Large Graph Mining: Power Tools and a Practitioner's guide

Task 5: Graphs over time & tensors

Faloutsos, Miller, Tsourakakis

CMU

Outline

- Introduction – Motivation
- Task 1: Node importance
- Task 2: Community detection
- Task 3: Recommendations
- Task 4: Connection sub-graphs
- **Task 5: Mining graphs over time**
- Task 6: Virus/influence propagation
- Task 7: Spectral graph theory
- Task 8: Tera/peta graph mining: hadoop
- Observations – patterns of real graphs
- Conclusions

Thanks to

- Tamara Kolda (Sandia)

for the foils on tensor definitions, and on TOPHITS

Detailed outline

- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining

Examples of Matrices: Authors and terms

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Motivation: Why tensors?

- Q: what is a tensor?

Motivation: Why tensors?

- A: N-D generalization of matrix:

KDD'09	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Motivation: Why tensors?

- A: N-D generalization of matrix:

KDD'07

KDD'08

KDD'09

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Tensors are useful for 3 or more modes

Terminology: 'mode' (or 'aspect'):

Mode#3

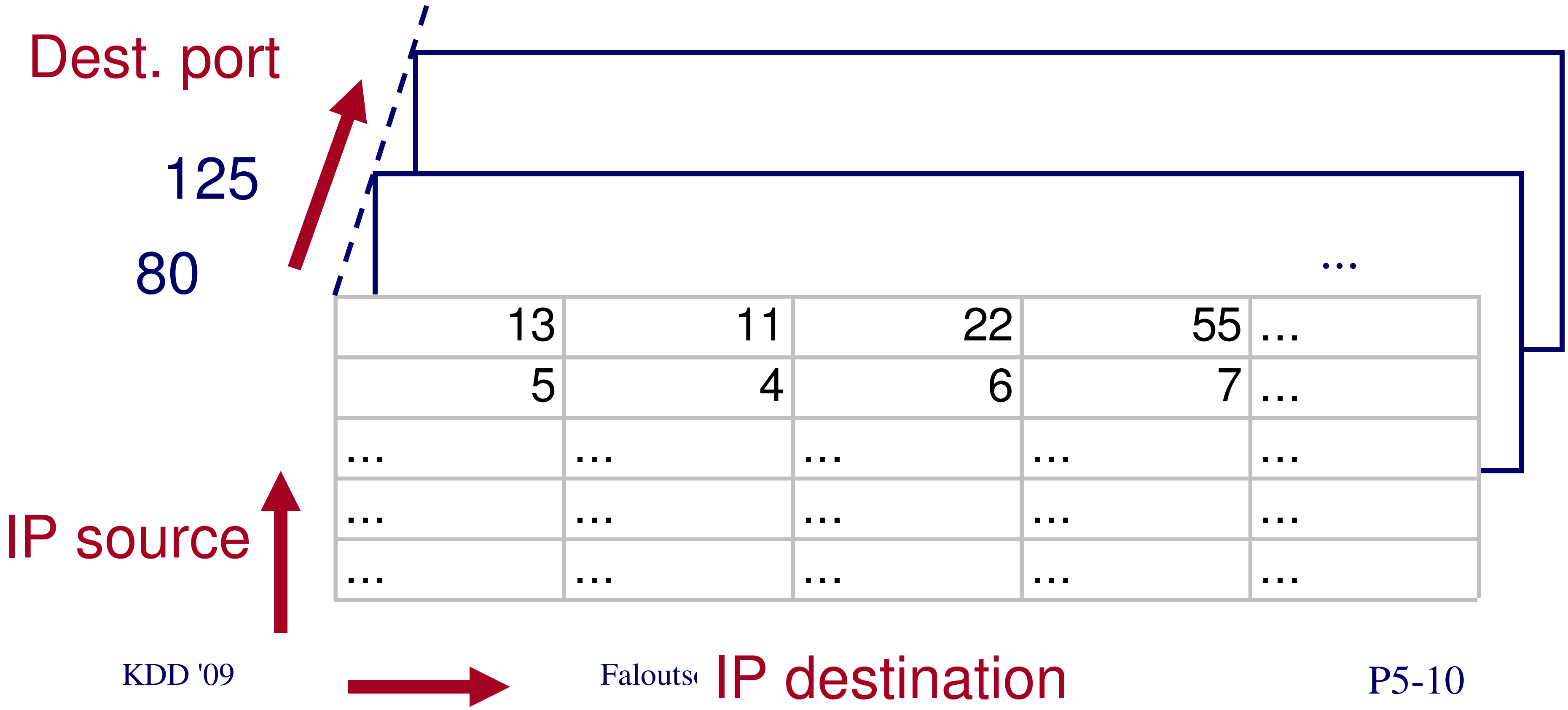
Mode#2

data	mining	classif.	tree	...
13	11	22	55	...
5	4	6	7	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Mode (== aspect) #1

Notice

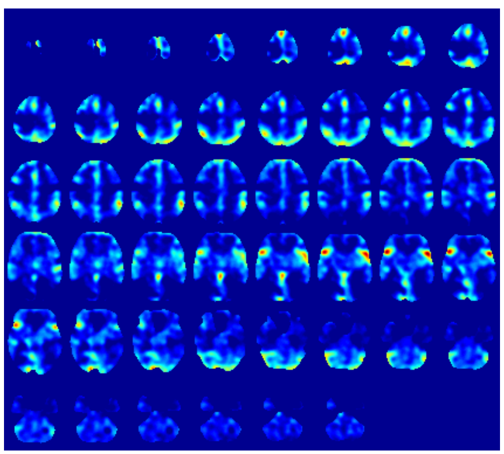
- 3^{rd} mode does not need to be time
- we can have more than 3 modes



13	11	22	55	...
5	4	6	7	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Notice

- 3rd mode does not need to be time
- we can have more than 3 modes
  - Eg, fFMRI: x,y,z, time, person-id, task-id



From DENLAB, Temple U.
(Prof. V. Megalooikonomou +)

http://denlab.temple.edu/bidms/cgi-bin/browse.cgi

Motivating Applications

- Why tensors are useful?
  - web mining (TOPHITS)
  - environmental sensors
  - Intrusion detection (src, dst, time, dest-port)
  - Social networks (src, dst, time, type-of-contact)
  - face recognition
  - etc …

Detailed outline

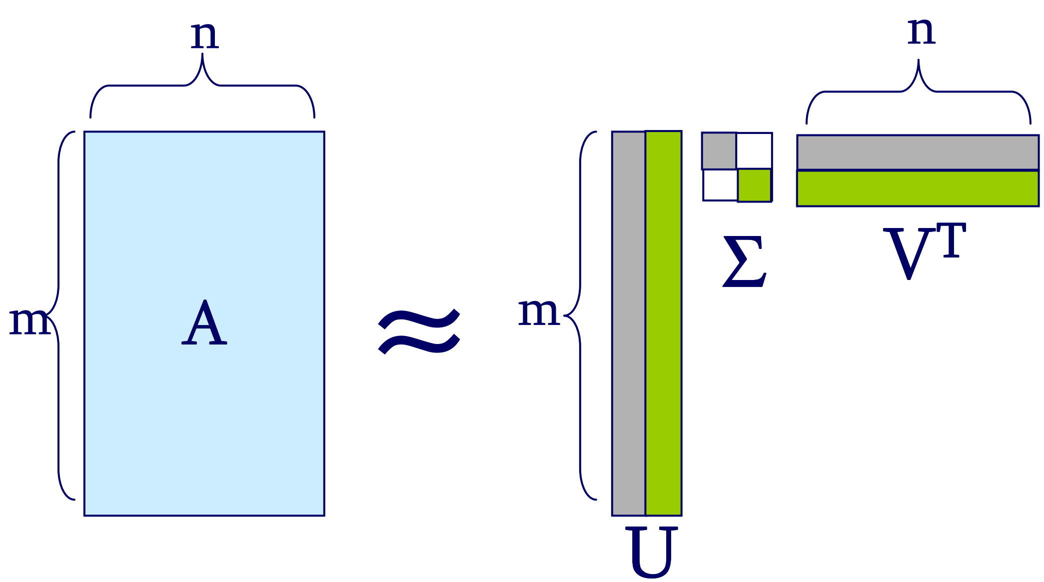
- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining

Tensor basics

- Multi-mode extensions of SVD – recall that:

Reminder: SVD

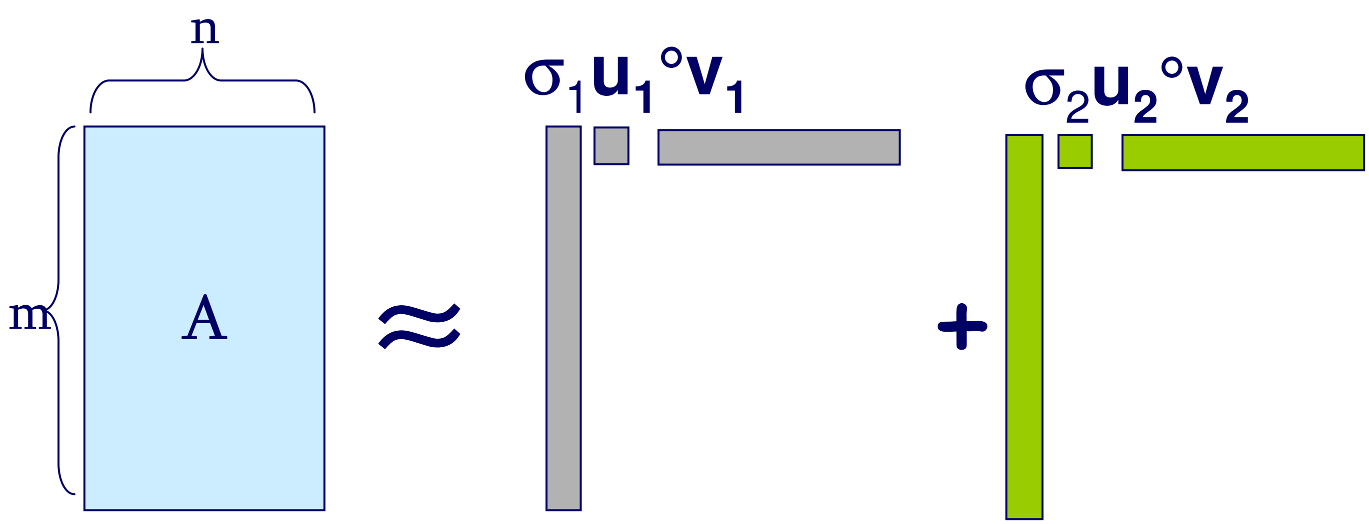
$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



– Best rank-k approximation in L2

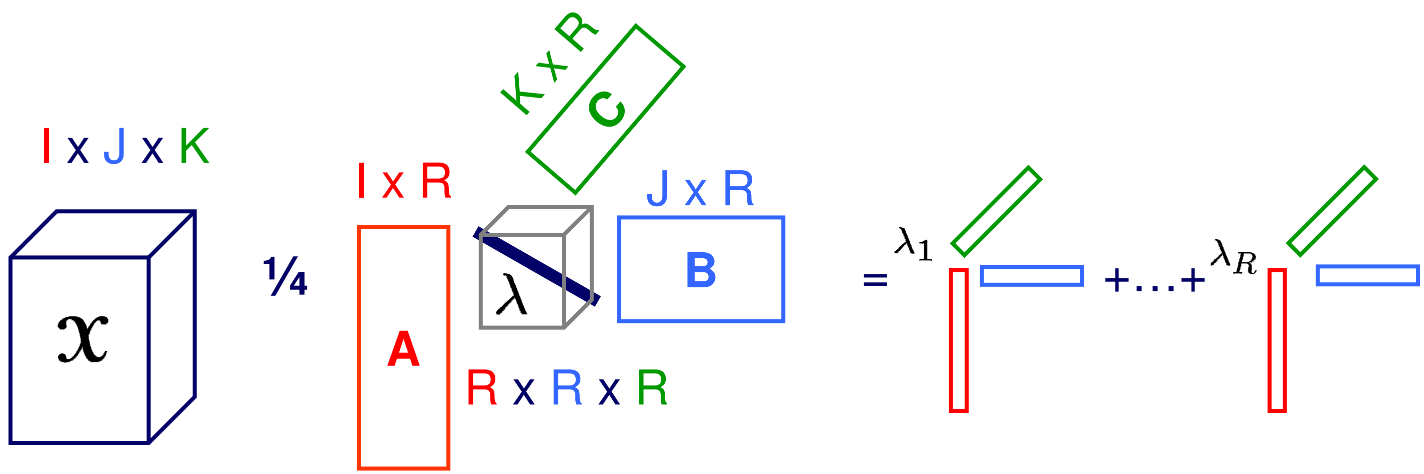
Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



– Best rank-k approximation in L2

Goal: extension to >=3 modes



$$\mathcal{X} \approx [\![\lambda\ ; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r\ \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Main points:

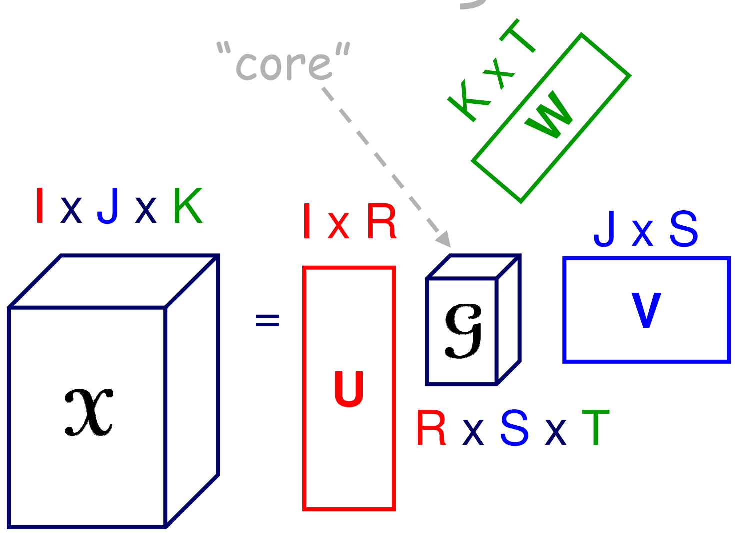
- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ``alternating least squares'' (ALS)

Specially Structured Tensors

- Tucker Tensor

$$\begin{aligned}\mathcal{X} &= \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} \\ &= \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t \\ &\equiv [\![\mathcal{G} \,;\, \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]\end{aligned}$$

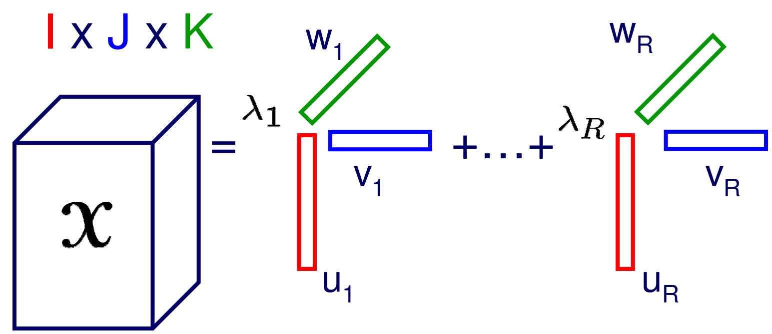
Our Notation



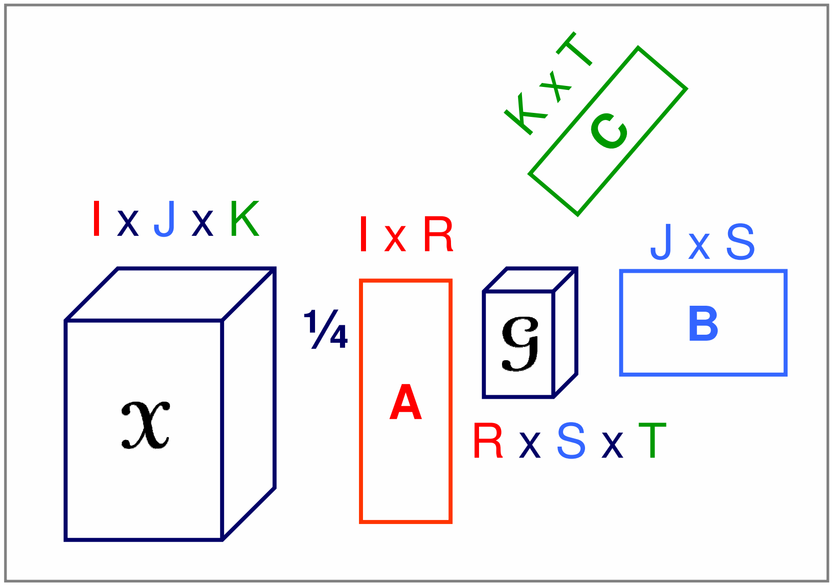
- Kruskal Tensor

$$\begin{aligned}\mathcal{X} &= \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \\ &\equiv [\![\lambda \,;\, \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]\end{aligned}$$

Our Notation



Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: conf. x conf-group
- **G**: how groups relate to each other

Intuition behind core tensor

- 2-d case: co-clustering
- [Dhillon et al. Information-Theoretic Co-clustering, KDD'03]

$$
\begin{array}{cc}
 & n \\
m & \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}
\end{array}
$$

eg, terms x documents

$$
\begin{array}{cc}
 & k \\
m & \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix}
\end{array}
\begin{array}{cc}
 & l \\
k & \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}
\end{array}
\begin{array}{cc}
 & n \\
l & \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix}
\end{array}
=
\begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}
$$

med. doc

cs doc

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

med. terms

cs terms

common terms

term group x doc. group

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \left[\begin{array}{ccc|ccc} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ \hline .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{array}\right]$$

term x term-group

doc x doc group

Tensor tools - summary

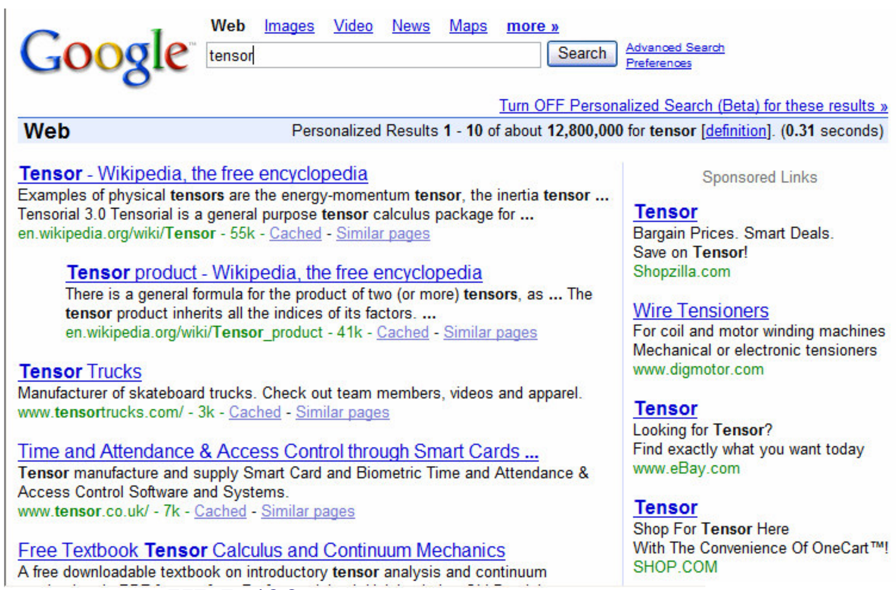
- Two main tools
  - PARAFAC
  - Tucker
- Both find row-, column-, tube-groups
  - but in PARAFAC the three groups are identical
- (To solve: Alternating Least Squares)

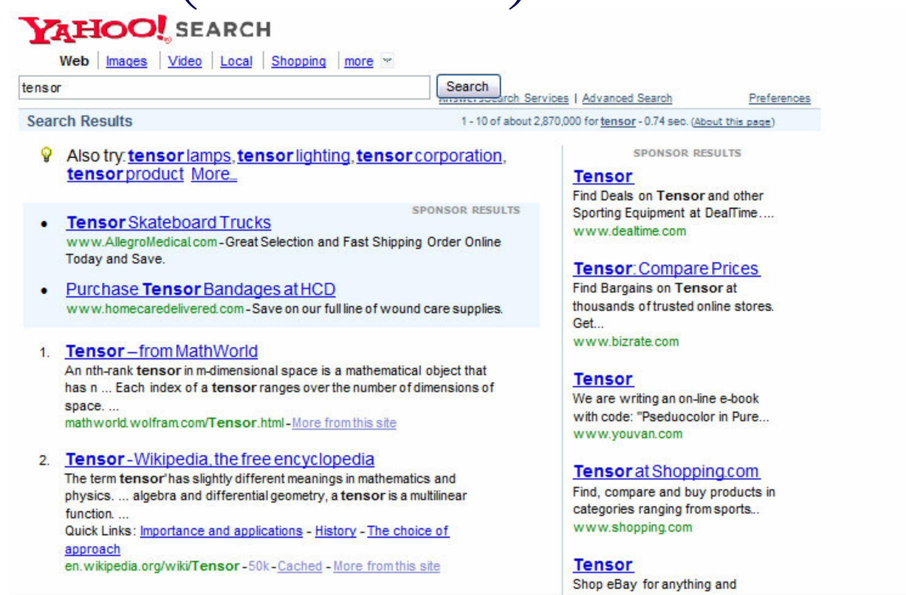
Detailed outline

- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining

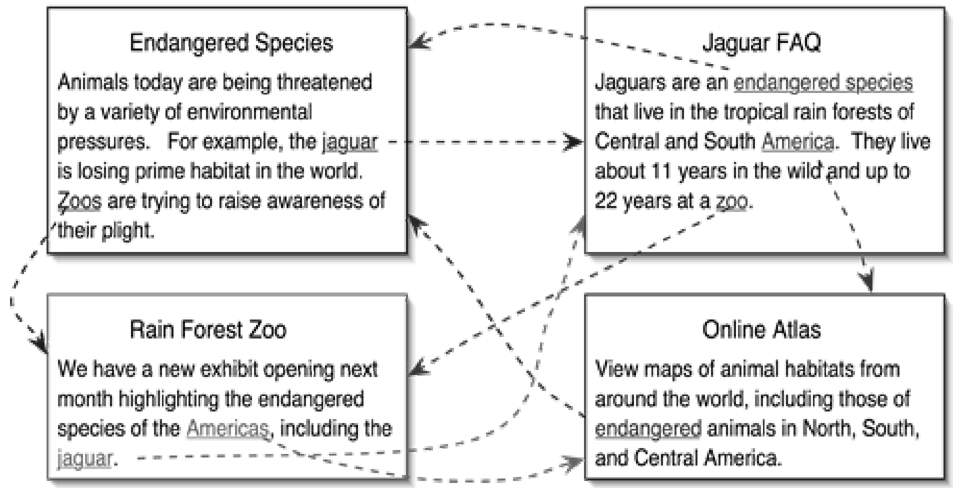
Web graph mining

- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)





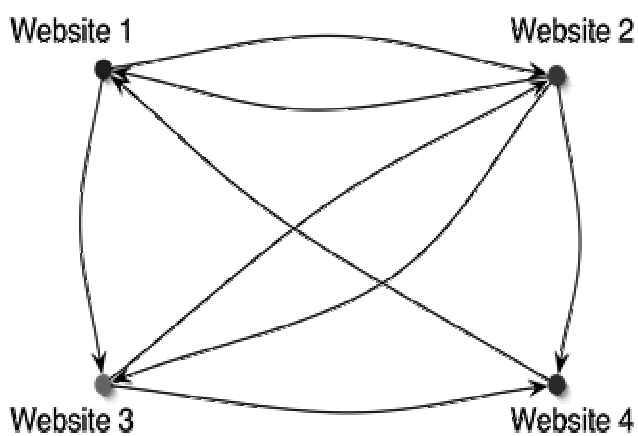
Kleinberg's Hubs and Authorities (the HITS method)

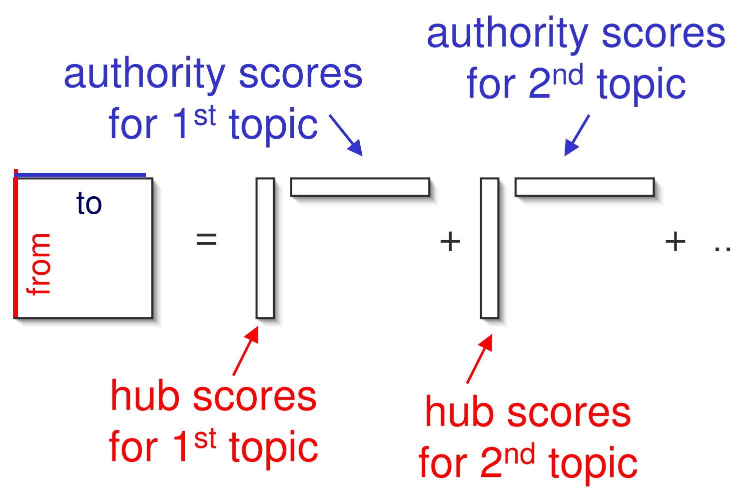


Sparse adjacency matrix and its SVD:

$$x_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{X} \approx \sum_r \sigma_r \, \mathbf{h}_r \circ \mathbf{a}_r$$





Kleinberg, JACM, 1999

HITS Authorities on Sample Data

1st Principal Factor	
.97	www.ibm.com
.24	www.alphaw
.08	www-128.ibm
.05	www.develop
.02	www.researc
.01	www.redbook
.01	news.com.c

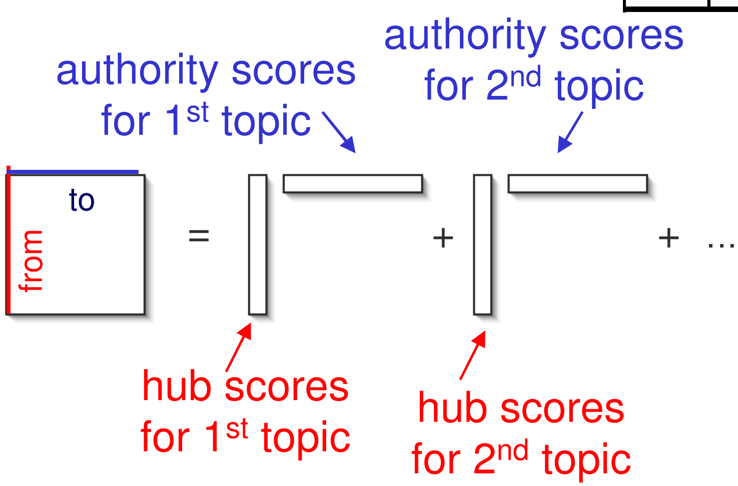
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh
.06	www.lehigha
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.
.02	lewisweb.cc.
.02	www.leo.lehi
.02	www.distanc
.02	fp1.cc.lehigh

3rd Principal Factor	
.75	java.sun.com
.38	www.sun.com
.36	developers.sun.
.24	see.sun.com
.16	www.samag.co
.13	docs.sun.com
.12	blogs.sun.com
.08	sunsolve.sun.c
.08	www.sun-catal
.08	news.com.com

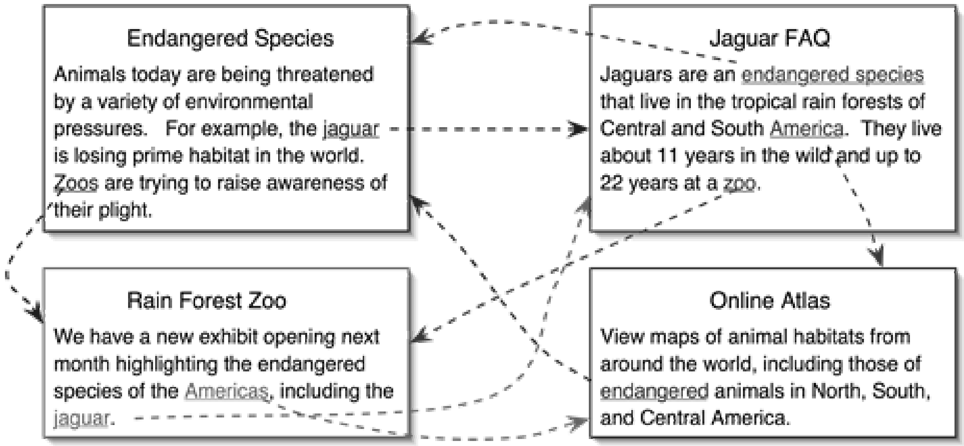
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.
.13	www.usdoj

6th Principal Factor	
.97	mathpost.asu.edu
.18	math.la.asu.edu
.17	www.asu.edu
.04	www.act.org
.03	www.eas.asu.edu
.02	archives.math.utk.edu
.02	www.geom.uiuc.edu
.02	www.fulton.asu.edu
.02	www.amstat.org
.02	www.maa.org

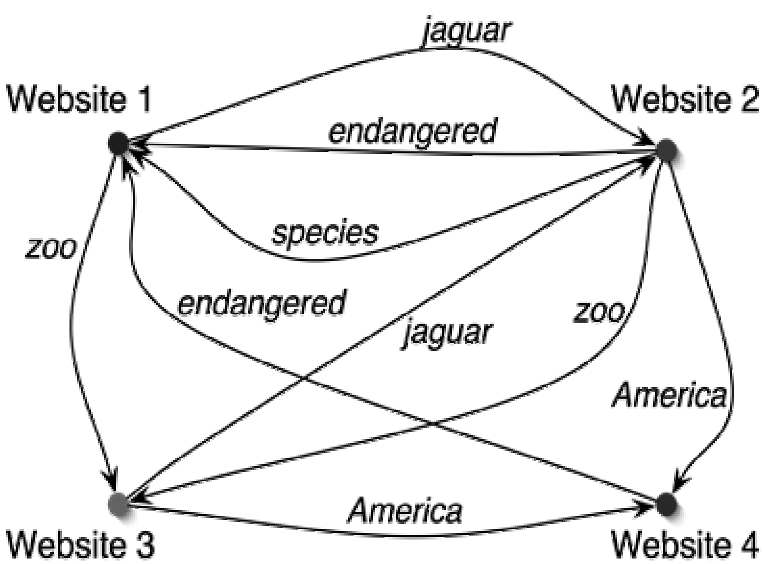
We started our crawl from http://www-neos.mcs.anl.gov/neos, and crawled 4700 pages, resulting in 560 cross-linked hosts.



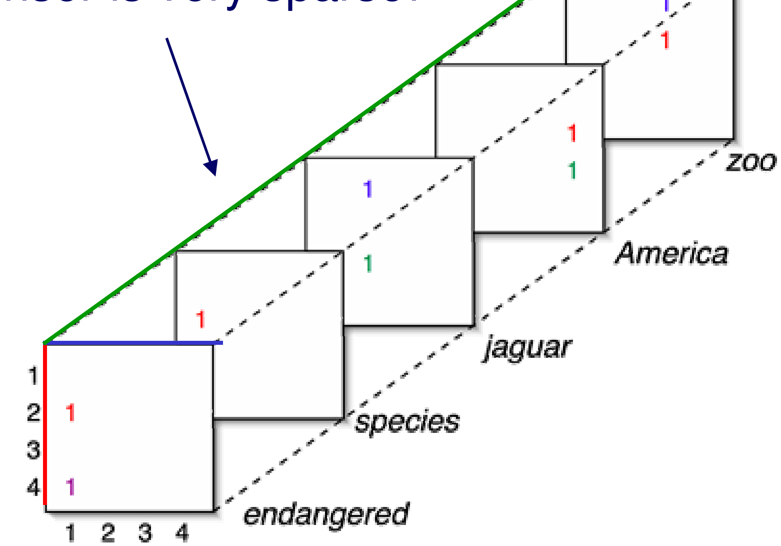
Three-Dimensional View of the Web



$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

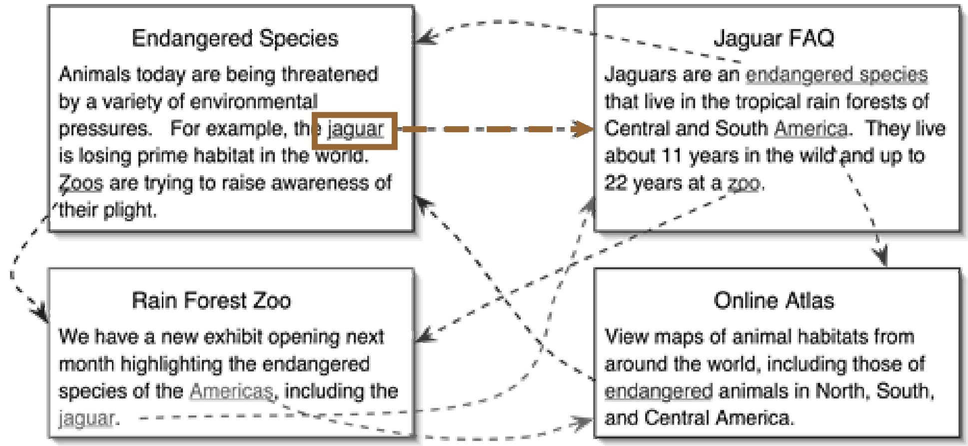


Observe that this tensor is very sparse!

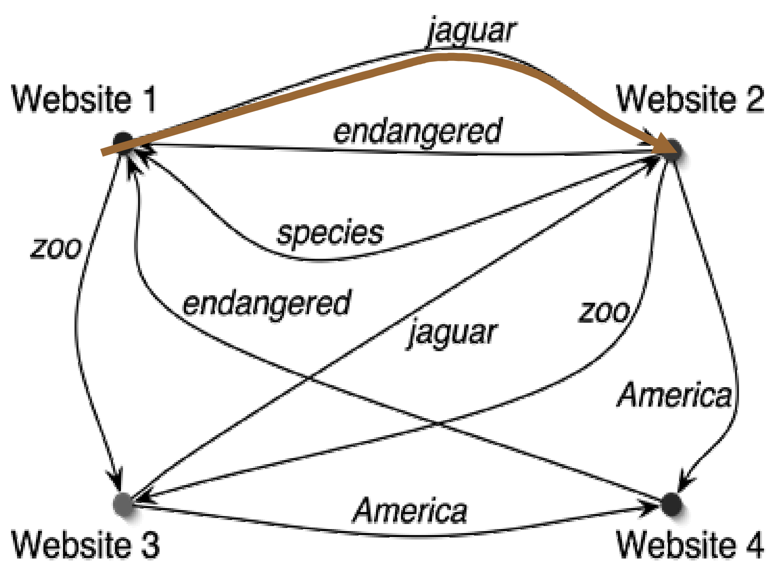


Kolda, Bader, Kenny, ICDM05

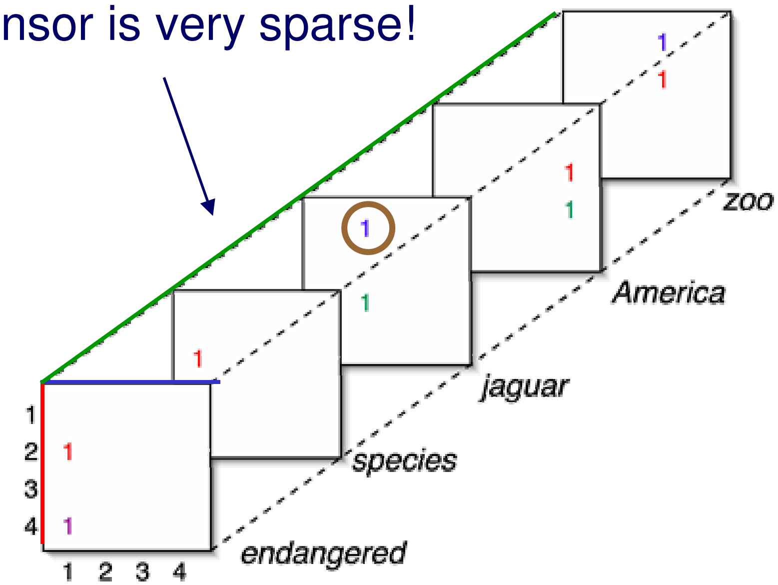
Three-Dimensional View of the Web



$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

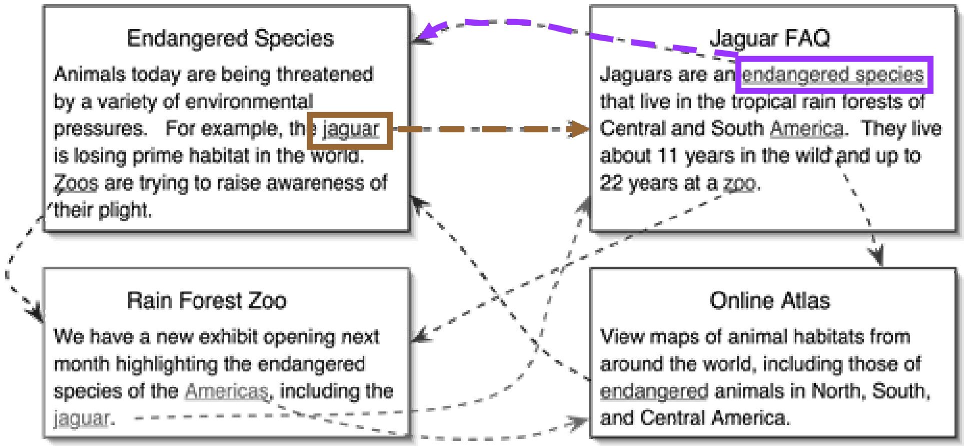


Observe that this tensor is very sparse!

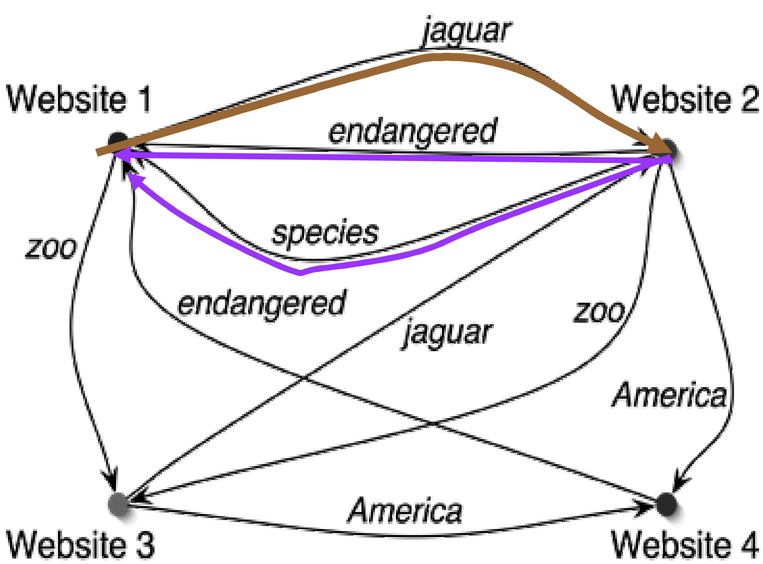


Kolda, Bader, Kenny, ICDM05

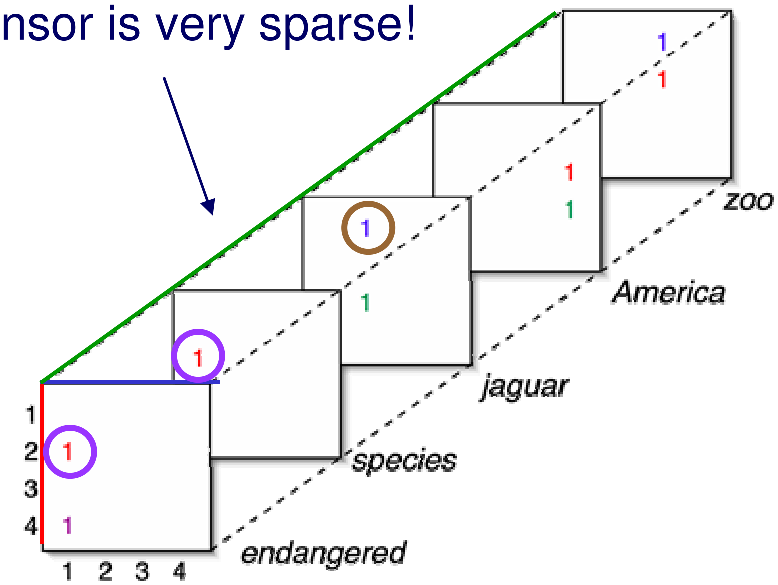
Three-Dimensional View of the Web



$$x_{ijk} = \begin{cases} 1 & \text{if page } i \longrightarrow \text{page } j \\ & \text{with term } k \\ 0 & \text{otherwise} \end{cases}$$



Observe that this tensor is very sparse!

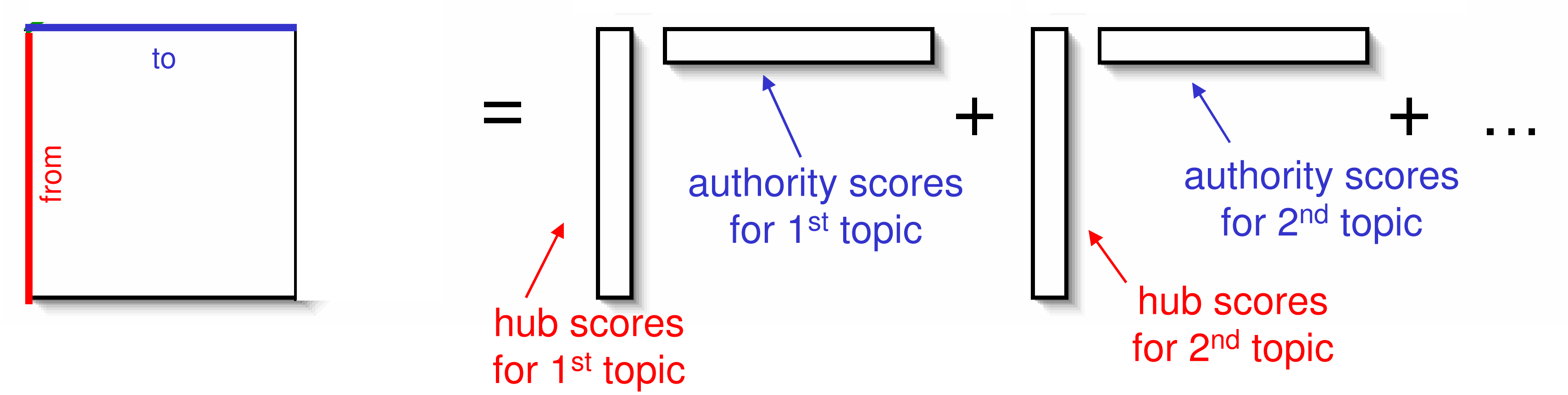


Kolda, Bader, Kenny, ICDM05

Topical HITS (TOPHITS)

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

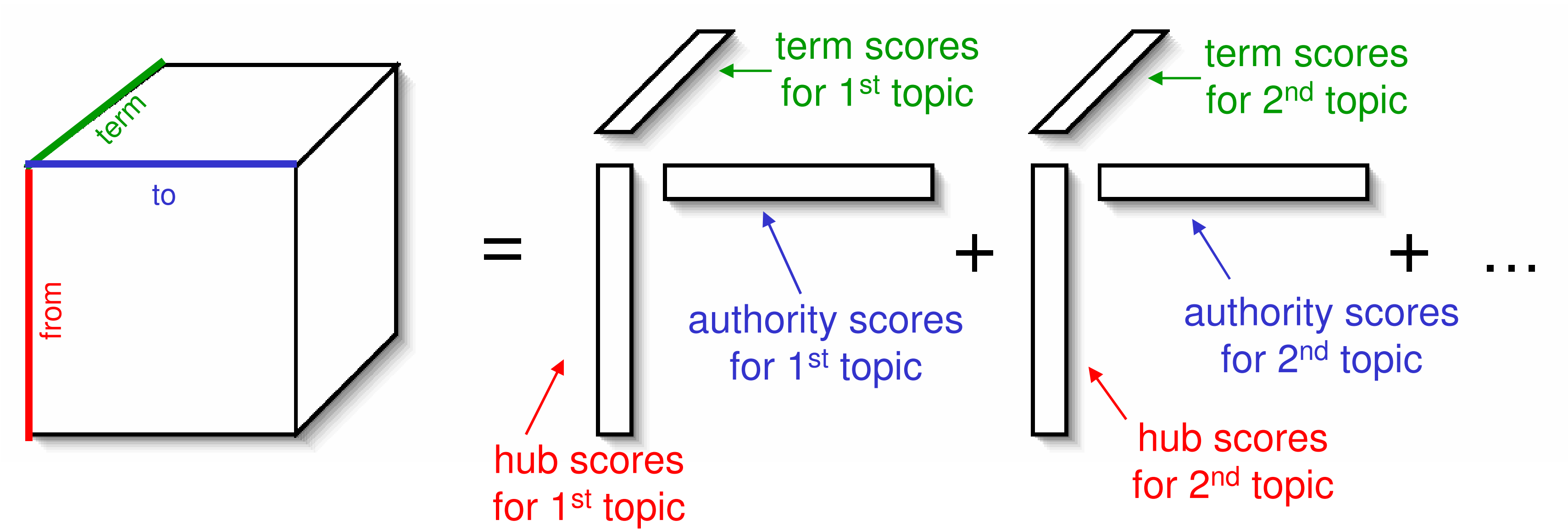
$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



Topical HITS (TOPHITS)

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r \circ \mathbf{t}_r$$



TOPHITS Terms & Authorities on Sample Data

1st Principal Factor

.23	JAVA	.86	java.sun.com
.18	SUN	.38	developers.sun.com
.17	PLATF		
.16	SOLAR		
.16	DEVEL		
.15	EDITIO		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-RE		

2nd Principal Factor

.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACUL		
.16	SEARC		
.16	NEWS		
.16	LIBRA		
.16	COMP		
.12	LEHIG		

3rd Principal Factor

.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM	.18	www.alphaworks.ibm.com
.12	SERVI		
.12	WEBS		
.12	WEB		
.11	DEVEL		
.11	LINUX		
.11	RESOU		
.11	TECHN		
.10	DOWN		

4th Principal Factor

.26	INFORMATION	.87	www.pueblo.gsa.gov
.24	FEDERAL	.24	www.irs.gov
.23	CITIZE		
.22	OTHER		
.19	CENTE		
.19	LANGU		
.15	U.S		
.15	PUBLI		
.14	CONS		
.13	FREE		

6th Principal Factor

.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHIT		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRES		
.11	OFFIC		

12th Principal Factor

.75	OPTIMIZATION	.35	www.palisade.com
.58	SOFTWARE	.35	www.solver.com
.08	DECIS		
.07	NEOS		
.06	TREE		
.05	GUIDE		
.05	SEARC		
.05	ENGIN		
.05	CONTR		
.05	ILOG		

13th Principal Factor

.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACRO		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

16th Principal Factor

.50	WEATHER	.81	www.weather.gov
.24	OFFICE	.41	www.spc.noaa.gov
.23	CENTER	.30	lwf.ncdc.noaa.gov
.19	NO-RE		
.17	ORGA		
.15	NWS		
.15	SEVER		
.15	FIRE		
.15	POLIC		
.14	CLIMA		

19th Principal Factor

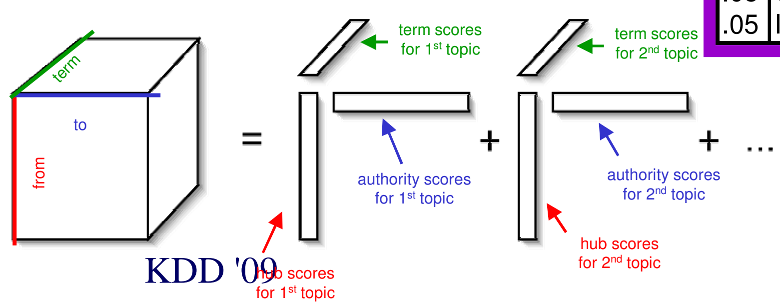
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

TOPHITS uses 3D analysis to find the dominant groupings of web pages <u>and</u> terms.

$$x_{ijk} = \begin{cases} \frac{1}{\log(w_k)+1} & \text{if } i \rightarrow j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

w_k = # unique links using term k

Tensor PARAFAC



KDD '09

Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Tensors provide elegant theory and algorithms
  - PARAFAC and Tucker: discover groups

References

- T. G. Kolda, B. W. Bader and J. P. Kenny. *Higher-Order Web Link Analysis Using Multilinear Algebra*. In: ICDM 2005, Pages 242-249, November 2005.
- Jimeng Sun, Spiros Papadimitriou, Philip Yu. *Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams*, Proc. of the Int. Conf. on Data Mining (ICDM), Hong Kong, China, Dec 2006

Resources

- See tutorial on tensors, KDD'07 (w/ Tamara Kolda and Jimeng Sun):

www.cs.cmu.edu/~christos/TALKS/KDD-07-tutorial

Tensor tools - resources

- Toolbox: from Tamara Kolda:

csmr.ca.sandia.gov/~tgkolda/TensorToolbox

• T. G. Kolda and B. W. Bader. ***Tensor Decompositions and Applications***. SIAM Review, Volume 51, Number 3, September 2009
csmr.ca.sandia.gov/~tgkolda/pubs/bibtgkfiles/TensorReview-preprint.pdf

• T. Kolda and J. Sun: Scalable Tensor Decomposition for Multi-Aspect Data Mining (ICDM 2008)