

Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Lecture #19: Tensor decompositions

C. Faloutsos

1

Carnegie Mellon



Problem

- Q: who-calls-whom-when – patterns?
 - Triplets (source-ip, dest-ip, port#)
 - KB (subject, verb, object)


15-826

Copyright (c) 2019 C. Faloutsos

2

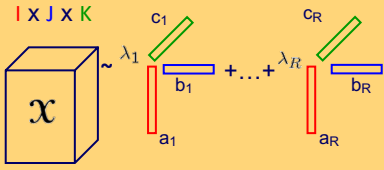
2

CarnegieMellon



Conclusions

- Q: who-calls-whom-when – patterns?
 - Triplets (source-ip, dest-ip, port#)
 - KB (subject, verb, object)
- A: Tensor analysis (PARAFAC)
 - <http://www.tensortoolbox.org/>



15-826 Copyright (c) 2019 C. Faloutsos 3

3

CarnegieMellon

Must-read Material

- [[Graph-Textbook](#)] Ch.16.
- Tensors survey: Papalexakis, Faloutsos, Sidiropoulos [Tensor for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms](#) ACM Trans. on Intelligent Systems and Technology, 8,2, Oct. 2016. ([local copy](#))

15-826 Copyright (c) 2019 C. Faloutsos 4

4

CarnegieMellon

Outline

Goal: 'Find **similar / interesting things**'

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

15-826 Copyright (c) 2019 C. Faloutsos 5

5

CarnegieMellon

Indexing - Detailed outline

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

15-826 Copyright (c) 2019 C. Faloutsos 6

6

Carnegie Mellon

Outline

- Motivation - Definitions
- Tensor tools
- Case studies



15-826 Copyright (c) 2019 C. Faloutsos 7

7

Carnegie Mellon

Most of foils by

- Dr. Tamara Kolda (Sandia N.L.)
• csmr.ca.sandia.gov/~tgkolda
- Prof. Jimeng Sun (GaTech)
• www.cc.gatech.edu/people/jimeng-sun



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

15-826 Copyright (c) 2019 C. Faloutsos 8

8

Motivation 1: Why “matrix”?

- Why matrices are important?

Examples of Matrices: Graph - social network

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

CarnegieMellon

Examples of Matrices: cloud of n-d points

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

15-826

Copyright (c) 2019 C. Faloutsos

11

11

CarnegieMellon

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
...

15-826

Copyright (c) 2019 C. Faloutsos

12

12

CarnegieMellon

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
...

15-826 Copyright (c) 2019 C. Faloutsos 13

13

CarnegieMellon

Examples of Matrices: Authors and terms

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

15-826 Copyright (c) 2019 C. Faloutsos 14

14

CarnegieMellon

Examples of Matrices: sensor-ids and time-ticks

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

15-826

Copyright (c) 2019 C. Faloutsos

15

15

CarnegieMellon

Motivation: Why tensors?

- Q: what is a tensor?

15-826

Copyright (c) 2019 C. Faloutsos

16

16

CarnegieMellon

Motivation 2: Why tensor?

- A: N-D generalization of matrix:

KDD' 17

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

15-826 Copyright (c) 2019 C. Faloutsos 17

17

CarnegieMellon

Motivation 2: Why tensor?

- A: N-D generalization of matrix:

KDD' 15
KDD' 16
KDD' 17

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

15-826 Copyright (c) 2019 C. Faloutsos 18

18

CarnegieMellon

Tensors are useful for 3 or more modes

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

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

15-826 19

19

CarnegieMellon

Motivating Applications

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

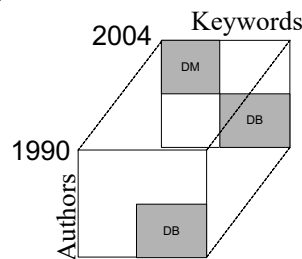
15-826 20

Copyright (c) 2019 C. Faloutsos

20

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



15-826

Copyright (c) 2019 C. Faloutsos

21

21

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

15-826

Copyright (c) 2019 C. Faloutsos

22

22

Carnegie Mellon

Outline

- Motivation –
Definitions
- **Tensor tools**
- Case studies

- Tensor Basics
- Tucker
- PARAFAC

15-826 Copyright (c) 2019 C. Faloutsos 23

23

Carnegie Mellon

Tensor Basics

24

Answer to both: tensor factorization

- Recall: (SVD) matrix factorization: finds blocks

15-826 Copyright (c) 2019 C. Faloutsos

25

Answer to both: tensor factorization

- PARAFAC decomposition

15-826 Copyright (c) 2019 C. Faloutsos

26

Carnegie Mellon

Answer: tensor factorization

- PARAFAC decomposition
- Results for who-calls-whom-when
 - 4M x 15 days ?? ?? ??

15-826 Copyright (c) 2019 C. Faloutsos 27

27

Carnegie Mellon

Goal: extension to ≥ 3 modes

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

Example of outer product 'o':

15-826 Copyright (c) 2019 C. Faloutsos 28

28

CarnegieMellon

Goal: extension to ≥ 3 modes

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

Suppose
 $R=1$
 $\mathbf{a}_1=(1,2,3,4)$
 $\mathbf{b}_1=(2,2,2)$
 $\mathbf{c}_1=(10,11)$
 $\lambda_1=7$

$X(1,1,1)=?$
 $X(3,1,2)=?$
 $X(5,1,1)=?$

Copyright (c) 2019 C. Faloutsos 29

29

CarnegieMellon

Goal: extension to ≥ 3 modes

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

Suppose
 $r=1$
 $\mathbf{a}_1=(1,2,3,4)$
 $\mathbf{b}_1=(2,2,2)$
 $\mathbf{c}_1=(10,11)$
 $\lambda_1=7$

$X(1,1,1)=7 * 1 * 2 * 10$
 $X(3,1,2)=?$
 $X(5,1,1)=?$

Copyright (c) 2019 C. Faloutsos 30

30

Goal: extension to ≥ 3 modes

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

Suppose
 $r=1$
 $\mathbf{a}_1 = (1, 2, 3, 4)$
 $\mathbf{b}_1 = (2, 2, 2)$
 $\mathbf{c}_1 = (10, 11)$
 $\lambda_1 = 7$

$X(1, 1, 1) = 7 * 1 * 2 * 10$
 $X(3, 1, 2) = 7 * 3 * 2 * 11$
 $X(5, 1, 1) = \text{N/A} - \text{TRICK QUESTION}$

31

31

Goal: extension to ≥ 3 modes

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

Suppose
 $r=1$
 $\mathbf{a}_1 = (1, 2, 3, 4)$
 $\mathbf{b}_1 = (2, 2, 2)$
 $\mathbf{c}_1 = (10, 11)$
 $\lambda_1 = 7$

$X(1, 1, 1) = 7 * 1 * 2 * 10$
 $X(3, 1, 2) = 7 * 3 * 2 * 11$
 $X(5, 1, 1) = ??$

32

32

CarnegieMellon

Goal: extension to ≥ 3 modes

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

Suppose
 $r=1$
 $\mathbf{a}_1=(1,2,3,4)$
 $\mathbf{b}_1=(2,2,2)$
 $\mathbf{c}_1=(10,11)$
 $\lambda_1=7$

$X(1,1,1)=7*1*2*10$
 $X(3,1,2)= 7*3*2*11$
 $X(5,1,1)= \text{N/A} - \text{TRICK QUESTION}$

Copyright (c) 2019 C. Faloutsos

33

33

CarnegieMellon

Main points:

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

Copyright (c) 2019 C. Faloutsos

34

34

Carnegie Mellon

Specially Structured Tensors

35

Carnegie Mellon

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

15-826
Copyright (c) 2019 C. Faloutsos
36

36

Carnegie Mellon
details

Specially Structured Tensors

- **Tucker Tensor**

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

In matrix form:

$$\begin{aligned} \mathbf{X}_{(1)} &= \mathbf{U} \mathbf{G}_{(1)} (\mathbf{W} \otimes \mathbf{V})^\top \\ \mathbf{X}_{(2)} &= \mathbf{V} \mathbf{G}_{(2)} (\mathbf{W} \otimes \mathbf{U})^\top \\ \mathbf{X}_{(3)} &= \mathbf{W} \mathbf{G}_{(3)} (\mathbf{V} \otimes \mathbf{U})^\top \end{aligned}$$

- **Kruskal Tensor**

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

In matrix form:

Let $\Lambda = \text{diag}(\lambda)$

$$\begin{aligned} \mathbf{X}_{(1)} &= \mathbf{U} \Lambda (\mathbf{W} \odot \mathbf{V})^\top \\ \mathbf{X}_{(2)} &= \mathbf{V} \Lambda (\mathbf{W} \odot \mathbf{U})^\top \\ \mathbf{X}_{(3)} &= \mathbf{W} \Lambda (\mathbf{V} \odot \mathbf{U})^\top \end{aligned}$$

$\text{vec}(\mathcal{X}) = (\mathbf{W} \otimes \mathbf{V} \otimes \mathbf{U}) \text{vec}(\mathcal{G})$
 $\text{vec}(\mathcal{X}) = (\mathbf{W} \odot \mathbf{V} \odot \mathbf{U}) \lambda$

15-826
Copyright (c) 2019 C. Faloutsos
37

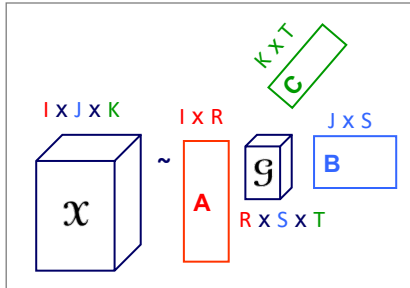
37

Carnegie Mellon

Tensor Decompositions

38

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!

15-826

Copyright (c) 2019 C. Faloutsos

39

39

Intuition behind core tensor

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

15-826

Copyright (c) 2019 C. Faloutsos

40

40

Carnegie Mellon

15-826 Copyright (c) 2019 C. Faloutsos 43

43

Carnegie Mellon

Tucker Decomposition

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

Given $\mathbf{A}, \mathbf{B}, \mathbf{C}$, 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
- \mathbf{A}, \mathbf{B} , and \mathbf{C} generally assumed to be orthonormal (generally assume they have full column rank)
- \mathcal{G} is not diagonal
- Not unique

Recall the equations for converting a tensor to a matrix

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})^\top$$

$$\mathbf{X}_{(2)} = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})^\top$$

$$\mathbf{X}_{(3)} = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})^\top$$

$$\text{vec}(\mathcal{X}) = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\text{vec}(\mathcal{G})$$

15-826 Copyright (c) 2019 C. Faloutsos 44

44

CarnegieMellon

Kronecker product

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 10 & 20 & 30 \end{bmatrix}$$

m1 x n1 *m2 x n2*

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} 1 * \mathbf{B} & 2 * \mathbf{B} \\ 3 * \mathbf{B} & 4 * \mathbf{B} \end{bmatrix}$$

*m1*m2 x n1*n2*

$$= \begin{bmatrix} 1 * 10 & 1 * 20 & 1 * 30 & 2 * 10 & 2 * 20 & 2 * 30 \\ 3 * 10 & 3 * 20 & 3 * 30 & 4 * 10 & 4 * 20 & 4 * 30 \end{bmatrix}$$

15-826
Copyright (c) 2019 C. Faloutsos
45

45

CarnegieMellon

Outline

- Motivation –
- Definitions
- Tensor tools
- Case studies

}

- Tensor Basics
- Tucker
- PARAFAC

15-826
Copyright (c) 2019 C. Faloutsos
46

46

CarnegieMellon

CANDECOMP/PARAFAC Decomposition

$$\mathcal{X} \approx [\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}] = \sum_{r=1}^R \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

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

15-826 Copyright (c) 2019 C. Faloutsos 47

47

CarnegieMellon

IMPORTANT

Tucker vs. PARAFAC Decompositions

<ul style="list-style-type: none"> • Tucker <ul style="list-style-type: none"> – Variable transformation in each mode – Core G may be dense – \mathbf{A}, \mathbf{B}, \mathbf{C} generally orthonormal – Not unique 	<ul style="list-style-type: none"> • PARAFAC <ul style="list-style-type: none"> – Sum of rank-1 components – No core, i.e., superdiagonal core – \mathbf{A}, \mathbf{B}, \mathbf{C} may have linearly dependent columns – Generally unique
---	---

15-826 Copyright (c) 2019 C. Faloutsos 48

48

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://www.tensortoolbox.org/>

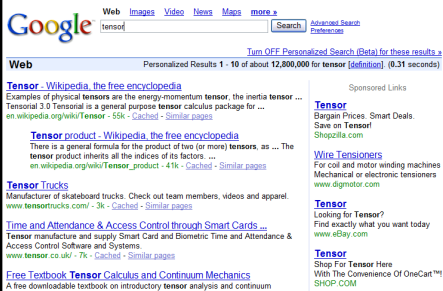
Outline

- Motivation - Definitions
- Tensor tools
- Case studies
 - ➔ – P1: web graph mining ('TOPHITS')
 - P2: phone-call patterns
 - P3: N.E.L.L. (never ending language learner)
 - P4: network traffic
 - P5: FaceBook activity

CarnegieMellon

P1: Web graph mining

- How to order the importance of web pages?
 - Kleinberg's algorithm HITS
 - PageRank



Google Web Images Video News Maps more »
tensor [Search] Advanced Search Preferences

Personalized Results 1 - 10 of about 12,800,000 for tensor [definition] (0.31 seconds)

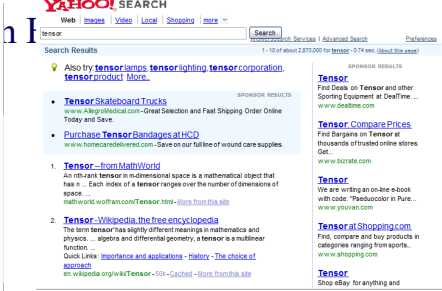
Tensor - Wikipedia, the free encyclopedia
Examples of physical tensors are the energy-momentum tensor, the inertia tensor ...
en.wikipedia.org/wiki/Tensor - 55k - Cached - Similar pages

Tensor product - Wikipedia, the free encyclopedia
There is a general formula for the product of two (or more) tensors, as ... The tensor product inherits all the indices of its factors. ...
en.wikipedia.org/wiki/Tensor_product - 41k - Cached - Similar pages

Tensor Trucks
Manufacturer of skateboard trucks. Check out team members, videos and apparel.
www.tensortrucks.com/ - 3k - Cached - Similar pages

Time and Attendance & Access Control through Smart Cards ...
Tensor manufacture and supply Smart Card and Biometric Time and Attendance & Access Control Software and Systems.
www.tensor.co.uk/ - 7k - Cached - Similar pages

Free Textbook Tensor Calculus and Continuum Mechanics
A free downloadable textbook on introductory tensor analysis and continuum ...
15-826



Yahoo! SEARCH Web Images Video Local Shopping more »
tensor [Search] Preferences

Search Results 1 - 10 of about 2,870,000 for tensor - 0.74 sec. (about this page)

Also try **tensor lamps**, **tensor lighting**, **tensor corporation**, **tensor product**, **block**.

Tensor Skateboard Trucks
www.4legged.com - Great Selection and Fast Shipping Order Online Today and Save

Purchase Tensor Bandages at HCD
www.homecaredelevered.com - Save on our fullline of wound care supplies.

1. **Tensor** - from MathWorld
An n-rank tensor in n-dimensional space is a mathematical object that has n ... Each index of a tensor ranges over the number of dimensions of space.
mathworld.wolfram.com/Tensor.html - More from this site

2. **Tensor** - Wikipedia, the free encyclopedia
The term tensor has slightly different meanings in mathematics and physics ... algebra and differential geometry, a tensor is a multilinear function ...
Quick Links: [importance and applications](#) - [history](#) - [the choice of associated](#)
en.wikipedia.org/wiki/Tensor - 50k - Cached - More from this site

Tensor
Find Deals on Tensor and other Sporting Equipment at DealTime ...
www.dealtime.com

Tensor Compare Prices
Find Bargains on Tensor at thousands of trusted online stores. Get ...
www.24mat.com

Tensor
We are selling an online eBook with code: "9adocctor in Pure."
www.youwant.com

Tensor at Shopping.com
Find, compare and buy products in categories ranging from airports ...
www.shopping.com

Tensor
Shop eBay for anything and ...

Copyright (c) 2019 C. Faloutsos 51

51

CarnegieMellon

P1: Web graph mining

- T. G. Kolda, B. W. Bader and J. P. Kenny, *Higher-Order Web Link Analysis Using Multilinear Algebra*, ICDM 2005: ICDM, pp. 242-249, November 2005, [doi:10.1109/ICDM.2005.77](https://doi.org/10.1109/ICDM.2005.77). [PDF]

15-826
Copyright (c) 2019 C. Faloutsos
52

52

26

CarnegieMellon

Kleinberg's Hubs and Authorities (the HITS method)

Endangered Species
Animals today are being threatened by a variety of environmental pressures. For example, the jaguar is losing prime habitat in the world. Zoos are trying to raise awareness of their plight.

Jaguar FAQ
Jaguars are an endangered species that live in the tropical rain forests of Central and South America. They live about 11 years in the wild and up to 22 years at a zoo.

Rain Forest Zoo
We have a new exhibit opening next month highlighting the endangered species of the Americas, including the jaguar.

Online Atlas
View maps of animal habitats from around the world, including those of endangered animals in North, South, and Central America.

Sparse adjacency matrix and its SVD:

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

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

authority scores for 1st topic + authority scores for 2nd topic

from = [] + [] + ...

hub scores for 1st topic + hub scores for 2nd topic

15-826
Copyright (c) 2019 C. Faloutsos
53

Kleinberg, JACM, 1999

53

CarnegieMellon

HITS Authorities on Sample Data

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

2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh.edu
.06	www.lehigha
.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-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.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

from = [] + [] + ...

hub scores for 1st topic + hub scores for 2nd topic

1
Copyright (c) 2019 C. Faloutsos
54

54

Carnegie Mellon

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

55

Carnegie Mellon

Topical HITS (TOPHITS)

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

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

from

to

=

hub scores for 1st topic

+

authority scores for 1st topic

+

authority scores for 2nd topic

+

hub scores for 2nd topic

+

...

Copyright (c) 2019 C. Faloutsos

56

CarnegieMellon

Topical HITS (TOPHITS)

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

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

15-826 Copyright (c) 2019 C. Faloutsos 57

57

CarnegieMellon

TOPHITS Terms & Authorities on Sample Data

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

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

$w_k = \# \text{ unique links using term } k$

Tensor PARAFAC

1st Principal Factor		
23	JAVA	.86 java.sun.com
18	SUN	.86 java.sun.com
17	PLATF	.99 www.lehigh.edu
16	SOLAF	.99 www.lehigh.edu
16	DEVEL	.97 www.ibm.com
16	SEARCH	.97 www.ibm.com
15	EDITIC	.18 www.alphaworks.ibm.com
15	NEWS	.18 www.alphaworks.ibm.com
15	DOWN	.18 www.alphaworks.ibm.com
16	LIBRA	.87 www.pueblo.gsa.gov
14	INFO	.87 www.pueblo.gsa.gov
12	SOFTV	.24 www.irs.gov
12	LEHIG	.24 www.irs.gov
12	NO-RE	.24 www.irs.gov
11	DEVEL	.87 www.whitehouse.gov
11	LINUX	.87 www.whitehouse.gov
11	RESOL	.18 www.irs.gov
11	CENTE	.18 www.irs.gov
11	TECHN	.18 www.irs.gov
10	DOWN	.35 www.palisade.com
15	US	.35 www.palisade.com
15	PUBLI	.35 www.solver.com
14	CONS	.35 www.solver.com
13	FREE	.99 www.adobe.com
15	HOUS	.99 www.adobe.com
13	BUDC	.99 www.adobe.com
13	PRES	.99 www.adobe.com
11	OFFIC	.99 www.adobe.com
05	SEAR	.81 www.weather.gov
05	ENGIN	.81 www.weather.gov
05	CONT	.41 www.spc.noaa.gov
05	ILOG	.41 www.spc.noaa.gov
05	DOWN	.73 www.irs.gov
17	ORGA	.43 travel.state.gov
15	NWS	.43 travel.state.gov
15	SEVER	.22 www.ssa.gov
15	FIRE	.22 www.ssa.gov
15	POLIC	.08 www.govbenefits.gov
14	CLIMA	.08 www.govbenefits.gov
14	INCOME	.06 www.usdoj.gov
14	STATE	.06 www.usdoj.gov
14	SERVICE	.03 www.census.gov
13	REVENUE	.03 www.census.gov
12	CREDIT	.02 www.usmint.gov
12	CREDIT	.02 www.usmint.gov
12	CREDIT	.02 www.nws.noaa.gov
12	CREDIT	.02 www.nws.noaa.gov
12	CREDIT	.02 www.gsa.gov
12	CREDIT	.02 www.gsa.gov
12	CREDIT	.01 www.annualcreditreport.com
12	CREDIT	.01 www.annualcreditreport.com

15-826 Copyright (c) 2019 C. Faloutsos 58

58

CarnegieMellon

Outline

- Motivation - Definitions
- Tensor tools
- Case studies
 - P1: web graph mining ('TOPHITS')
 - ➔ – P2: phone-call patterns
 - P3: N.E.L.L. (never ending language learner)
 - P4: network traffic
 - P5: FaceBook activity

15-826 Copyright (c) 2019 C. Faloutsos 59

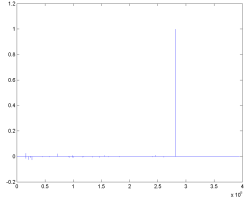
59

CarnegieMellon

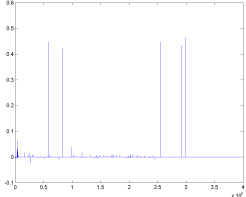
P2: Anomaly detection in time-evolving graphs

- Anomalous communities in phone call data:
 - European country, 4M clients, data over 2 weeks

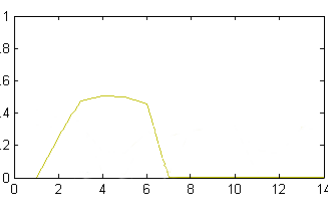
1 caller



5 receivers



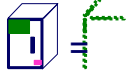
4 days of activity



~200 calls to EACH receiver on EACH day!

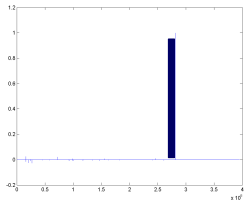
15-826 Copyright (c) 2019 C. Faloutsos 60

60

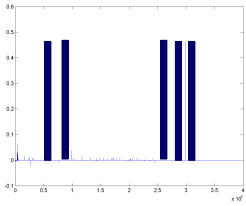
CarnegieMellon **P2: Anomaly detection in time-evolving graphs** 

- Anomalous communities in phone call data:
 - European country, 4M clients, data over 2 weeks

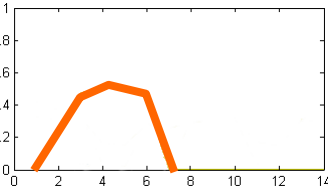
1 caller



5 receivers



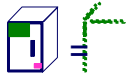
4 days of activity




~200 calls to EACH receiver on EACH day!

15-826
Copyright (c) 2019 C. Faloutsos
61

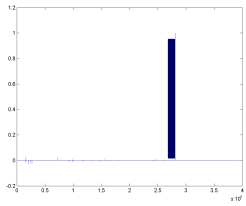
61

CarnegieMellon **P2: Anomaly detection in time-evolving graphs** 

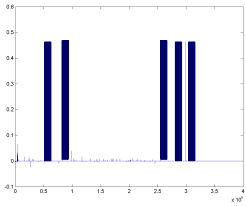
- Anomalous communities in phone call data:
 - European country, 4M clients, data over 2 weeks



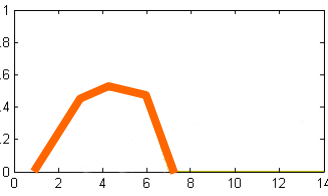
1 caller



5 receivers



4 days of activity



~200 calls to EACH receiver on EACH day!

15-826
Copyright (c) 2019 C. Faloutsos
62

62

CarnegieMellon

P2: Anomaly detection in time-evolving graphs

- Anomalous communities in phone call data:
 - European country, 4M clients, data over 2 weeks



Miguel Araujo, Spiros Papadimitriou, Stephan Günemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, Evangelos Papalexakis, Danai Koutra. *Com2: Fast Automatic Discovery of Temporal (Comet) Communities*. PAKDD 2014, Tainan, Taiwan.

63

CarnegieMellon

Outline

- Motivation - Definitions
- Tensor tools
- Case studies
 - P1: web graph mining ('TOPHITS')
 - P2: phone-call patterns
 - ➔ – P3: N.E.L.L. (never ending language learner)
 - P4: network traffic
 - P5: FaceBook activity

15-826

Copyright (c) 2019 C. Faloutsos

64

64

CarnegieMellon

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'

15-826
Copyright (c) 2019 C. Faloutsos
67

67

CarnegieMellon

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'

15-826
Copyright (c) 2019 C. Faloutsos
68

68

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

15-826
Copyright (c) 2019 C. Faloutsos
69

69

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

15-826
Copyright (c) 2019 C. Faloutsos
70

70

CarnegieMellon


Outline

- Motivation - Definitions
- Tensor tools
- Case studies
 - P1: web graph mining ('TOPHITS')
 - P2: phone-call patterns
 - P3: N.E.L.L. (never ending language learner)
 - ➔ – P4: network traffic
 - P5: FaceBook activity

15-826 Copyright (c) 2019 C. Faloutsos 71

71

CarnegieMellon




ParCube: Sparse Parallelizable Tensor Decompositions

Evangelos E. Papalexakis, Christos Faloutsos, Nikos Sidiropoulos,
ECML/PKDD 2012

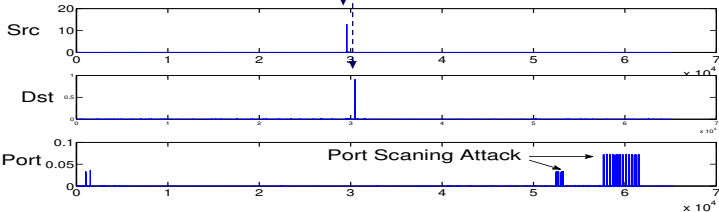
Evangelos E. Papalexakis
Email: epapalex@cs.ucr.edu
Web: <http://www.cs.ucr.edu/~epapalex>

72

CarnegieMellon


P4: LBNL Network Data


1 src 1 dst



- Modes: src IP, dst IP, port #
- ~ Port Scanning Attack

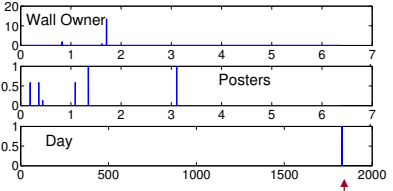
15-826
Copyright (c) 2019 C. Faloutsos
73

73

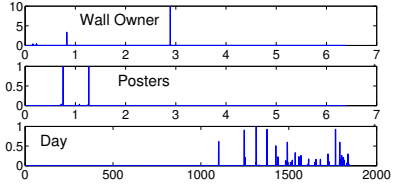
CarnegieMellon


P5: FACEBOOK Wall posts


1 Wall ↙



(a) FACEBOOK anomaly (Wall owner's birthday)



(b) FACEBOOK normal activity



1 day

- Modes: wall-owner, poster, timestamp
- Discovery: birthday-like event.

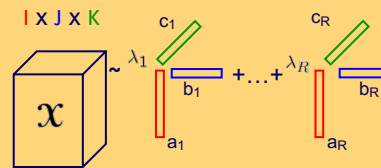
15-826
Copyright (c) 2019 C. Faloutsos
74

74



Conclusions

- Q: who-calls-whom-when – patterns?
 - Triplets (source-ip, dest-ip, port#)
 - KB (subject, verb, object)
- A: Tensor analysis (PARAFAC)
 - <http://www.tensortoolbox.org/>



15-826

Copyright (c) 2019 C. Faloutsos

75

75

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

15-826

Copyright (c) 2019 C. Faloutsos

76

76