



15-826: Multimedia Databases and Data Mining

Lecture #28: Data Mining - OLAP

C. Faloutsos



Must-Read Material

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

15-826

Copyright: C. Faloutsos (2011)

2



Outline

Goal: 'Find similar / interesting things'

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

15-826

Copyright: C. Faloutsos (2011)

3

 CMU SCS

Data Mining - Detailed outline

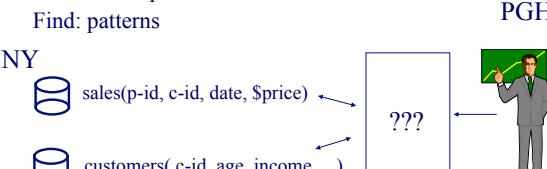
- ➡ data warehouses; data cubes; OLAP
- classifiers
- association rules

15-826 Copyright: C. Faloutsos (2011) 4

 CMU SCS

Data Warehousing + OLAP

Problem:
Given: multiple data sources
Find: patterns



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

SF

15-826 Copyright: C. Faloutsos (2011) 5

 CMU SCS

Data Warehousing

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

15-826 Copyright: C. Faloutsos (2011) 6

 CMU SCS

Data Warehousing

Problem:

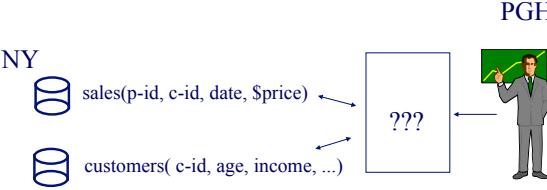
Given: multiple data sources
 Find: patterns (such as?)
 • classifiers ('supervised learning')
 • 'association rules'; clusters ('unsup. learning')
 ↘
 bread, milk -> butter

15-826 Copyright: C. Faloutsos (2011) 7

 CMU SCS

Data Warehousing

P1: how to collect the data ?



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

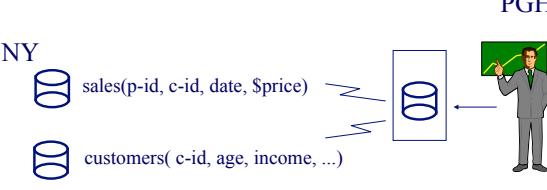
SF

Copyright: C. Faloutsos (2011) 8

 CMU SCS

Data Warehousing

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

Copyright: C. Faloutsos (2011) 9



Data Warehousing

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 (2011)

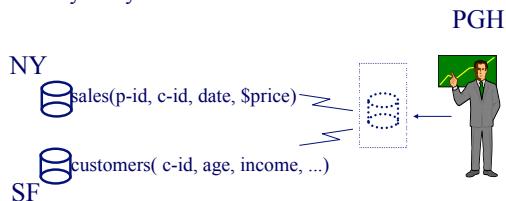
10



Data Warehousing

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

A: may delay OLTP machines



15-826

Copyright: C. Faloutsos (2011)

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 (2011)

12

 CMU SCS

D/W - OLAP

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

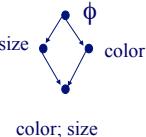
sales	ci-d	p-id	Size	Color	\$	C / S	S	M	L	TOT
						Red	20	3	5	28
	C10	Shirt	L	Blue	30	Blue	3	3	8	14
	-C10	Pants	XL	Red	50	Gray	0	0	5	5
	C20	Shirt	XL	White	20	TOT	23	6	18	47
	...									

15-826 Copyright: C. Faloutsos (2011) 13

 CMU SCS

DataCubes

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



size φ color
 color; size

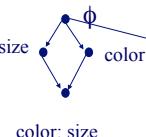
	C / S	S	M	L	TOT
Red	20	3	5	28	
Blue	3	3	8	14	
Gray	0	0	5	5	
TOT	23	6	18	47	

15-826 Copyright: C. Faloutsos (2011) 14

 CMU SCS

DataCubes

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



size φ color
 color; size

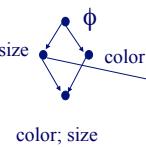
	C / S	S	M	L	TOT
Red	20	3	5	28	
Blue	3	3	8	14	
Gray	0	0	5	5	
TOT	23	6	18	47	

15-826 Copyright: C. Faloutsos (2011) 15

 CMU SCS

DataCubes

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



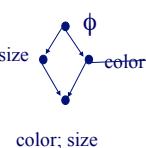
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

15-826 Copyright: C. Faloutsos (2011) 16

 CMU SCS

DataCubes

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



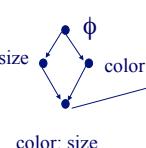
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

15-826 Copyright: C. Faloutsos (2011) 17

 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

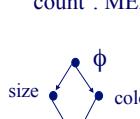
15-826 Copyright: C. Faloutsos (2011) 18


 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 (2011)

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  
...
```

Copyright: C. Faloutsos (2011)

 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 (2011)

21



DataCubes

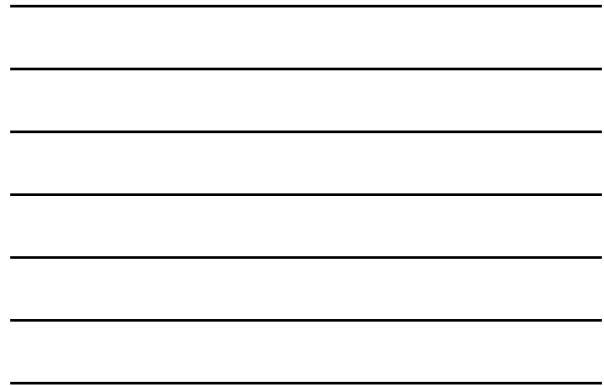
(some additional concepts:

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

15-826

Copyright: C. Faloutsos (2011)

22



DataCubes

Q1: How to store a dataCube

Q2: What operations should we support?

Q3: How to index a dataCube?



DataCubes

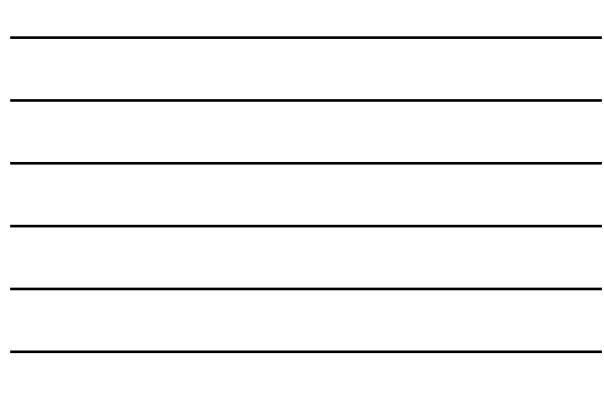
Q1: How to store a dataCube?

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 (2011)

24



 CMU SCS

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 (2011) 25

 CMU SCS

DataCubes

Q1: How to store a dataCube?

A2: Multi-dimensional (M-OLAP)

A3: Hybrid (H-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 (2011) 26

 CMU SCS

DataCubes

Pros/Cons:

ROLAP strong points: (DSS, Metacube)

15-826 Copyright: C. Faloutsos (2011) 27



DataCubes

Pros/Cons:

ROLAP strong points: (DSS, Metacube)

- use existing RDBMS technology
- scale up better with dimensionality

15-826

Copyright: C. Faloutsos (2011)

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 (2011)

29



DataCubes

Q1: How to store a dataCube

► Q2: What operations should we support?

Q3: How to index a dataCube?

15-826

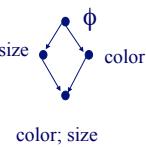
Copyright: C. Faloutsos (2011)

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

color; size

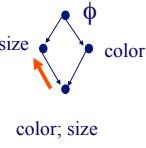
15-826 Copyright: C. Faloutsos (2011) 31

 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

color; size

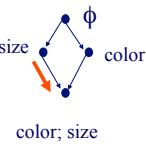
15-826 Copyright: C. Faloutsos (2011) 32

 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

color; size

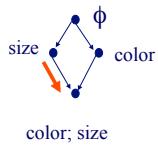
15-826 Copyright: C. Faloutsos (2011) 33



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 (2011)

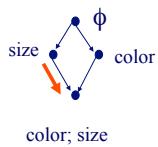
34



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 (2011)

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 (2011)

36

 CMU SCS

 details

DataCubes

Q1: How to store a dataCube

Q2: What operations should we support?

→ Q3: How to index a dataCube?

15-826

Copyright: C. Faloutsos (2011)

37

CMU SCS

 details

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 (2011)

38


CMU SCS


details

DataCubes

Q3: How to index a dataCube?

A1: Bitmaps

S	M	L	Red	Blue	Gray	C / S	S	M	L	TOT
1			1			Red	20	3	5	28
1				1		Blue	3	3	8	14
	1				1	Gray	0	0	5	5
...	TOT	23	6	18	47

15-826

Copyright: C. Faloutsos (2011)

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 (2011) 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 (2011) 41

 CMU SCS

DataCubes

Drill:

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

15-826 Copyright: C. Faloutsos (2011) 42



DataCubes

Drill:

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

15-826

Copyright: C. Faloutsos (2011)

43



D/W - OLAP - Conclusions

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

15-826

Copyright: C. Faloutsos (2011)

44
