

CMU SCS

15-826: Multimedia Databases and Data Mining

Data Mining - DB concepts
C. Faloutsos

CMU SCS

Outline

Goal: 'Find **similar** / **interesting** things'

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

15-826 Copyright: C. Faloutsos (2005) 2

CMU SCS

Data Mining - Detailed outline

- Statistics
- AI - decision trees
- ➔ • DB
 - data warehouses; data cubes; OLAP
 - classifiers
 - association rules
 - misc. topics:
 - reconstruction of info
 - network databases; time sequence forecasting

15-826 Copyright: C. Faloutsos (2005) 3

CMU SCS

Data Ware-housing + OLAP

Problem:
Given: multiple data sources
Find: patterns

NY
SF

sales(p-id, c-id, date, \$price)
customers(c-id, age, income, ...)

PGH

???

15-826 Copyright: C. Faloutsos (2005) 4

CMU SCS

Data Ware-housing

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

15-826 Copyright: C. Faloutsos (2005) 5

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

15-826 Copyright: C. Faloutsos (2005) 6

Data Ware-housing

Sub-problems:

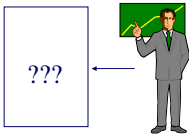
- ➔ P1: how to collect the data (-> Data Warehousing)
- P1.1: how to collect counts (-> OLAP; datacubes)
- P2: Decision trees
- P3: Association rules
- P4: Clustering

15-826 Copyright: C. Faloutsos (2005) 7

Data Ware-housing

P1: how to collect the data ?

NY
 sales(p-id, c-id, date, \$price)
 customers(c-id, age, income, ...)
 SF

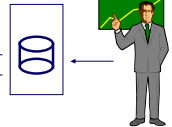
PGH


15-826 Copyright: C. Faloutsos (2005) 8

Data Ware-housing

P1: how to collect the data ?
 A: one solution: make local (summarized) copy

NY
 sales(p-id, c-id, date, \$price)
 customers(c-id, age, income, ...)
 SF

PGH


15-826 Copyright: C. Faloutsos (2005) 9

Data Ware-housing

P1: how to collect the data ?
 A: one solution: make local (summarized) copy

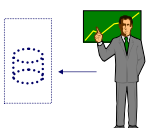
- how often to update?
- what/how to summarize?
- ‘wrappers’ and ‘mediators’: s/w modules to automate conversions and smooth discrepancies
- Q: how about a ‘virtual’ D/W?

15-826 Copyright: C. Faloutsos (2005) 10

Data Ware-housing

Q: how about a ‘virtual’ D/W? (ie., ‘views’)
 A: may delay OLTP machines (but: ‘Cerebellum’)

NY
 sales(p-id, c-id, date, \$price)
 customers(c-id, age, income, ...)
 S
 F

PGH


15-826 Copyright: C. Faloutsos (2005) 11

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?”

15-826 Copyright: C. Faloutsos (2005) 12

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

15-826 Copyright: C. Faloutsos (2005) 13

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

15-826 Copyright: C. Faloutsos (2005) 14

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

15-826 Copyright: C. Faloutsos (2005) 15

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

15-826 Copyright: C. Faloutsos (2005) 16

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

15-826 Copyright: C. Faloutsos (2005) 17

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

15-826 Copyright: C. Faloutsos (2005) 18

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

DataCube

15-826 Copyright: C. Faloutsos (2005) 19

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
```
- ...

15-826 Copyright: C. Faloutsos (2005) 20

**DataCubes**

SQL query to generate DataCube:

- with 'cube by' keyword:  

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

15-826 Copyright: C. Faloutsos (2005) 21

**DataCubes**

(some additional concepts:

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

15-826 Copyright: C. Faloutsos (2005) 22

**DataCubes**

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

15-826 Copyright: C. Faloutsos (2005) 23

**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  |

15-826 Copyright: C. Faloutsos (2005) 24

**DataCubes**

Q1: How to store a dataCube?  
A1: Relational (R-OLAP)

| Color | Size  | count | C / S | S  | M | L  | TOT |
|-------|-------|-------|-------|----|---|----|-----|
| 'all' | 'all' | 47    | Red   | 20 | 3 | 5  | 28  |
| Blue  | 'all' | 14    | Blue  | 3  | 3 | 8  | 14  |
| Blue  | M     | 3     | Gray  | 0  | 0 | 5  | 5   |
| ...   |       |       | TOT   | 23 | 6 | 18 | 47  |

15-826 Copyright: C. Faloutsos (2005) 25

**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  |

15-826 Copyright: C. Faloutsos (2005) 26

**DataCubes**

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

15-826 Copyright: C. Faloutsos (2005) 27

**DataCubes**

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

- use existing RDBMS technology
- scale up better with dimensionality

15-826 Copyright: C. Faloutsos (2005) 28

**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

15-826 Copyright: C. Faloutsos (2005) 29

**DataCubes**

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

15-826 Copyright: C. Faloutsos (2005) 30

**DataCubes**

Q2: What operations should we support?

| 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  |

15-826 Copyright: C. Faloutsos (2005) 31

**DataCubes**

Q2: What operations should we support?

**Roll-up**

| 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  |

15-826 Copyright: C. Faloutsos (2005) 32

**DataCubes**

Q2: What operations should we support?

**Drill-down**

| 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  |

15-826 Copyright: C. Faloutsos (2005) 33

**DataCubes**

Q2: What operations should we support?

**Slice**

| 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  |

15-826 Copyright: C. Faloutsos (2005) 34

**DataCubes**

Q2: What operations should we support?

**Dice**

| 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  |

15-826 Copyright: C. Faloutsos (2005) 35

**DataCubes**

Q2: What operations should we support?

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

15-826 Copyright: C. Faloutsos (2005) 36

**DataCubes**

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

15-826 Copyright: C. Faloutsos (2005) 37

**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  |

15-826 Copyright: C. Faloutsos (2005) 38

**DataCubes**

Q3: How to index a dataCube?  
 A1: Bitmaps

| S   | M   | L   | Red | Blue | Gray |
|-----|-----|-----|-----|------|------|
| 1   |     |     | 1   |      |      |
| 1   |     |     |     | 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  |

15-826 Copyright: C. Faloutsos (2005) 39

**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  |

15-826 Copyright: C. Faloutsos (2005) 40

**DataCubes**

Parallelism - 'measure' classes:

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

15-826 Copyright: C. Faloutsos (2005) 41

**DataCubes**

Drill:

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

15-826 Copyright: C. Faloutsos (2005) 42



CMU SCS

## DataCubes

Drill:

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

15-826 Copyright: C. Faloutsos (2005) 43



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)

15-826 Copyright: C. Faloutsos (2005) 44



CMU SCS

## Reference

- Han + Kamber, chapter 2.1-2.4

15-826 Copyright: C. Faloutsos (2005) 45