dimensionality

CMU SCS

DataCubes

Pros/Cons:
MOLAP strong points: (EssBase/hyperion.com)
- faster indexing

(careful with: high-dimensionality; sparseness)

HOLAP: (MS SQL server OLAP services)
- detail data in ROLAP; summaries in MOLAP

CMU SCS

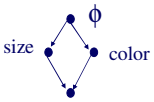
DataCubes

Q1: How to store a dataCube
Q2: What operations should we support?
Q3: How to index a dataCube?

CMU SCS

DataCubes

Q2: What operations should we support?



color; size

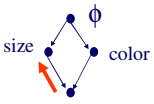
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q2: What operations should we support?

Roll-up



color; size

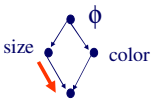
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q2: What operations should we support?

Drill-down



color; size

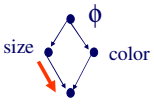
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q2: What operations should we support?

Slice



color; size

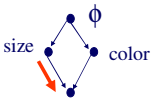
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q2: What operations should we support?

Dice



color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q2: What operations should we support?

- Roll-up
- Drill-down
- Slice
- Dice
- (Pivot/rotate; drill-across; drill-through
- top N
- moving averages, etc)

CMU SCS

DataCubes

Q1: How to store a dataCube

Q2: What operations should we support?

Q3: How to index a dataCube?

CMU SCS

DataCubes

Q3: How to index a dataCube?

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q3: How to index a dataCube?

A1: Bitmaps

S	M	L
1		
1		
	1	
…	…	…

Red	Blue	Gray
1		
	1	
		1
…	…	…

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Q3: How to index a dataCube?

A2: Join indices (see [Han+Kamber])

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

CMU SCS

DataCubes

Parallelism - 'measure' classes:

- distributive (eg., 'sum') -> easily combined
- algebraic (eg., 'avg') -> combine-able
- holistic (eg., 'median') -> nope!

CMU SCS

DataCubes

Drill:

- 'count'?
- 'max', 'min'?
- '90-percentile'?
- standard deviation?

CMU SCS

DataCubes

Drill:

• 'count'?	distributive
• 'max', 'min'?	distributive
• '90-percentile'?	holistic
• standard deviation?	algebraic

CMU SCS

D/W - OLAP - Conclusions

- D/W: copy (summarized) data + analyze
- OLAP - concepts:
  - DataCube
  - R/M/H-OLAP servers
  - 'dimensions'; 'measures'
  - concept hierarchies (day->month->year)