CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #21: Tensor decompositions

C. Faloutsos

CMU SCS

Must-read Material

- Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. Technical Report SAND2007-6702, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, November 2007

2

CMU SCS

Outline

Goal: 'Find similar / interesting things'

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

3

CMU SCS

Indexing - Detailed outline

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

4

CMU SCS

Most of foils by

- Dr. Tamara Kolda (Sandia N.L.)
- csmr.ca.sandia.gov/~tgkolda

- Dr. Jimeng Sun (CMU -> IBM)
- www.cs.cmu.edu/~jimeng

3h tutorial: www.cs.cmu.edu/~christos/TALKS/SDM-tut-07/

5

CMU SCS

Outline

- Motivation - Definitions
- Tensor tools
- Case studies

6

CMU SCS

Motivation 0: Why "matrix"?

- Why matrices are important?

CMU SCS

Examples of Matrices: Graph - social network

	John	Peter	Mary	Nick	...
John	0	11	22	55	...
Peter	5	0	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

CMU SCS

Examples of Matrices: cloud of n-d points

	chol#	blood#	age	..	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

CMU SCS

Examples of Matrices: Market basket

- **market basket** as in Association Rules

	milk	bread	choc.	wine	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

10

CMU SCS

Examples of Matrices: Documents and terms

	data	mining	classif.	tree	...
Paper#1	13	11	22	55	...
Paper#2	5	4	6	7	...
Paper#3	...	...	...	...	...
Paper#4	...	...	...	...	...
...	...	...	...	...	...

11

CMU SCS

Examples of Matrices: Authors and terms

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

12

CMU SCS

Examples of Matrices:
sensor-ids and time-ticks

	temp1	temp2	humid.	pressure	...
t1	13	11	22	55	...
t2	5	4	6	7	...
t3	...	...	...	...	...
t4	...	...	...	...	...
...	...	...	...	...	...

13

CMU SCS

Motivation: Why tensors?

- Q: what is a tensor?

14

CMU SCS

Motivation 2: Why tensor?

- A: N-D generalization of matrix:

KDD'07	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

15

Motivation 2: Why tensor?

- A: N-D generalization of matrix:

KDD'05
KDD'06
KDD'07

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

16

Tensors are useful for 3 or more modes

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

Mode#3

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

Mode#2

Mode (== aspect) #1

17

Motivating Applications

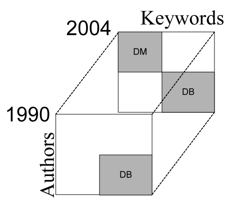
- Why matrices are important?
- Why tensors are useful?
  - P1: social networks
  - P2: web mining

18

CMU SCS

P1: Social network analysis

- Traditionally, people focus on static networks and find community structures
- We plan to monitor the change of the community structure over time



19

CMU SCS

P2: Web graph mining

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

20

CMU SCS

Outline

- Motivation – Definitions
- Tensor tools
  - Tensor Basics
  - Tucker
  - PARAFAC
- Case studies

21

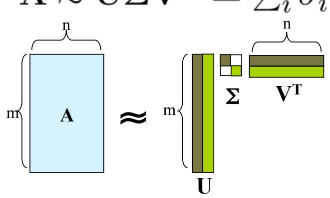
CMU SCS

Tensor Basics

CMU SCS

Reminder: SVD

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



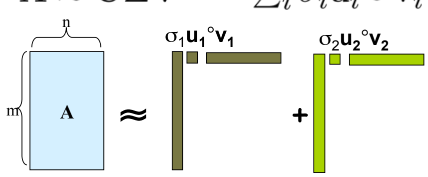
– Best rank-k approximation in L2

23

CMU SCS

Reminder: SVD

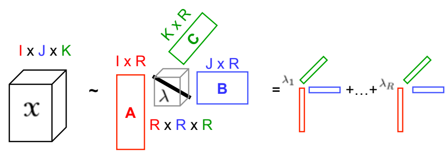
$$\mathbf{A} \approx \mathbf{U \Sigma V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



– Best rank-k approximation in L2

24

Goal: extension to >=3 modes



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

Main points:

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

Specially Structured Tensors

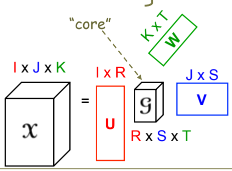
CMU SCS

Specially Structured Tensors

- Tucker Tensor

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

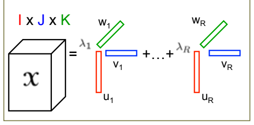
Our Notation



- Kruskal Tensor

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

Our Notation



28

CMU SCS

details

Specially Structured Tensors

- Tucker Tensor

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

In matrix form:

$$\mathbf{X}_{(1)} = \mathbf{U}\mathbf{G}_{(1)}(\mathbf{W} \otimes \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\mathbf{G}_{(2)}(\mathbf{W} \otimes \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\mathbf{G}_{(3)}(\mathbf{V} \otimes \mathbf{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathbf{W} \otimes \mathbf{V} \otimes \mathbf{U})\,\mathrm{vec}(\mathcal{G})$$

- Kruskal Tensor

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

In matrix form:

Let $\mathbf{\Lambda} = \mathrm{diag}(\lambda)$

$$\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{\Lambda}\,(\mathbf{W} \odot \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{\Lambda}\,(\mathbf{W} \odot \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{\Lambda}\,(\mathbf{V} \odot \mathbf{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathbf{W} \odot \mathbf{V} \odot \mathbf{U})\,\lambda$$

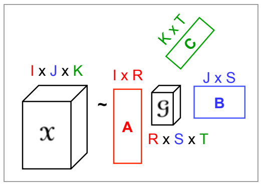
29

CMU SCS

Tensor Decompositions

CMU SCS

Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: conf. x conf-group
- $\mathcal{G}$: how groups relate to each other

Needs elaboration!

31

CMU SCS

Intuition behind core tensor

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

32

CMU SCS

eg, terms x documents

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

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

33

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

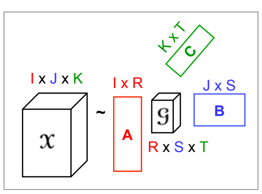
term group x doc. group

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

doc x doc group

term x term-group

Tucker Decomposition



$$\mathcal{X} \approx [\![\mathcal{G} \,; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Given A, B, C, the optimal core is:

$$\mathcal{G} = [\![\mathcal{X} \,; \mathbf{A}^{\dagger}, \mathbf{B}^{\dagger}, \mathbf{C}^{\dagger}]\!]$$

- Proposed by Tucker (1966)
- AKA: Three-mode factor analysis, three-mode PCA, orthogonal array decomposition
- **A**, **B**, and **C** generally assumed to be orthonormal (generally assume they have full column rank)
- $\mathcal{G}$ is <u>not</u> diagonal
- Not unique

Recall the equations for converting a tensor to a matrix

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathrm{vec}(\mathcal{X}) = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathcal{G})$$

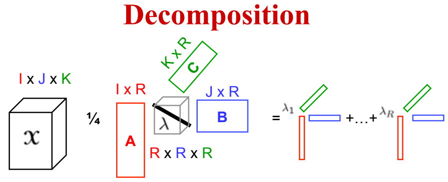
CMU SCS

Outline

- Motivation – Definitions
- Tensor tools
  - Tensor Basics
  - Tucker
  - PARAFAC
- Case studies

CMU SCS

CANDECOMP/PARAFAC Decomposition



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

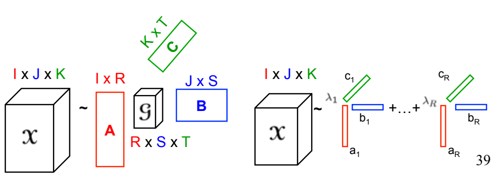
- CANDECOMP = Canonical Decomposition (Carroll & Chang, 1970)
- PARAFAC = Parallel Factors (Harshman, 1970)
- Core is diagonal (specified by the vector λ)
- Columns of **A**, **B**, and **C** are not orthonormal
- If R is *minimal*, then R is called the **rank** of the tensor (Kruskal 1977)
- Can have rank($\mathcal{X}$) > min{I,J,K}

CMU SCS

IMPORTANT

Tucker vs. PARAFAC Decompositions

- Tucker
  - Variable transformation in each mode
  - Core G may be dense
  - A, B, C generally orthonormal
  - Not unique
- PARAFAC
  - Sum of rank-1 components
  - No core, i.e., superdiagonal core
  - A, B, C may have linearly dependent columns
  - Generally unique



CMU SCS

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

- Toolbox: from Tamara Kolda:

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

40

CMU SCS

Outline

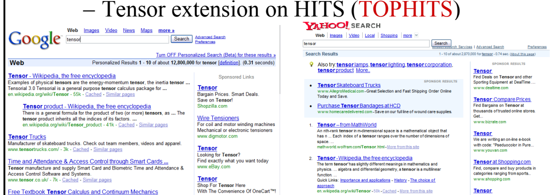
- Motivation - Definitions
- Tensor tools
- Case studies

41

CMU SCS

P1: Web graph mining

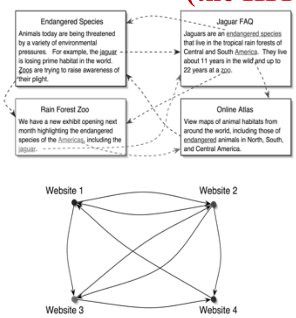
- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)



42

CMU SCS

Kleinberg's Hubs and Authorities (the HITS method)



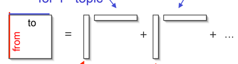
Sparse adjacency matrix and its SVD:

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

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

authority scores for 1st topic

authority scores for 2nd topic



hub scores for 1st topic

hub scores for 2nd topic

43

Kleinberg, JACM, 1999

CMU SCS

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.co

2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh
.06	www.lehighal
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.c
.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-catalo
.08	news.com.com

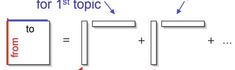
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.g
.13	www.usdoj

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.

authority scores for 1st topic

authority scores for 2nd topic



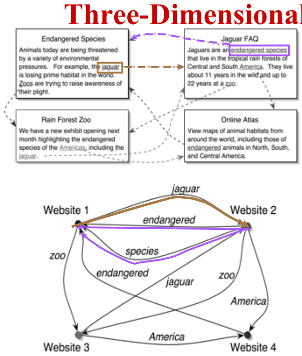
hub scores for 1st topic

hub scores for 2nd topic

44

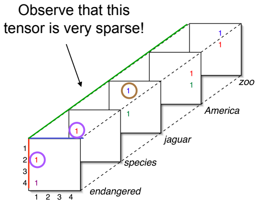
CMU SCS

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!



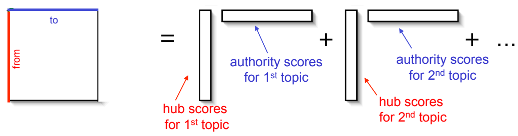
45

Kolda, Bader, Kenny, ICDM05

Topical HITS (TOPHITS)

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

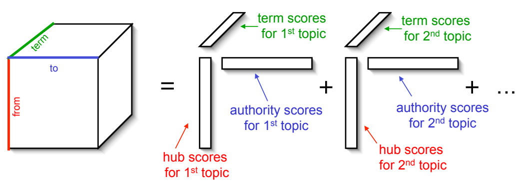
$$\mathfrak{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



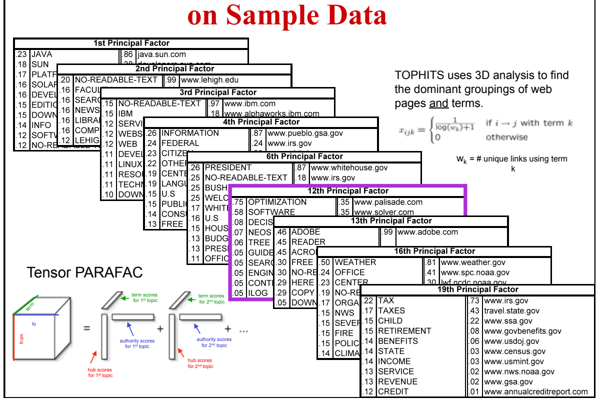
Topical HITS (TOPHITS)

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

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



TOPHITS Terms & Authorities on Sample Data



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

CMU SCS

GigaTensor: Scaling Tensor Analysis Up By 100 Times – Algorithms and Discoveries

U Kang   Evangelos Papalexakis   Abhay Harpale   Christos Faloutsos

School of Computer Science
Carnegie Mellon University

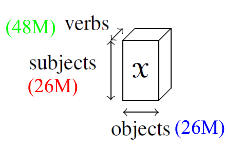
CMU SCS

P2: N.E.L.L. analysis

- NELL: Never Ending Language Learner
  - Q1: dominant concepts / topics?
  - Q2: synonyms for a given new phrase?

"Eric Clapton plays guitar"

"Barrack Obama is the president of U.S."



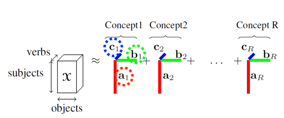
NELL (Never Ending Language Learner)
Nonzeros =144M

50

CMU SCS

A1: Concept Discovery

- Concept Discovery in Knowledge Base



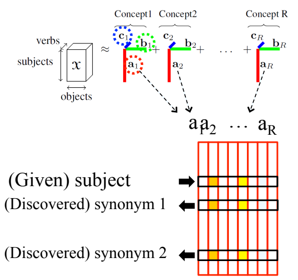
Noun Phrase 1	Noun Phrase 2	Context
Concept 1: "Web Protocol"		
internet	protocol	'np1' 'stream' 'np2'
file	software	'np1' 'marketing' 'np2'
data	suite	'np1' 'dating' 'np2'
Concept 2: "Credit Cards"		
credit	information	'np1' 'card' 'np2'
Credit	debt	'np1' 'report' 'np2'
library	number	'np1' 'cards' 'np2'
Concept 3: "Health System"		
health	provider	'np1' 'care' 'np2'
child	providers	'np' 'insurance' 'np2'
home	system	'np1' 'service' 'np2'
Concept 4: "Family Life"		
life	rest	'np2' 'of' 'my' 'np1'
family	part	'np2' 'of' 'his' 'np1"
body	years	'np2' 'of' 'her' 'np1'

A1: Concept Discovery

Noun Phrase 1	Noun Phrase 2	Context
Concept 1: "Web Protocol"		
internet	protocol	'np1' 'stream' 'np2'
file	software	'np1' 'marketing' 'np2'
data	suite	'np1' 'dating' 'np2'
Concept 2: "Credit Cards"		
credit	information	'np1' 'card' 'np2'
Credit	debt	'np1' 'report' 'np2'
library	number	'np1' 'cards' 'np2'
Concept 3: "Health System"		
health	provider	'np1' 'care' 'np2'
child	providers	'np' 'insurance' 'np2'
home	system	'np1' 'service' 'np2'
Concept 4: "Family Life"		
life	rest	'np2' 'of' 'my' 'np1'
family	part	'np2' 'of' 'his' 'np1"
body	years	'np2' 'of' 'her' 'np1'

A2: Synonym Discovery

- Synonym Discovery in Knowledge Base



(Given) Noun Phrase	(Discovered) Potential Synonyms
pollutants	dioxin, sulfur dioxide, greenhouse gases, particulates, nitrogen oxide, air pollutants, cholesterol
disabilities	infections, dizziness, injuries, diseases, drowsiness, stiffness, injuries
vodafone	verizon, comcast
Christian history	European history, American history, Islamic history, history
disbelief	dismay, disgust, astonishment
cyberpunk	online-gaming
soul	body

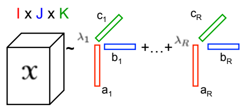
A2: Synonym Discovery

(Given) Noun Phrase	(Discovered) Potential Synonyms
pollutants	dioxin, sulfur dioxide, greenhouse gases, particulates, nitrogen oxide, air pollutants, cholesterol
disabilities	infections, dizziness, injuries, diseases, drowsiness, stiffness, injuries
vodafone	verizon, comcast
Christian history	European history, American history, Islamic history, history
disbelief	dismay, disgust, astonishment
cyberpunk	online-gaming
soul	body

CMU SCS

Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Matrices and tensors provide elegant theory and algorithms



55

CMU SCS

References

- Inderjit S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha: Information-theoretic co-clustering. KDD 2003: 89-98
- 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

56