



15-826: Multimedia Databases and Data Mining

SVD - part II (case studies)

C. Faloutsos



Outline

Goal: ‘Find similar / interesting things’

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

15-826

Copyright: C. Faloutsos (2006)

2



Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- multimedia
- ...

15-826

Copyright: C. Faloutsos (2006)

3



SVD - Detailed outline

- Motivation
- Definition - properties
- Interpretation
- Complexity
- Case studies
- SVD properties
- Conclusions

15-826

Copyright: C. Faloutsos (2006)

4



SVD - Case studies

- • multi-lingual IR; LSI queries
- compression
- PCA - ‘ratio rules’
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826

Copyright: C. Faloutsos (2006)

5



Case study - LSI

- Q1: How to do queries with LSI?
 Q2: multi-lingual IR (english query, on spanish text?)

15-826

Copyright: C. Faloutsos (2006)

6



Case study - LSI

Q1: How to do queries with LSI?

Problem: Eg., find documents with ‘data’

$$\begin{array}{c} \text{retrieval} \\ \text{data inf} \downarrow \text{brain lung} \\ \uparrow \quad \downarrow \\ \text{CS} \\ \downarrow \\ \text{MD} \end{array} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0.58 & 0.58 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \\ 0.58 & 0.58 \\ 0.58 & 0.58 \\ 0 & 0 \\ 0 & 0.27 \\ 0.71 & 0.71 \end{bmatrix}$$

15-826

Copyright: C. Faloutsos (2006)

7



Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into ‘concept space’ – how?

$$\begin{array}{c} \text{retrieval} \\ \text{data inf} \downarrow \text{brain lung} \\ \uparrow \quad \downarrow \\ \text{CS} \\ \downarrow \\ \text{MD} \end{array} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0.58 & 0.58 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \\ 0.58 & 0.58 \\ 0.58 & 0.58 \\ 0 & 0 \\ 0 & 0.27 \\ 0.71 & 0.71 \end{bmatrix}$$

15-826

Copyright: C. Faloutsos (2006)

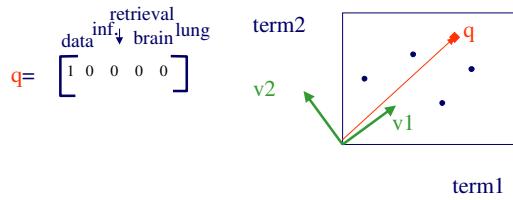
8



Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into ‘concept space’ – how?



15-826

Copyright: C. Faloutsos (2006)

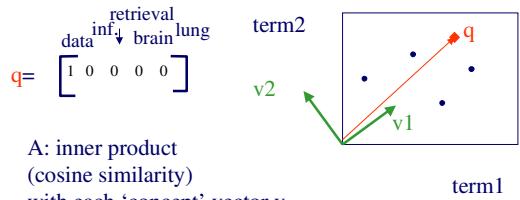
9



Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into ‘concept space’ – how?



15-826

Copyright: C. Faloutsos (2006)

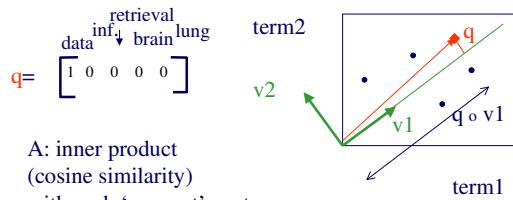
10



Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into ‘concept space’ – how?



A: inner product
(cosine similarity)
with each ‘concept’ vector v_i

15-826

Copyright: C. Faloutsos (2006)

11



Case study - LSI

compactly, we have:

$$q_{\text{concept}} = q V$$

Eg:

$$q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}$$

↓

CS-concept

term-to-concept
similarities

15-826

Copyright: C. Faloutsos (2006)

12



Case study - LSI

Drill: how would the document ('information', 'retrieval') handled by LSI?

term-to-concept
similarities

15-826

Copyright: C. Faloutsos (2006)

13



Case study - LSI

Drill: how would the document ('information', 'retrieval') handled by LSI? A: SAME:

$$\begin{aligned} d_{\text{concept}} &= d \mathbf{V} \\ \text{Eg: } d &= \begin{bmatrix} \text{data} & \substack{\text{inf} \downarrow \\ \text{retrieval}} & \text{brain} & \text{lung} \end{bmatrix} \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{bmatrix} 1.16 & 0 \end{bmatrix} \\ d &= \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} \end{aligned}$$

term-to-concept
similarities

15-826

Copyright: C. Faloutsos (2006)

14



Case study - LSI

Observation: document ('information', 'retrieval') will be retrieved by query ('data'), although it does not contain 'data'!!

CS-concept

$$\begin{aligned} d &= \begin{bmatrix} \text{data} & \substack{\text{inf} \downarrow \\ \text{retrieval}} & \text{brain} & \text{lung} \end{bmatrix} \rightarrow \begin{bmatrix} 1.16 & 0 \end{bmatrix} \\ q &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.58 & 0 \end{bmatrix} \end{aligned}$$

15-826

Copyright: C. Faloutsos (2006)

15



Case study - LSI

Q1: How to do queries with LSI?

→ Q2: multi-lingual IR (english query, on spanish text?)

15-826

Copyright: C. Faloutsos (2006)

16



Case study - LSI

- Problem:
 - given many documents, translated to both languages (eg., English and Spanish)
 - answer queries across languages

15-826

Copyright: C. Faloutsos (2006)

17



Case study - LSI

- Solution: ~ LSI

$$\begin{array}{c} \text{informacion} \\ \text{datos} \\ \uparrow \quad \downarrow \\ \text{data} \quad \text{brain} \quad \text{lung} \\ \uparrow \quad \downarrow \\ \text{CS} \quad \text{MD} \\ \uparrow \quad \downarrow \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 4 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

15-826

Copyright: C. Faloutsos (2006)

18

Case study - LSI

- Solution: ~ LSI

$$M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

15-826 Copyright: C. Faloutsos (2006) 19

SVD - Case studies

- multi-lingual IR; LSI queries
- compression
- PCA - ‘ratio rules’
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826 Copyright: C. Faloutsos (2006) 20

Case study: compression

[Korn+97]

Problem:

- given a matrix
- compress it, but maintain ‘random access’

(surprisingly, its solution leads to data mining and visualization...)

15-826 Copyright: C. Faloutsos (2006) 21

Problem - specs

- ~10**6 rows; ~10**3 columns; no updates;
- random access to any cell(s); small error: OK

customer	W _o	T _n	F _r	S _a	S _u
	7/10/98	7/11/98	7/12/98	7/13/98	7/14/98
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

15-826 Copyright: C. Faloutsos (2006) 22

Idea

15-826 Copyright: C. Faloutsos (2006) 23

SVD - reminder

- space savings: 2:1
- minimum RMS error

15-826 Copyright: C. Faloutsos (2006) 24

Case study: compression

outliers?

A: treat separately
(SVD with 'Deltas')

15-826 Copyright: C. Faloutsos (2006) 25

Compression - Performance

- 3 pass algo (-> scalability) (HOW?)
- random cell(s) reconstruction
- 10:1 compression with < 2% error

15-826 Copyright: C. Faloutsos (2006) 26

Performance - scaleup

15-826 Copyright: C. Faloutsos (2006) 27

Compression - Visualization

- no Gaussian clusters; Zipf-like distribution

15-826 Copyright: C. Faloutsos (2006) 28

SVD - Case studies

- multi-lingual IR; LSI queries
- compression
- **→ PCA - 'ratio rules'**
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826 Copyright: C. Faloutsos (2006) 29

PCA - 'Ratio Rules'

[Korn+00]

Typically: 'Association Rules' (eg., {bread, milk} → {butter})

But:

- which set of rules is 'better'?
- how to reconstruct missing/corrupted values?
- need binary/bucketized values

15-826 Copyright: C. Faloutsos (2006) 30

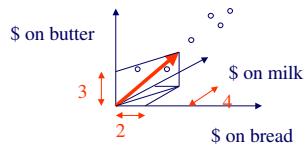


PCA - 'Ratio Rules'

Idea: try to find 'concepts':

- singular vectors dictate rules about ratios:

$$\text{bread:milk:butter} = 2:4:3$$



15-826

Copyright: C. Faloutsos (2006)

31



PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- Q1: which set of rules is 'better'?
- - Q2: how to reconstruct missing/corrupted values?
- Q3: is there need for binary/bucketized values?
- Q4: how to interpret the rules (= 'principal components')?

15-826

Copyright: C. Faloutsos (2006)

32



PCA - 'Ratio Rules'

Q2: how to reconstruct missing/corrupted values?

Eg:

- rule: bread:milk = 3:4
- a customer spent \$6 on bread - how about milk?

15-826

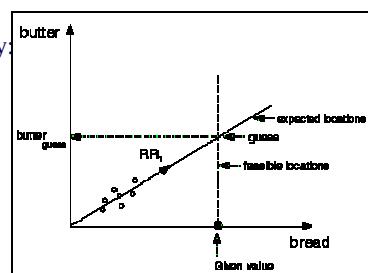
Copyright: C. Faloutsos (2006)

33



PCA - 'Ratio Rules'

pictorially:



15-826

Copyright: C. Faloutsos (2006)

34



PCA - 'Ratio Rules'

harder cases: overspecified/underspecified

over-specified:

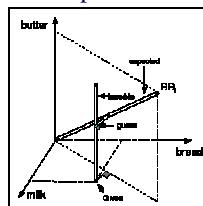
$$\text{•milk:bread:butter} = 1:2:3$$

•a customer got

- \$2 bread and \$4 milk

•how much milk?

Answer: minimize distance between 'feasible' and 'expected' values (using SVD...)



15-826

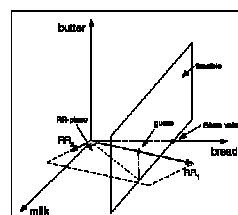
Copyright: C. Faloutsos (2006)

35



PCA - 'Ratio Rules'

harder cases: underspecified



15-826

Copyright: C. Faloutsos (2006)

36



PCA - 'Ratio Rules'

bottom line: we can reconstruct any count of missing values

This is very useful:

- can spot outliers (how?)
- can measure the ‘goodness’ of a set of rules (how?)

15-826

Copyright: C. Faloutsos (2006)

37



PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- ➡ – Q1: which set of rules is ‘better’?
- ✓ – Q2: how to reconstruct missing/corrupted values?
- Q3: is there need for binary/bucketized values?
- Q4: how to interpret the rules (= ‘principal components’)?

15-826

Copyright: C. Faloutsos (2006)

38



PCA - 'Ratio Rules'

- Q1: which set of rules is ‘better’?
- A: the ones that needs the fewest outliers:
 - pretend we don’t know a value (eg., \$ of ‘Smith’ on ‘bread’)
 - reconstruct it
 - and sum up the squared errors, for all our entries
- (other Answers are also reasonable)

15-826

Copyright: C. Faloutsos (2006)

39



PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- ✓ – Q1: which set of rules is ‘better’?
- ✓ – Q2: how to reconstruct missing/corrupted values?
- ➡ – Q3: is there need for binary/bucketized values?
- Q4: how to interpret the rules (= ‘principal components’)?

15-826

Copyright: C. Faloutsos (2006)

40



PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- ✓ – Q1: which set of rules is ‘better’?
- ✓ – Q2: how to reconstruct missing/corrupted values?
- ✓ – Q3: is there need for binary/bucketized values? NO
- ➡ – Q4: how to interpret the rules (= ‘principal components’)?

15-826

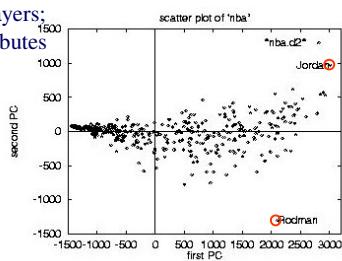
Copyright: C. Faloutsos (2006)

41



PCA - Ratio Rules

NBA dataset
~500 players;
~30 attributes



15-826

Copyright: C. Faloutsos (2006)

42



PCA - Ratio Rules

- PCA: get singular vectors v_1, v_2, \dots
- ignore entries with small abs. value
- try to interpret the rest

15-826

Copyright: C. Faloutsos (2006)

43



PCA - Ratio Rules

NBA dataset - V matrix (term to ‘concept’ similarities)

field	RR_1	RR_2	RR_3
minutes played	.808	-.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		-.489	.602
assists			-.486
steals			-.07

 v_1

15-826

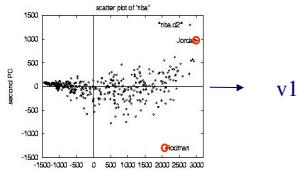
Copyright: C. Faloutsos (2006)

44



Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?



15-826

Copyright: C. Faloutsos (2006)

45



Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?
- A: ‘goodness’ of player

15-826

Copyright: C. Faloutsos (2006)

46



Ratio Rules - example

- RR2: points: rebounds negatively correlated(!)

field	RR_1	RR_2	RR_3
minutes played	.808	-.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		-.489	.602
assists			-.486
steals			-.07

15-826

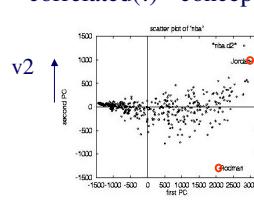
Copyright: C. Faloutsos (2006)

47



Ratio Rules - example

- RR2: points: rebounds negatively correlated(!) - concept?



15-826

Copyright: C. Faloutsos (2006)

48



Ratio Rules - example

- RR2: points: rebounds negatively correlated(!) - concept?
- A: position: offensive/defensive

15-826

Copyright: C. Faloutsos (2006)

49



SVD - Case studies

- multi-lingual IR; LSI queries
- compression
- PCA - ‘ratio rules’
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826

Copyright: C. Faloutsos (2006)

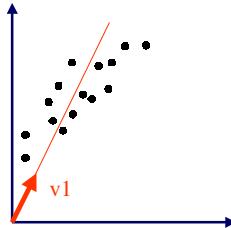
50



K-L transform

[Duda & Hart]; [Fukunaga]

A subtle point:
SVD will give vectors that
go through the origin



15-826

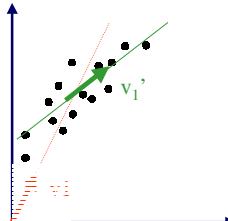
Copyright: C. Faloutsos (2006)

51



K-L transform

A subtle point:
SVD will give vectors that
go through the origin
Q: how to find v_1' ?



15-826

Copyright: C. Faloutsos (2006)

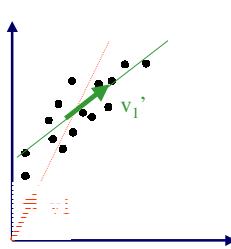
52



K-L transform

A subtle point:
SVD will give vectors that
go through the origin
Q: how to find v_1' ?

A: ‘centered’ PCA, ie.,
move the origin to center
of gravity



15-826

Copyright: C. Faloutsos (2006)

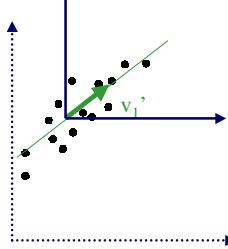
53



K-L transform

A subtle point:
SVD will give vectors that
go through the origin
Q: how to find v_1' ?

A: ‘centered’ PCA, ie.,
move the origin to center
of gravity
and THEN do SVD



15-826

Copyright: C. Faloutsos (2006)

54



K-L transform

- How to ‘center’ a set of vectors (= data matrix)?
- What is the covariance matrix?
- A: see textbook
- (‘whitening transformation’)

15-826

Copyright: C. Faloutsos (2006)

55



Conclusions

- SVD: popular for dimensionality reduction / compression
- SVD is the ‘engine under the hood’ for PCA (principal component analysis)
- ... as well as the Karhunen-Lowe transform
- (and there is more to come ...)

15-826

Copyright: C. Faloutsos (2006)

56



References

- Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. New York, Wiley.
- Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, Academic Press.
- Jolliffe, I. T. (1986). Principal Component Analysis, Springer Verlag.

15-826

Copyright: C. Faloutsos (2006)

57



References

- Korn, F., H. V. Jagadish, et al. (May 13-15, 1997). Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. ACM SIGMOD, Tucson, AZ.
- Korn, F., A. Labrinidis, et al. (1998). Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. VLDB, New York, NY.

15-826

Copyright: C. Faloutsos (2006)

58



References

- Korn, F., A. Labrinidis, et al. (2000). "Quantifiable Data Mining Using Ratio Rules." VLDB Journal 8(3-4): 254-266.
- Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in C, Cambridge University Press.

15-826

Copyright: C. Faloutsos (2006)

59