CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #19: SVD - part III (more case studies)

C. Faloutsos

CMU SCS

Must-read Material

- Textbook Appendix D
- Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms.
- Brin, S. and L. Page (1998). Anatomy of a Large-Scale Hypertextual Web Search Engine. 7th Intl World Wide Web Conf.

CMU SCS

Outline

Goal: 'Find similar / interesting things'

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

CMU SCS

Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- Singular Value Decomposition (SVD)
- multimedia
- ...

CMU SCS

SVD - Detailed outline

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

CMU SCS

SVD - detailed outline

- ...
- Case studies
- SVD properties
- more case studies
  - google/Kleinberg algorithms
  - query feedbacks
- Conclusions

CMU SCS

SVD - Other properties - summary

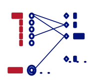
- can produce orthogonal basis (obvious) (who cares?)
- can solve over- and under-determined linear problems (see C(1) property)
- can compute ‘fixed points’ (= ‘steady state prob. in Markov chains’) (see C(4) property)

CMU SCS

IMPORTANT!

Properties – sneak preview:

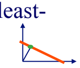
A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{\mathbf{T}}_{[r \times m]}$



B(5): $(\mathbf{A}^{T} \mathbf{A})^{k} \mathbf{v'} \sim (\text{constant})\ \mathbf{v}_1$

C(1): $\mathbf{A}_{[n \times m]} \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$

then, $\mathbf{x}_0 = \mathbf{V} \mathbf{\Lambda}^{(-1)} \mathbf{U}^{T} \mathbf{b}$: shortest, actual or least-squares solution



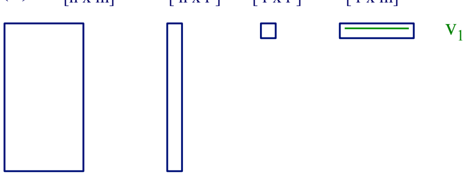
C(4): $\mathbf{A}^{\mathbf{T}} \mathbf{A}\ \mathbf{v}_1 = \lambda_1^{2}\ \mathbf{v}_1$

CMU SCS

IMPORTANT!

Properties – sneak preview:

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{\mathbf{T}}_{[r \times m]}$



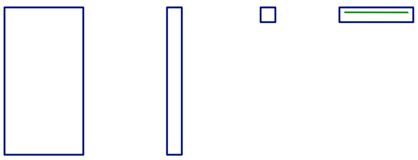
Document-term Doc-concept Concept-term

CMU SCS

IMPORTANT!

Properties – sneak preview:

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \, \mathbf{\Lambda}_{[r \times r]} \, \mathbf{V^T}_{[r \times m]}$



Libraries-books Libraries-concepts Concepts-books

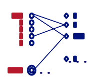
CMU SCS

IMPORTANT!

Properties – sneak preview:

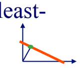
A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \, \mathbf{\Lambda}_{[r \times r]} \, \mathbf{V^T}_{[r \times m]}$

B(5): $(\mathbf{A}^T \mathbf{A})^k \, \mathbf{v'} \sim (\text{constant}) \, \mathbf{v}_1$



C(1): $\mathbf{A}_{[n \times m]} \, \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$

then, $\mathbf{x}_0 = \mathbf{V} \, \mathbf{\Lambda}^{(-1)} \, \mathbf{U}^T \, \mathbf{b}$**:** shortest, actual or least-squares solution



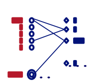
C(4): $\mathbf{A}^T \mathbf{A} \, \mathbf{v}_1 = \lambda_1^2 \, \mathbf{v}_1$

CMU SCS

IMPORTANT!

Properties – sneak preview:

Libraries Books



B(5): $(\mathbf{A}^T \mathbf{A})^k \, \mathbf{v'} \sim (\text{constant}) \, \mathbf{v}_1$

Convergence

CMU SCS

IMPORTANT!

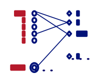
Properties – sneak preview:

B(5): $(\mathbf{A}^T \mathbf{A})^k \mathbf{v'} \sim$ (constant) $\mathbf{v}_1$

Fixed point

C(4): $\mathbf{A}^T \mathbf{A} \mathbf{v}_1 = \lambda_1^2 \mathbf{v}_1$

Libraries Books



book1

book…

15-826 Copyright: C. Faloutsos (2013) 13

CMU SCS

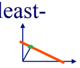
IMPORTANT!

Properties – sneak preview:

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^T_{[r \times m]}$

C(1): $\mathbf{A}_{[n \times m]} \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$

then, $\mathbf{x}_0 = \mathbf{V} \mathbf{\Lambda}^{(-1)} \mathbf{U}^T \mathbf{b}$: shortest, actual or least-squares solution



15-826 Copyright: C. Faloutsos (2013) 14

CMU SCS

SVD -outline of properties

- (A): obvious
- (B): less obvious
- (C): least obvious (and most powerful!)

15-826 Copyright: C. Faloutsos (2013) 15

CMU SCS

Properties - by defn.:

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V^T}_{[r \times m]}$

A(1): $\mathbf{U}^T_{[r \times n]} \mathbf{U}_{[n \times r]} = \mathbf{I}_{[r \times r]}$ (identity matrix)
A(2): $\mathbf{V}^T_{[r \times n]} \mathbf{V}_{[n \times r]} = \mathbf{I}_{[r \times r]}$
A(3): $\mathbf{\Lambda}^k = \text{diag}(\lambda_1^k, \lambda_2^k, \dots \lambda_r^k)$ (k: ANY real number)
A(4): $\mathbf{A^T} = \mathbf{V} \mathbf{\Lambda} \mathbf{U^T}$

CMU SCS

Less obvious properties

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V^T}_{[r \times m]}$

B(1): $\mathbf{A}_{[n \times m]} (\mathbf{A}^T)_{[m \times n]} = ??$

CMU SCS

Less obvious properties

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V^T}_{[r \times m]}$
B(1): $\mathbf{A}_{[n \times m]} (\mathbf{A}^T)_{[m \times n]} = \mathbf{U} \mathbf{\Lambda}^2 \mathbf{U}^T$
symmetric; Intuition?

CMU SCS

Less obvious properties

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \, \mathbf{\Lambda}_{[r \times r]} \, \mathbf{V^T}_{[r \times m]}$

B(1): $\mathbf{A}_{[n \times m]} \, (\mathbf{A}^T)_{[m \times n]} = \mathbf{U} \, \mathbf{\Lambda}^2 \, \mathbf{U}^T$

symmetric; Intuition?

'document-to-document' similarity matrix

B(2): symmetrically, for '**V**'

$(\mathbf{A}^T)_{[m \times n]} \, \mathbf{A}_{[n \times m]} = \mathbf{V} \, \mathbf{\Lambda}^2 \, \mathbf{V}^T$

Intuition?

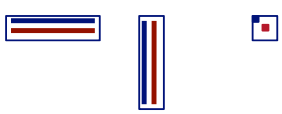
CMU SCS

Reminder: 'column orthonormal'

- $\mathbf{V}^T \quad \mathbf{V} \quad = \quad \mathbf{I}_{[r \times r]}$

v1

$\mathbf{v}_2$



$\mathbf{v_1}^T \times \mathbf{v_1} = 1$

$\mathbf{v_1}^T \times \mathbf{v_2} = 0$

CMU SCS

Less obvious properties

A: term-to-term similarity matrix

B(3): $((\mathbf{A}^T)_{[m \times n]} \, \mathbf{A}_{[n \times m]})^k = \mathbf{V} \, \mathbf{\Lambda}^{2k} \, \mathbf{V}^T$

and

B(4): $(\mathbf{A}^T \mathbf{A})^k \sim \mathbf{v}_1 \, \lambda_1^{2k} \, \mathbf{v}_1^T$ for k>>1

where

$\mathbf{v}_1$: [m x 1] first column (singular-vector) of **V**

λ_1: strongest singular value

CMU SCS

Proof of (B4)?

CMU SCS

Less obvious properties

B(4): $(\mathbf{A}^T \mathbf{A})^k \sim \mathbf{v}_1 \lambda_1^{2k} \mathbf{v}_1^T$ for k>>1

$\boxed{\text{B(5): } (\mathbf{A}^T \mathbf{A})^k \mathbf{v'} \sim \text{(constant) } \mathbf{v}_1}$

ie., for (almost) any **v'**, it converges to a vector parallel to $\mathbf{v}_1$

Thus, useful to compute first singular vector/ value (as well as the next ones, too...)

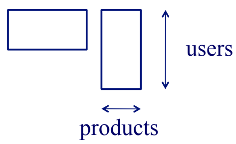
CMU SCS

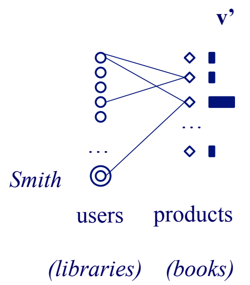
Proof of (B5)?

CMU SCS

Property (B5)

- Intuition:
  - $(A^T A)\ v'$
  - $(A^T A)^k\ v'$

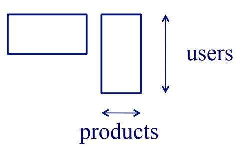


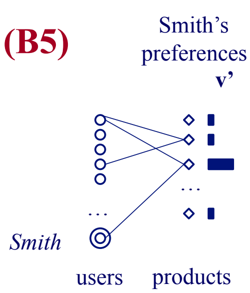


CMU SCS

Property (B5)

- Intuition:
  - $(A^T A)\ v'$
  - $(A^T A)^k\ v'$

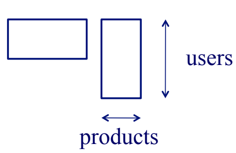


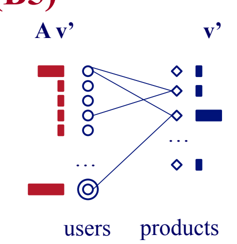


CMU SCS

Property (B5)

- Intuition:
  - $(A^T A)\ v'$
  - $(A^T A)^k\ v'$

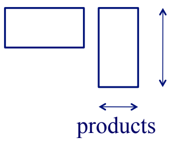


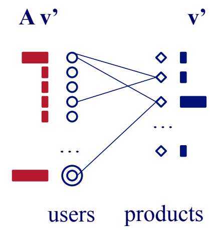


Property (B5)

- Intuition:
  - $(\mathbf{A}^T \mathbf{A})\ \mathbf{v'}$
  - $(\mathbf{A}^T \mathbf{A})^k\ \mathbf{v'}$

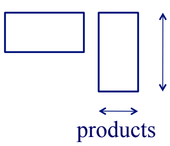
similarities to Smith

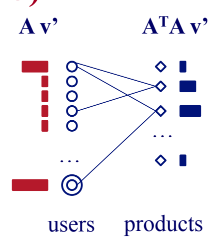




Property (B5)

- Intuition:
  - $(\mathbf{A}^T \mathbf{A})\ \mathbf{v'}$
  - $(\mathbf{A}^T \mathbf{A})^k\ \mathbf{v'}$





Property (B5)

- Intuition:
  - $(\mathbf{A}^T \mathbf{A})\ \mathbf{v'}$ what Smith's 'friends' like
  - $(\mathbf{A}^T \mathbf{A})^k\ \mathbf{v'}$ what k-step-away-friends like

(ie., after k steps, we get what everybody likes, and Smith's initial opinions don't count)

Less obvious properties - repeated:

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{T}_{[r \times m]}$

B(1): $\mathbf{A}_{[n \times m]} (\mathbf{A}^T)_{[m \times n]} = \mathbf{U} \mathbf{\Lambda}^2 \mathbf{U}^T$
B(2): $(\mathbf{A}^T)_{[m \times n]} \mathbf{A}_{[n \times m]} = \mathbf{V} \mathbf{\Lambda}^2 \mathbf{V}^T$
B(3): $((\mathbf{A}^T)_{[m \times n]} \mathbf{A}_{[n \times m]})^k = \mathbf{V} \mathbf{\Lambda}^{2k} \mathbf{V}^T$
B(4): $(\mathbf{A}^T \mathbf{A})^k \sim v_1 \lambda_1^{2k} v_1^T$
B(5): $(\mathbf{A}^T \mathbf{A})^k \mathbf{v'} \sim (\text{constant})\, \mathbf{v}_1$

Least obvious properties

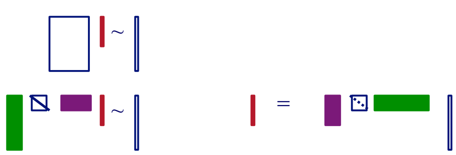
A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{T}_{[r \times m]}$

C(1): $\mathbf{A}_{[n \times m]} \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$
let $\mathbf{x}_0 = \mathbf{V} \mathbf{\Lambda}^{(-1)} \mathbf{U}^T \mathbf{b}$
if under-specified, $\mathbf{x}_0$ gives 'shortest' solution
if over-specified, it gives the 'solution' with the smallest least squares error

(see Num. Recipes, p. 62)

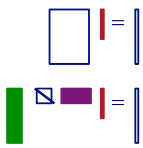
Least obvious properties

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{T}_{[r \times m]}$
C(1): $\mathbf{A}_{[n \times m]} \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$
let $\mathbf{x}_0 = \mathbf{V} \mathbf{\Lambda}^{(-1)} \mathbf{U}^T \mathbf{b}$



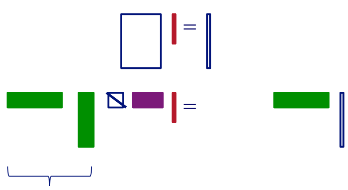
CMU SCS

Slowly:



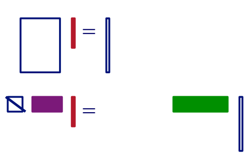
CMU SCS

Slowly:



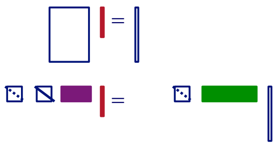
CMU SCS

Slowly:



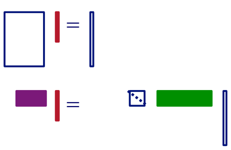
CMU SCS

Slowly:



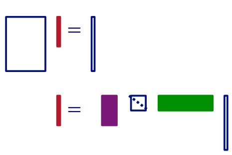
CMU SCS

Slowly:



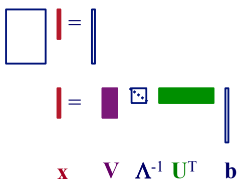
CMU SCS

Slowly:



CMU SCS

Slowly:

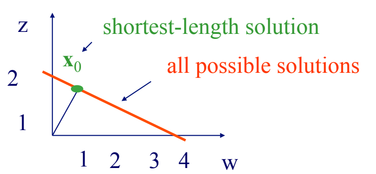


CMU SCS

Least obvious properties

Illustration: under-specified, eg
[1 2] [w z]^T = 4 (ie, 1 w + 2 z = 4)

A=??
b=??



CMU SCS

Exercise

Verify formula:

A = [1 2] **b** = [4]
A = **U Λ V**^{T}
U = ??
Λ = ??
V= ??
$\mathbf{x}_0 = \mathbf{V}\,\mathbf{\Lambda}^{(-1)}\,\mathbf{U}^T\,\mathbf{b}$

CMU SCS

Exercise

Verify formula:

$\mathbf{A} = [1\ 2]$ $\mathbf{b} = [4]$

$\mathbf{A} = \mathbf{U}\, \mathbf{\Lambda}\, \mathbf{V}^T$

$\mathbf{U} = [1]$

$\mathbf{\Lambda} = [\ \mathrm{sqrt}(5)\ ]$

$\mathbf{V} = [\ 1/\mathrm{sqrt}(5) \quad 2/\mathrm{sqrt}(5)\ ]^T$

$\mathbf{x_0} = \mathbf{V}\, \mathbf{\Lambda}^{(-1)}\, \mathbf{U}^T\, \mathbf{b}$

15-826 Copyright: C. Faloutsos (2013) 43

CMU SCS

Exercise

Verify formula:

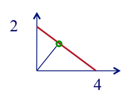
$\mathbf{A} = [1\ 2]$ $\mathbf{b} = [4]$

$\mathbf{A} = \mathbf{U}\, \mathbf{\Lambda}\, \mathbf{V}^T$

$\mathbf{U} = [1]$

$\mathbf{\Lambda} = [\ \mathrm{sqrt}(5)\ ]$

$\mathbf{V} = [\ 1/\mathrm{sqrt}(5) \quad 2/\mathrm{sqrt}(5)\ ]^T$



$\mathbf{x_0} = \mathbf{V}\, \mathbf{\Lambda}^{(-1)}\, \mathbf{U}^T\, \mathbf{b} = [\ 1/5 \quad 2/5]^T\ [4]$

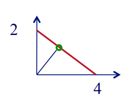
$= [4/5 \quad 8/5]^T$: $w = 4/5$, $z = 8/5$

15-826 Copyright: C. Faloutsos (2013) 44

CMU SCS

Exercise

Verify formula:



Show that $w = 4/5$, $z = 8/5$ is

(a) A solution to $1*w + 2*z = 4$ and

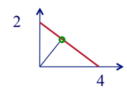
(b) Minimal (wrt Euclidean norm)

15-826 Copyright: C. Faloutsos (2013) 45

CMU SCS

Exercise

Verify formula:



Show that w= 4/5, z = 8/5 is

(a) A solution to 1*w + 2*z = 4 and

A: easy

(b) Minimal (wrt Euclidean norm)

A: [4/5 8/5] is perpenticular to [2 -1]

15-826 46

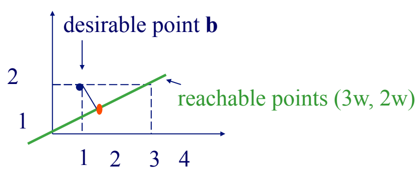
CMU SCS

Least obvious properties – cont'd

Illustration: over-specified, eg

$[3\ 2]^T\ [w] = [1\ 2]^T$ (ie, 3 w = 1; 2 w = 2)

A=??
b=??



15-826 47

CMU SCS

Exercise

Verify formula:

$\mathbf{A} = [3\ 2]^T \quad \mathbf{b} = [\ 1\ \ 2]^T$

$\mathbf{A} = \mathbf{U}\, \boldsymbol{\Lambda}\, \mathbf{V}^T$

$\mathbf{U}$ = ??

$\boldsymbol{\Lambda}$ = ??

$\mathbf{V}$ = ??

$\mathbf{x_0} = \mathbf{V}\, \boldsymbol{\Lambda}^{(-1)}\ \mathbf{U}^T\, \mathbf{b}$

15-826 48

CMU SCS

Exercise

Verify formula:

$\mathbf{A} = [3\ 2]^T$ $\mathbf{b} = [\,1\ \ 2]^T$

$\mathbf{A} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{V}^T$

$\mathbf{U} = [\ 3/\mathrm{sqrt}(13)\quad 2/\mathrm{sqrt}(13)\ ]^T$

$\mathbf{\Lambda} = [\ \mathrm{sqrt}(13)\ ]$

$\mathbf{V} = [\ 1\ ]$

$\mathbf{x_0} = \mathbf{V}\,\mathbf{\Lambda}^{(-1)}\ \mathbf{U}^T\,\mathbf{b} = [\ 7/13\ ]$

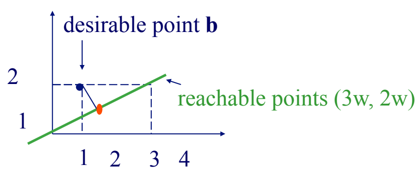
CMU SCS

Exercise

Verify formula:

$[3\ 2]^T\ \ [7/13] = [1\ 2]^T$

$[21/13\ \ 14/13\ ]^T$ -> ‘red point’



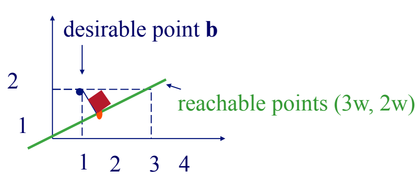
CMU SCS

Exercise

Verify formula:

$[3\ 2]^T\ \ [7/13] = [1\ 2]^T$

$[21/13\ \ 14/13\ ]^T$ -> ‘red point’ - perpenticular?

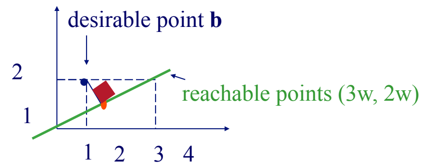


CMU SCS

Exercise

Verify formula:

A: [3 2] . ([1 2] – [21/13 14/13]) =
[3 2] . [-8/13 12/13] = [3 2] . [-2 3] = 0



CMU SCS

Least obvious properties - cont’d

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \boldsymbol{\Lambda}_{[r \times r]} \mathbf{V}^{\mathbf{T}}_{[r \times m]}$

C(2): $\mathbf{A}_{[n \times m]} \mathbf{v_1}_{[m \times 1]} = \lambda_1 \mathbf{u_1}_{[n \times 1]}$
where $\mathbf{v}_1$, $\mathbf{u}_1$ the first (column) vectors of **V**, **U**. ($\mathbf{v}_1$ == right-singular-vector)
C(3): symmetrically: $\mathbf{u}_1^{\mathbf{T}} \mathbf{A} = \lambda_1 \mathbf{v}_1^{\mathbf{T}}$
$\mathbf{u}_1$ == left-singular-vector
Therefore:

CMU SCS

Least obvious properties - cont’d

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \boldsymbol{\Lambda}_{[r \times r]} \mathbf{V}^{\mathbf{T}}_{[r \times m]}$

$\boxed{\text{C(4): } \mathbf{A}^{\mathbf{T}} \mathbf{A}\, \mathbf{v}_1 = \lambda_1^2\, \mathbf{v}_1}$
(**fixed point** - the dfn of eigenvector for a symmetric matrix)

Least obvious properties - altogether

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{\mathbf{T}}_{[r \times m]}$

C(1): $\mathbf{A}_{[n \times m]} \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$
then, $\mathbf{x}_0 = \mathbf{V} \mathbf{\Lambda}^{(-1)} \mathbf{U}^T \mathbf{b}$**:** shortest, actual or least-squares solution

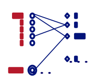
C(2): $\mathbf{A}_{[n \times m]} \mathbf{v}_{\mathbf{1}\,[m \times 1]} = \lambda_1 \mathbf{u}_{\mathbf{1}\,[n \times 1]}$

C(3): $\mathbf{u}_1^{\mathbf{T}} \mathbf{A} = \lambda_1 \mathbf{v}_1^{\mathbf{T}}$

C(4): $\mathbf{A}^{\mathbf{T}} \mathbf{A} \mathbf{v}_1 = \lambda_1^2 \mathbf{v}_1$

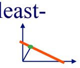
Properties - conclusions

A(0): $\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \mathbf{\Lambda}_{[r \times r]} \mathbf{V}^{\mathbf{T}}_{[r \times m]}$



B(5): $(\mathbf{A}^T \mathbf{A})^k \mathbf{v}' \sim (\text{constant})\, \mathbf{v}_1$

C(1): $\mathbf{A}_{[n \times m]} \mathbf{x}_{[m \times 1]} = \mathbf{b}_{[n \times 1]}$
then, $\mathbf{x}_0 = \mathbf{V} \mathbf{\Lambda}^{(-1)} \mathbf{U}^T \mathbf{b}$**:** shortest, actual or least-squares solution



C(4): $\mathbf{A}^{\mathbf{T}} \mathbf{A} \mathbf{v}_1 = \lambda_1^2 \mathbf{v}_1$

SVD - detailed outline

- ...
- Case studies
- SVD properties
- more case studies
  - Kleinberg/google algorithms
  - query feedbacks
- Conclusions

CMU SCS

Kleinberg's algo (HITS)

Kleinberg, Jon (1998). *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms.

CMU SCS

Kleinberg's algorithm

- Problem dfn: given the web and a query
- find the most 'authoritative' web pages for this query

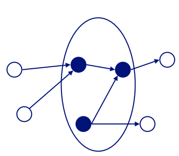
Step 0: find all pages containing the query terms

Step 1: expand by one move forward and backward

CMU SCS

Kleinberg's algorithm

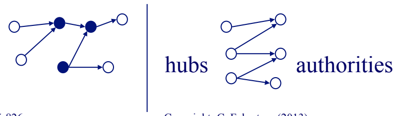
- Step 1: expand by one move forward and backward



CMU SCS

Kleinberg’s algorithm

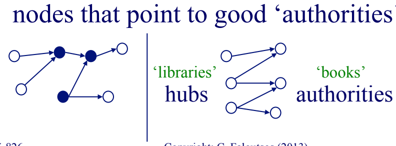
- on the resulting graph, give high score (= ‘authorities’) to nodes that many important nodes point to
- give high importance score (‘hubs’) to nodes that point to good ‘authorities’)



CMU SCS

Kleinberg’s algorithm

- on the resulting graph, give high score (= ‘authorities’) to nodes that many important nodes point to
- give high importance score (‘hubs’) to nodes that point to good ‘authorities’)



CMU SCS

Kleinberg’s algorithm

observations

- recursive definition!
- each node (say, ‘i’-th node) has both an authoritativeness score a_i and a hubness score h_i

CMU SCS

Kleinberg's algorithm

Let E be the set of edges and **A** be the adjacency matrix:

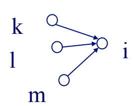
the (i,j) is 1 if the edge from i to j exists

Let h and a be [n x 1] vectors with the 'hubness' and 'authoritativiness' scores.

Then:

CMU SCS

Kleinberg's algorithm



Then:

$$a_i = h_k + h_l + h_m$$

that is

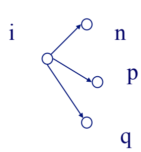
a_i = Sum (h_j) over all j that (j,i) edge exists

or

$\mathbf{a} = \mathbf{A}^T \mathbf{h}$

CMU SCS

Kleinberg's algorithm



symmetrically, for the 'hubness':

$$h_i = a_n + a_p + a_q$$

that is

h_i = Sum (q_j) over all j that (i,j) edge exists

or

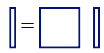
$\mathbf{h} = \mathbf{A}\ \mathbf{a}$

Kleinberg's algorithm

In conclusion, we want vectors **h** and **a** such that:

$$\mathbf{h} = \mathbf{A}\ \mathbf{a}$$

$$\mathbf{a} = \mathbf{A}^T\ \mathbf{h}$$



Recall properties:

C(2): $\mathbf{A}_{[n \times m]}\ \mathbf{v_1}_{[m \times 1]} = \lambda_1\ \mathbf{u_1}_{[n \times 1]}$

C(3): $\mathbf{u_1}^T \mathbf{A} = \lambda_1\ \mathbf{v_1}^T$

Kleinberg's algorithm

In short, the solutions to

$$\mathbf{h} = \mathbf{A}\ \mathbf{a}$$

$$\mathbf{a} = \mathbf{A}^T\ \mathbf{h}$$

are the left- and right- singular-vectors of the adjacency matrix **A.**

Starting from random **a'** and iterating, we'll eventually converge

(Q: to which of all the singular-vectors? why?)

Kleinberg's algorithm

(Q: to which of all the singular-vectors? why?)

A: to the ones of the strongest singular-value, because of property B(5):

B(5): $(\mathbf{A}^T \mathbf{A})^k\ \mathbf{v'} \sim (\text{constant})\ \mathbf{v}_1$

Kleinberg's algorithm - results

Eg., for the query 'java':

0.328 www.gamelan.com

0.251 java.sun.com

0.190 www.digitalfocus.com ("the java developer")

Kleinberg's algorithm - discussion

- 'authority' score can be used to find 'similar pages' (how?)
- closely related to 'citation analysis', social networs / 'small world' phenomena

SVD - detailed outline

- ...
- Case studies
- SVD properties
- more case studies
  - Kleinberg/google algorithms
  - query feedbacks
- Conclusions

CMU SCS

PageRank (google)

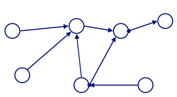
Larry Page   Sergey Brin

- Brin, Sergey and Lawrence Page (1998). *Anatomy of a Large-Scale Hypertextual Web Search Engine*. 7th Intl World Wide Web Conf.

CMU SCS

Problem: PageRank

Given a directed graph, find its most interesting/central node



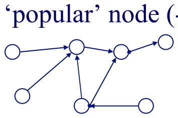
A node is important, if it is connected with important nodes (recursive, but OK!)

CMU SCS

Problem: PageRank - solution

Given a directed graph, find its most interesting/central node

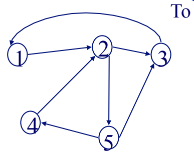
Proposed solution: Random walk; spot most 'popular' node (-> steady state prob. (ssp))



A node has high ssp, if it is connected with high ssp nodes (recursive, but OK!)

(Simplified) PageRank algorithm

- Let **A** be the adjacency matrix;
- let **B** be the transition matrix: transpose, column-normalized - then

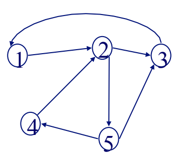


From / To

$$\mathbf{B}\quad \begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix} \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix} = \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}$$

(Simplified) PageRank algorithm

- **B p = p**



$$\underset{\mathbf{B}}{\begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix}} \underset{\mathbf{p}}{\begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}} = \underset{\mathbf{p}}{\begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}}$$

(Simplified) PageRank algorithm

- **B p** = 1 * **p**
- thus, **p** is the **eigenvector** that corresponds to the highest eigenvalue (=1, since the matrix is column-normalized)
- Why does such a **p** exist?
  - **p** exists if **B** is nxn, nonnegative, irreducible [Perron–Frobenius theorem]

CMU SCS

(Simplified) PageRank algorithm

- **B p** = 1 * **p**
- thus, **p** is the **eigenvector(*)** that corresponds to the highest eigenvalue (=1, since the matrix is column-normalized)
- Why does such a **p** exist?
  - **p** exists if **B** is nxn, nonnegative, irreducible [Perron–Frobenius theorem]

All<-> all

(*) dfn: a few foils later

 79

CMU SCS

(Simplified) PageRank algorithm

- In short: imagine a particle randomly moving along the edges
- compute its steady-state probabilities (ssp)

Full version of algo: with occasional random jumps

Why? To make the matrix irreducible

15-826 80

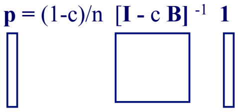
CMU SCS

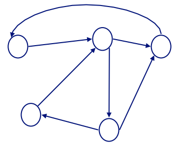
Full Algorithm

- With probability *1-c*, fly-out to a random node
- Then, we have

$\mathbf{p} = c\, \mathbf{B}\, \mathbf{p} + (1-c)/n\ \mathbf{1} \Rightarrow$

$\mathbf{p} = (1-c)/n\ [\mathbf{I} - c\, \mathbf{B}]^{-1}\ \mathbf{1}$

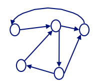




15-826 81

Full Algorithm

- With probability *1-c*, fly-out to a random node
- Then, we have



$\mathbf{p} = c\, \mathbf{B}\, \mathbf{p} + (1\text{-}c)/n\ \mathbf{1} \Rightarrow$

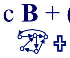
$\mathbf{p} = (1\text{-}c)/n\ [\mathbf{I} - c\, \mathbf{B}]^{-1}\ \mathbf{1}$

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Alternative notation – eigenvector viewpoint

M Modified transition matrix

$\mathbf{M} = c\, \mathbf{B} + (1\text{-}c)/n\ \mathbf{1}\ \mathbf{1}^T$



CLIQUE

Then

$\mathbf{p} = \mathbf{M}\, \mathbf{p}$

That is: the steady state probabilities =

PageRank scores form the *first eigenvector* of the 'modified transition matrix'

Parenthesis: intuition behind eigenvectors

- Definition
- 3 properties
- intuition

Formal definition

If **A** is a (n x n) square matrix
(λ , **x**) is an **eigenvalue/eigenvector** pair of **A** if

$$\boxed{\mathbf{A}\ \mathbf{x} = \lambda\ \mathbf{x}}$$

CLOSELY related to singular values:

Property #1: Eigen- vs singular-values

if

$$\mathbf{B}_{[n \times m]} = \mathbf{U}_{[n \times r]}\ \mathbf{\Lambda}_{[r \times r]}\ (\mathbf{V}_{[m \times r]})^T$$

then $\mathbf{A} = (\mathbf{B}^T\mathbf{B})$ is symmetric and

C(4): $\mathbf{B}^T\ \mathbf{B}\ \mathbf{v}_i = \lambda_i^2\ \mathbf{v}_i$

ie, $\mathbf{v}_1$, $\mathbf{v}_2$, ...: eigenvectors of $\mathbf{A} = (\mathbf{B}^T\mathbf{B})$

Property #2

- If $\mathbf{A}_{[nxn]}$ is a real, symmetric matrix
- Then it has *n* real eigenvalues

(if **A** is not symmetric, some eigenvalues may be complex)

CMU SCS

Property #3

- If $\mathbf{A}_{[nxn]}$ is a real, symmetric matrix
- Then it has *n* real eigenvalues
- And they agree with its *n* singular values, except possibly for the sign

CMU SCS

Parenthesis: intuition behind eigenvectors

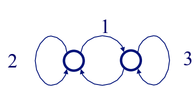
- Definition
- 3 properties
- **intuition**

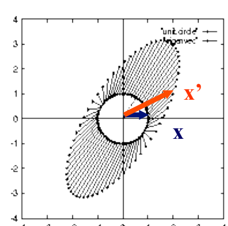
CMU SCS

Intuition

- **A** as vector transformation

$$\underset{\mathbf{x'}}{\begin{bmatrix} 2 \\ 1 \end{bmatrix}} = \underset{\mathbf{A}}{\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}$$



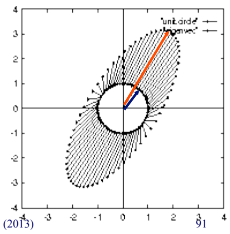


CMU SCS

Intuition

- By defn., eigenvectors remain parallel to themselves (‘**fixed points**’)

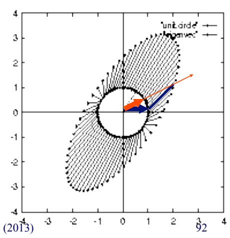
$$\underset{\lambda_1}{3.62} * \underset{\mathbf{v_1}}{\begin{bmatrix}0.52\\0.85\end{bmatrix}} = \underset{\mathbf{A}}{\begin{bmatrix}2 & 1\\1 & 3\end{bmatrix}} \underset{\mathbf{v_1}}{\begin{bmatrix}0.52\\0.85\end{bmatrix}}$$



CMU SCS

Convergence

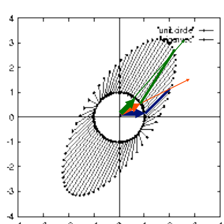
- Usually, fast:



CMU SCS

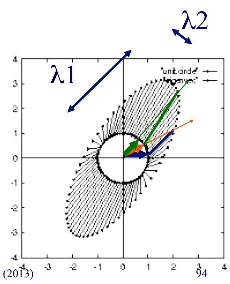
Convergence

- Usually, fast:



Convergence

- Usually, fast:
- depends on ratio
  $\lambda 1 : \lambda 2$



Closing the parenthesis wrt intuition behind eigenvectors

Kleinberg/PageRank - conclusions

SVD helps in graph analysis:

hub/authority scores: strongest left- and right-singular-vectors of the adjacency matrix

random walk on a graph: steady state probabilities are given by the strongest eigenvector of the transition matrix

CMU SCS

SVD - detailed outline

- ...
- Case studies
- SVD properties
- more case studies
  - google/Kleinberg algorithms
  - query feedbacks
- Conclusions

CMU SCS

Query feedbacks

[Chen & Roussopoulos, sigmod 94]

Sample problem:

estimate selectivities (e.g., '*how many movies were made between 1940 and 1945*?'

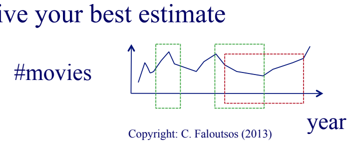
for query optimization,

LEARNING from the query results so far!!

CMU SCS

Query feedbacks

- Given: past queries and their results
  - #movies(1925,1935) = 52
  - #movies(1948, 1990) = 123
  - …
  - And a new query, say #movies(1979,1980)?
- Give your best estimate



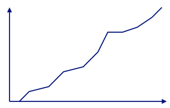
CMU SCS

Query feedbacks

Idea #1: consider a function for the CDF (cummulative distr. function), eg., 6-th degree polynomial (or splines, or anything else)

PDF

count, so far



year

CMU SCS

Query feedbacks

For example

F(x) = # movies made until year ‘x’

$= a_1 + a_2 * x + a_3 * x^2 + \ldots a_7 * x^6$

CMU SCS

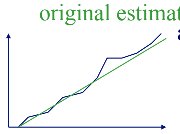
Query feedbacks

GREAT idea #2: adapt your model, as you see the actual counts of the actual queries

original estimate

actual

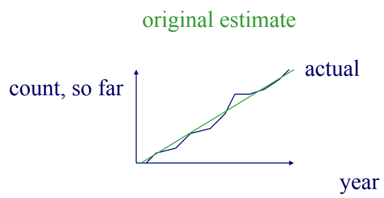
count, so far



year

CMU SCS

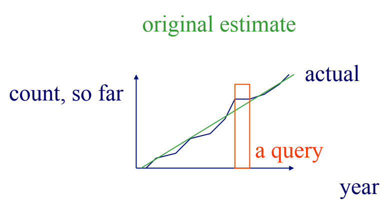
Query feedbacks



15-826 Copyright: C. Faloutsos (2013) 103

CMU SCS

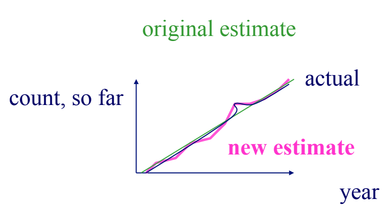
Query feedbacks



15-826 Copyright: C. Faloutsos (2013) 104

CMU SCS

Query feedbacks



15-826 Copyright: C. Faloutsos (2013) 105

CMU SCS

Query feedbacks

Eventually, the problem becomes:

- estimate the parameters a_1, ... a_7 of the model
- to minimize the least squares errors from the real answers so far.

Formally:

15-826 Copyright: C. Faloutsos (2013) 106

CMU SCS

Query feedbacks

Formally, with *n* queries and 6-th degree polynomials:

$$\begin{bmatrix} X11 & X12 & & & X17 \\ & & & & \\ & & & & \\ Xn1 & Xn2 & & & Xn7 \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ \\ \\ a7 \end{bmatrix} = \begin{bmatrix} b1 \\ b2 \\ \\ \\ bn \end{bmatrix}$$

15-826 Copyright: C. Faloutsos (2013) 107

CMU SCS

Query feedbacks

where $x_{i,j}$ such that Sum ($x_{i,j}$ * a_i) = our estimate for the # of movies and b_j: the actual

$$\begin{bmatrix} X11 & X12 & & & X17 \\ & & & & \\ & & & & \\ Xn1 & Xn2 & & & Xn7 \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ \\ \\ a7 \end{bmatrix} = \begin{bmatrix} b1 \\ b2 \\ \\ \\ bn \end{bmatrix}$$

15-826 Copyright: C. Faloutsos (2013) 108

Query feedbacks

For example, for query '*find the count of movies during (1920-1932)*':

$a_1 + a_2 * 1932 + a_3 * 1932**2 + \ldots$

-

$(a_1 + a_2 * 1920 + a_3 * 1920**2 + \ldots)$

$$\begin{bmatrix} X11 & X12 & & & X17 \\ & & & & \\ & & & & \\ & & & & \\ Xn1 & Xn2 & & & Xn7 \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ \\ \\ a7 \end{bmatrix} = \begin{bmatrix} b1 \\ b2 \\ \\ \\ bn \end{bmatrix}$$

Query feedbacks

And thus X11 = 0; X12 = 1932-1920, etc

$a_1 + a_2 * 1932 + a_3 * 1932**2 + \ldots$

-

$(a_1 + a_2 * 1920 + a_3 * 1920**2 + \ldots)$

$$\begin{bmatrix} X11 & X12 & & & X17 \\ & & & & \\ & & & & \\ & & & & \\ Xn1 & Xn2 & & & Xn7 \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ \\ \\ a7 \end{bmatrix} = \begin{bmatrix} b1 \\ b2 \\ \\ \\ bn \end{bmatrix}$$

Query feedbacks

In matrix form:

X **a** = **b**

$$\begin{array}{l} \text{1st query} \\ \\ \\ \\ \text{n-th query} \end{array} \begin{bmatrix} X11 & X12 & & & X17 \\ & & & & \\ & & & & \\ & & & & \\ Xn1 & Xn2 & & & Xn7 \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ \\ \\ a7 \end{bmatrix} = \begin{bmatrix} b1 \\ b2 \\ \\ \\ bn \end{bmatrix}$$

CMU SCS

Query feedbacks

In matrix form:

$$\mathbf{X}\,\mathbf{a} = \mathbf{b}$$

and the least-squares estimate for **a** is

$$\mathbf{a} = \mathbf{V}\,\mathbf{\Lambda}^{(-1)}\,\mathbf{U}^{\mathbf{T}}\,\mathbf{b}$$

according to property C(1)

(let $\mathbf{X} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{V}^{\mathbf{T}}$)

15-826 Copyright: C. Faloutsos (2013) 112

CMU SCS

Query feedbacks - enhancements

The solution

$$\mathbf{a} = \mathbf{V}\,\mathbf{\Lambda}^{(-1)}\,\mathbf{U}^{\mathbf{T}}\,\mathbf{b}$$

works, but needs expensive SVD each time a new query arrives

GREAT Idea #3: Use 'Recursive Least Squares', to adapt **a** incrementally.

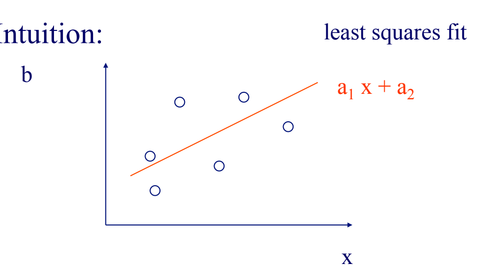
Details: in paper - intuition:

15-826 Copyright: C. Faloutsos (2013) 113

CMU SCS

Query feedbacks - enhancements

Intuition:

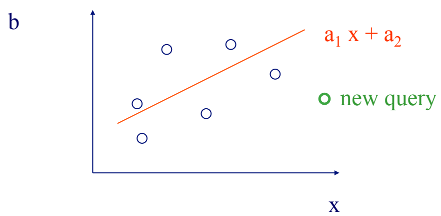


15-826 Copyright: C. Faloutsos (2013) 114

CMU SCS

Query feedbacks - enhancements

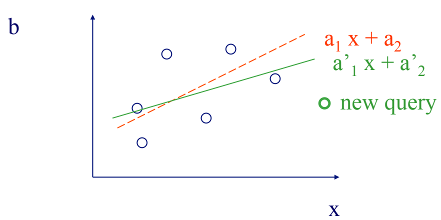
Intuition: least squares fit



CMU SCS

Query feedbacks - enhancements

Intuition: least squares fit



CMU SCS

Query feedbacks - enhancements

the new coefficients can be quickly computed from the old ones, plus statistics in a (7x7) matrix

(no need to know the details, although the RLS is a brilliant method)

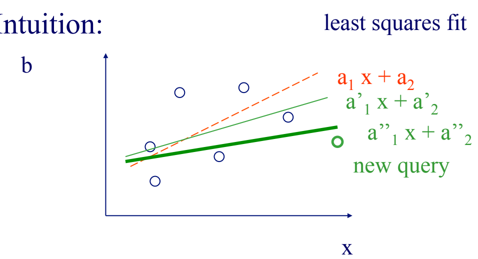
Query feedbacks - enhancements

GREAT idea #4: 'forgetting' factor - we can even down-play the weight of older queries, since the data distribution might have changed.

(comes for 'free' with RLS...)

Query feedbacks - enhancements

Intuition:



Query feedbacks - conclusions

SVD helps find the Least Squares solution, to adapt to query feedbacks

(RLS = Recursive Least Squares is a great method to incrementally update least-squares fits)

CMU SCS

SVD - detailed outline

- ...
- Case studies
- SVD properties
- more case studies
  - google/Kleinberg algorithms
  - query feedbacks
- Conclusions

CMU SCS

Conclusions

- SVD: a **valuable** tool
- given a document-term matrix, it finds 'concepts' (LSI)
- ... and can reduce dimensionality (KL)
- ... and can find rules (PCA; RatioRules)

CMU SCS

Conclusions cont'd

- ... and can find fixed-points or steady-state probabilities (google/ Kleinberg/ Markov Chains)
- ... and can solve optimally over- and under-constraint linear systems (least squares / query feedbacks)

CMU SCS

References

- Brin, S. and L. Page (1998). Anatomy of a Large-Scale Hypertextual Web Search Engine. 7th Intl World Wide Web Conf.
- Chen, C. M. and N. Roussopoulos (May 1994). Adaptive Selectivity Estimation Using Query Feedback. Proc. of the ACM-SIGMOD , Minneapolis, MN.

CMU SCS

References cont’d

- Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms.
- Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in C, Cambridge University Press.