

# 15-826: Multimedia Databases and Data Mining

*Data Mining - DB concepts*  
C. Faloutsos

---

---

---

---

---

---

---

---



## Outline

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

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

15-826 Copyright: C. Faloutsos (2007) 2

---

---

---


---

---

---

---

---



## 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 (2007) 3

---

---

---

---

---

---

---

---

CMU SCS

## Data Ware-housing + OLAP

Problem:  
 Given: multiple data sources  
 Find: patterns

NY

sales(p-id, c-id, date, \$price)

SF

customers( c-id, age, income, ...)

???

PGH

15-826 Copyright: C. Faloutsos (2007) 4

---

---

---

---

---

---

---

---

CMU SCS

## Data Ware-housing

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

15-826 Copyright: C. Faloutsos (2007) 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 (2007) 6

---

---

---

---

---

---

---

---

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

15-826 Copyright: C. Faloutsos (2007) 7

---

---

---

---

---

---

---


---


CMU SCS

## 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 (2007) 8

---

---

---

---

---

---

---

---


CMU SCS


## 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 (2007) 9

---

---

---

---

---

---

---

---

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?

15-826 Copyright: C. Faloutsos (2007) 10

---

---

---

---

---

---

---

---

CMU SCS

## Data Ware-housing

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

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

PGH

NY  
sales(p-id, c-id, date, \$price)

SF  
customers( c-id, age, income, ...)

15-826 Copyright: C. Faloutsos (2007) 11

---

---

---

---

---

---

---

---

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?"

15-826 Copyright: C. Faloutsos (2007) 12

---

---

---

---

---

---

---

---

**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	\$	C / S	S	M	L	TOT
						Red	20	3	5	28
	C10	Shirt	L	Blue	30	Blue	3	3	8	14
	<del>C10</del>	<del>Pants</del>	<del>XL</del>	<del>Red</del>	<del>50</del>	Gray	0	0	5	5
	C20	Shirt	XL	White	20	TOT	23	6	18	47
...										

15-826 Copyright: C. Faloutsos (2007) 13

---

---

---

---

---

---

---

---

---

---

**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

15-826 Copyright: C. Faloutsos (2007) 14

---

---

---

---

---

---

---

---

---

---

**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

15-826 Copyright: C. Faloutsos (2007) 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 (2007) 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 (2007) 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 (2007) 18

---

---

---

---

---

---

---

---

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

DataCube

15-826 Copyright: C. Faloutsos (2007) 19

---

---

---

---

---

---

---

---

---

---

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
...
```

15-826 Copyright: C. Faloutsos (2007) 20

---

---

---

---

---

---

---

---

---

---

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
```

15-826 Copyright: C. Faloutsos (2007) 21

---

---

---

---

---

---

---

---

---

---

CMU SCS

## 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 (2007) 22

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

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

15-826 Copyright: C. Faloutsos (2007) 23

---

---

---

---

---

---

---

---

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

15-826 Copyright: C. Faloutsos (2007) 24

---

---

---

---

---

---

---

---



CMU SCS

## DataCubes

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

Color	Size	count	C / S	S	M	L	TOT
Red	'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 (2007) 25

---

---

---

---

---

---

---

---

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

15-826 Copyright: C. Faloutsos (2007) 26

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

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

15-826 Copyright: C. Faloutsos (2007) 27

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

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

- use existing RDBMS technology
- scale up better with dimensionality

15-826 Copyright: C. Faloutsos (2007) 28

---

---

---

---

---

---

---

---

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

15-826 Copyright: C. Faloutsos (2007) 29

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

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

15-826 Copyright: C. Faloutsos (2007) 30

---

---

---

---

---

---

---

---

CMU SCS

## 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 (2007) 31

---

---

---

---

---

---

---

---

---

---

CMU SCS

## 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 (2007) 32

---

---

---

---

---

---

---

---

---

---

CMU SCS

## 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 (2007) 33

---

---

---

---

---

---

---

---

---

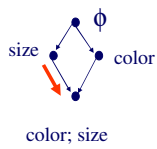
---

CMU SCS

## 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 (2007) 34

---

---

---

---

---

---

---

---

---

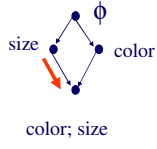
---

CMU SCS

## 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 (2007) 35

---

---

---

---

---

---

---

---

---

---

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)

15-826 Copyright: C. Faloutsos (2007) 36

---

---

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

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

15-826 Copyright: C. Faloutsos (2007) 37

---

---

---

---

---

---

---

---

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

15-826 Copyright: C. Faloutsos (2007) 38

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

Q3: How to index a dataCube?  
 A1: Bitmaps

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

S	M	L
1		
1		
	1	
...	...	...

Red	Blue	Gray
1		
	1	
		1
...	...	...

15-826 Copyright: C. Faloutsos (2007) 39

---

---

---

---

---

---

---

---

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

15-826 Copyright: C. Faloutsos (2007) 40

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

Parallelism - 'measure' classes:

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

15-826 Copyright: C. Faloutsos (2007) 41

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

Drill:

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

15-826 Copyright: C. Faloutsos (2007) 42

---

---

---

---

---

---

---

---

CMU SCS

## DataCubes

Drill:

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

15-826                    Copyright: C. Faloutsos (2007)                    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 (2007)                    44

---

---

---

---

---

---

---

---

CMU SCS

## Reference

- Han + Kamber, chapter 2.1-2.4

15-826                    Copyright: C. Faloutsos (2007)                    45

---

---

---

---

---

---

---

---