Towards Interactive Construction of Topical Hierarchy

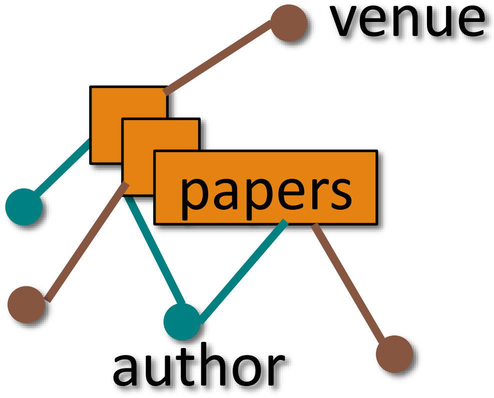
Chi Wang

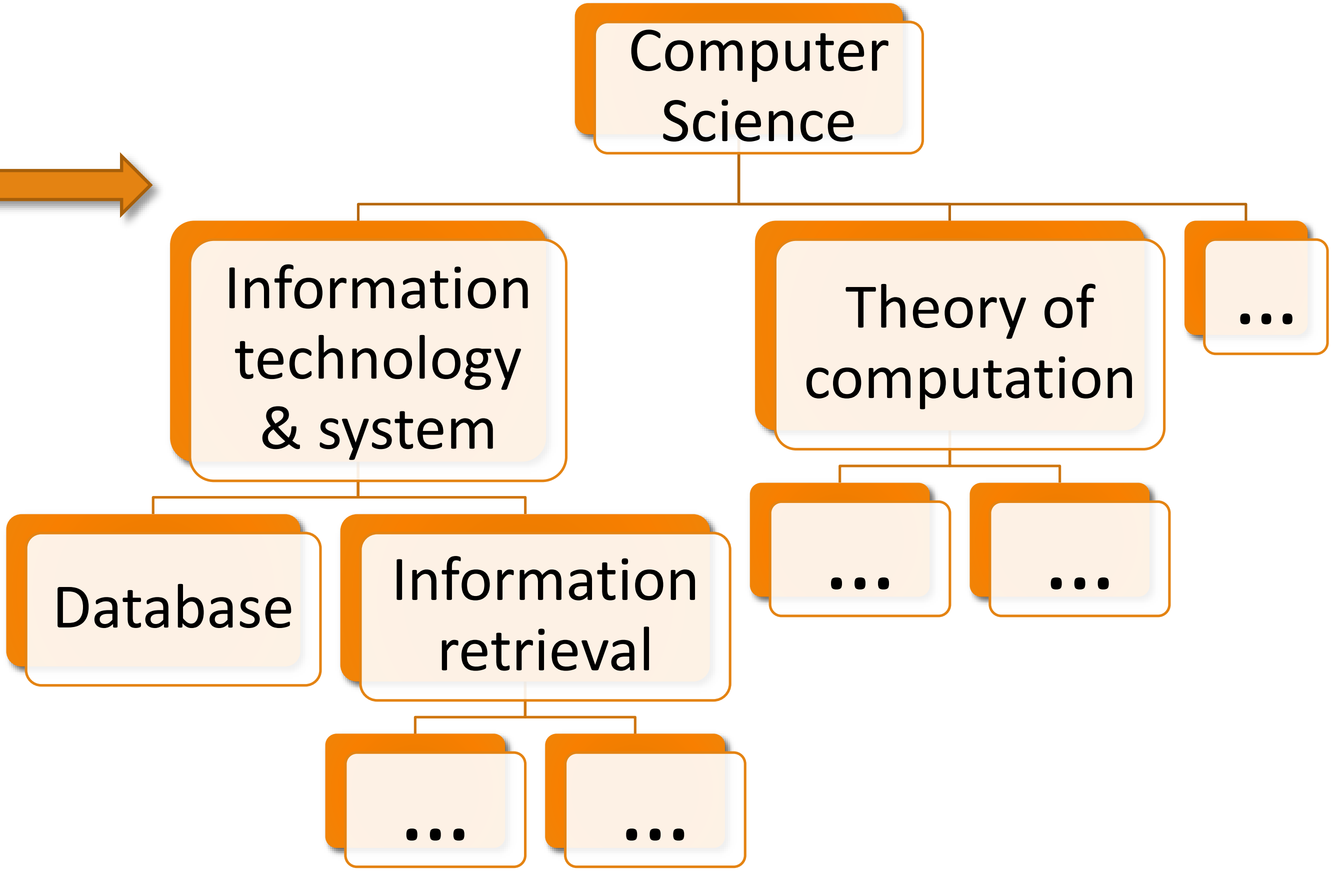
Microsoft Research, Redmond

Xueqing Liu, Yanglei Song, Jiawei Han

University of Illinois at Urbana-Champaign

Topic Hierarchy: Summarize the Data with Multiple Granularity

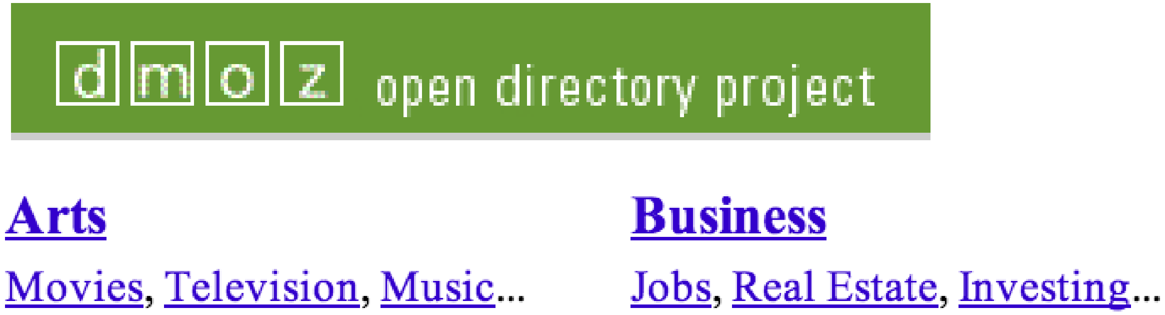




- ❑ Top 10 researchers in data mining?
  - And their specializations?
- ❑ Important research areas in SIGIR conference?

Construct A Topical Hierarchy: Manual or Automated?

MANUAL APPROACH


dmoz open directory project

Arts

Movies, Television, Music...

Business

Jobs, Real Estate, Investing...

4,249,724 sites
89,312 editors
over 1,020,274 categories
1998-2014

□ Quality; Labor-intensive

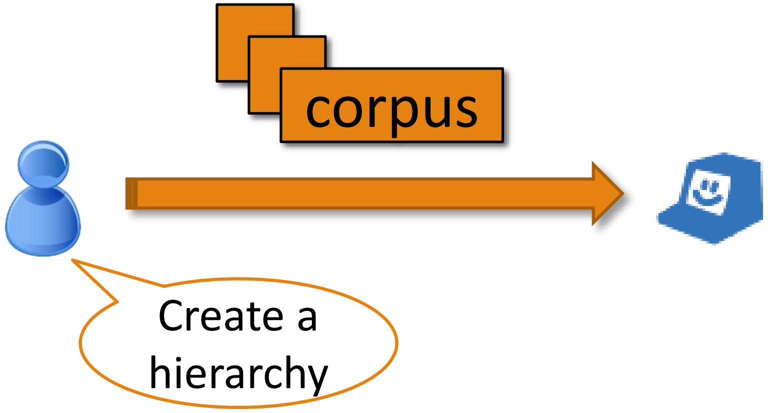
AUTOMATED APPROACH (TOPIC MODELING)

nested Chinese Restaurant Process [Griffiths 04]
Pachinko Allocation Model [Li & McCallum 06]
hierarchical Pachinko Allocation [Mimno 07]
recursive Chinese Restaurant Process [Kim 12]
nested Chinese Restaurant Franchise [Ahmed 13]
splitLDA [Pujara & Skomoroch 12]

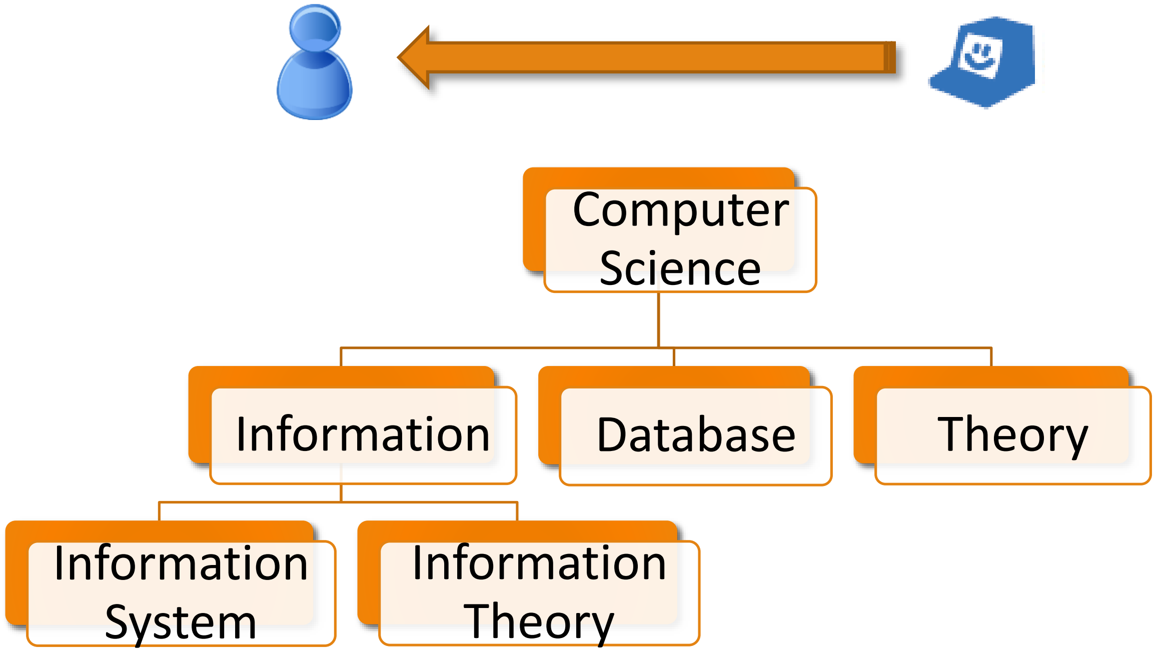
□ Low human effort; Low quality

Interactive Approach: Arm Human Curators with Automated Operators

- A human curator runs an operator

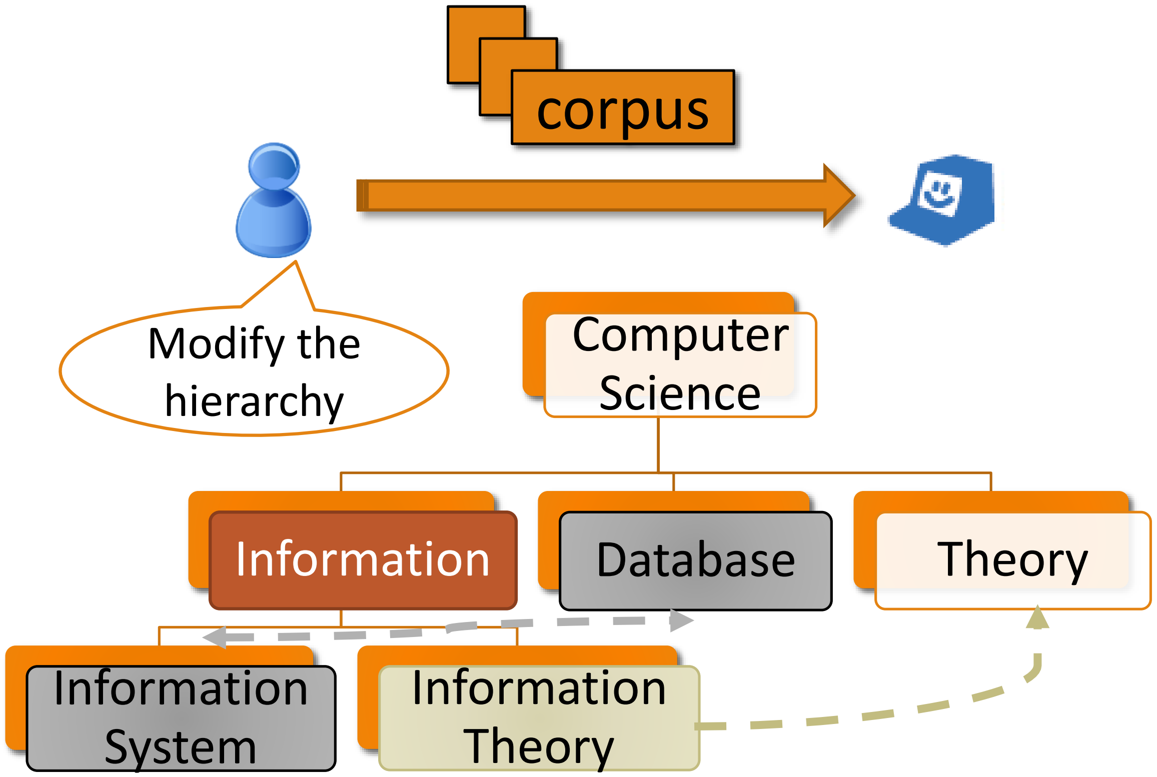


- The operator returns an initial hierarchy

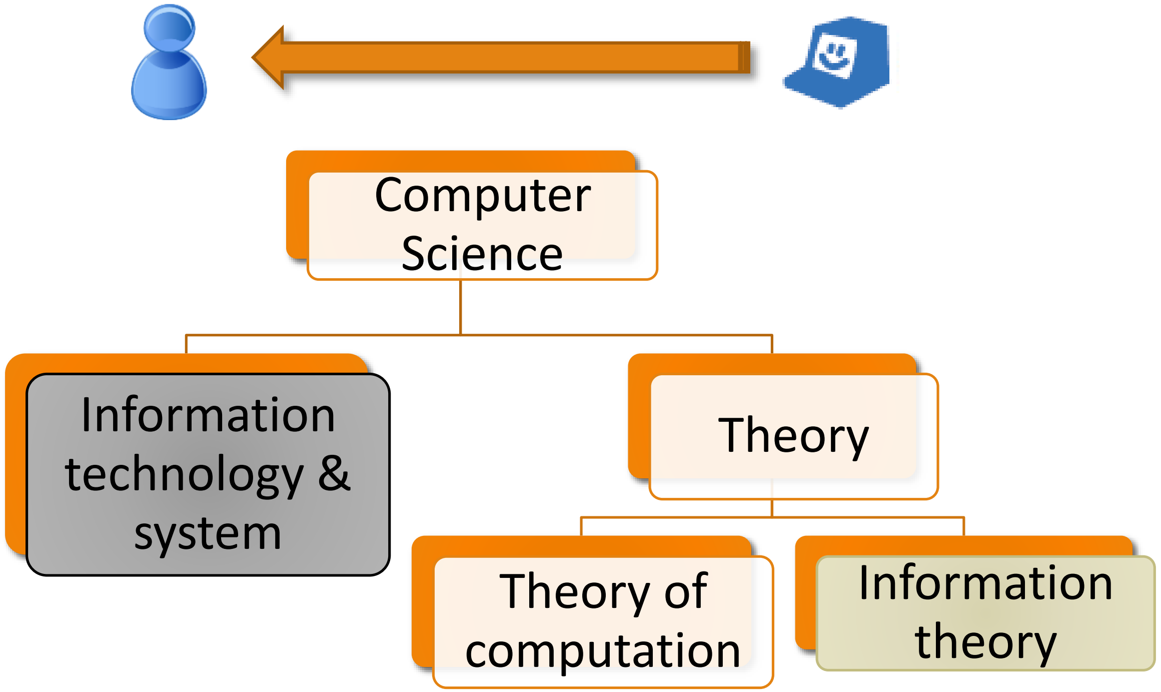


Interactive Approach: Arm Human Curators with Automated Operators (cont’d)

- ❑ A human curator runs an operator

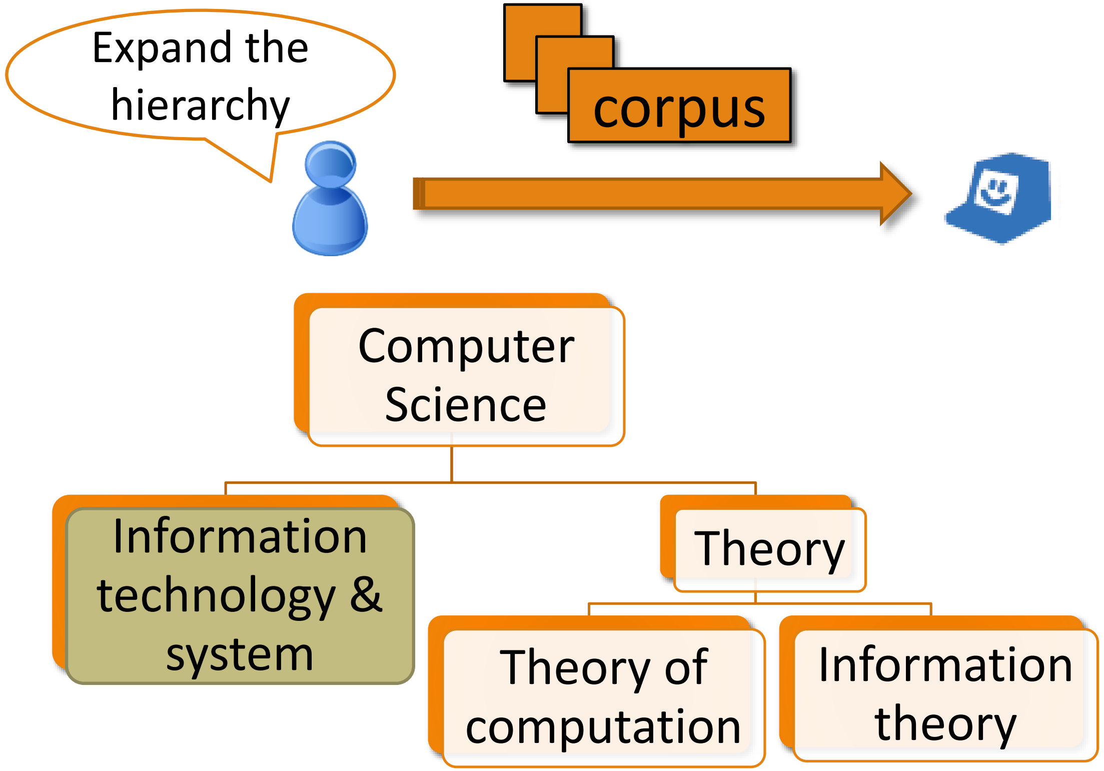


- ❑ The operator returns a modified hierarchy



Interactive Approach: Arm Human Curators with Automated Operators (cont'd)

- ❑ A human curator runs an operator



Expand the hierarchy

corpus

Computer Science

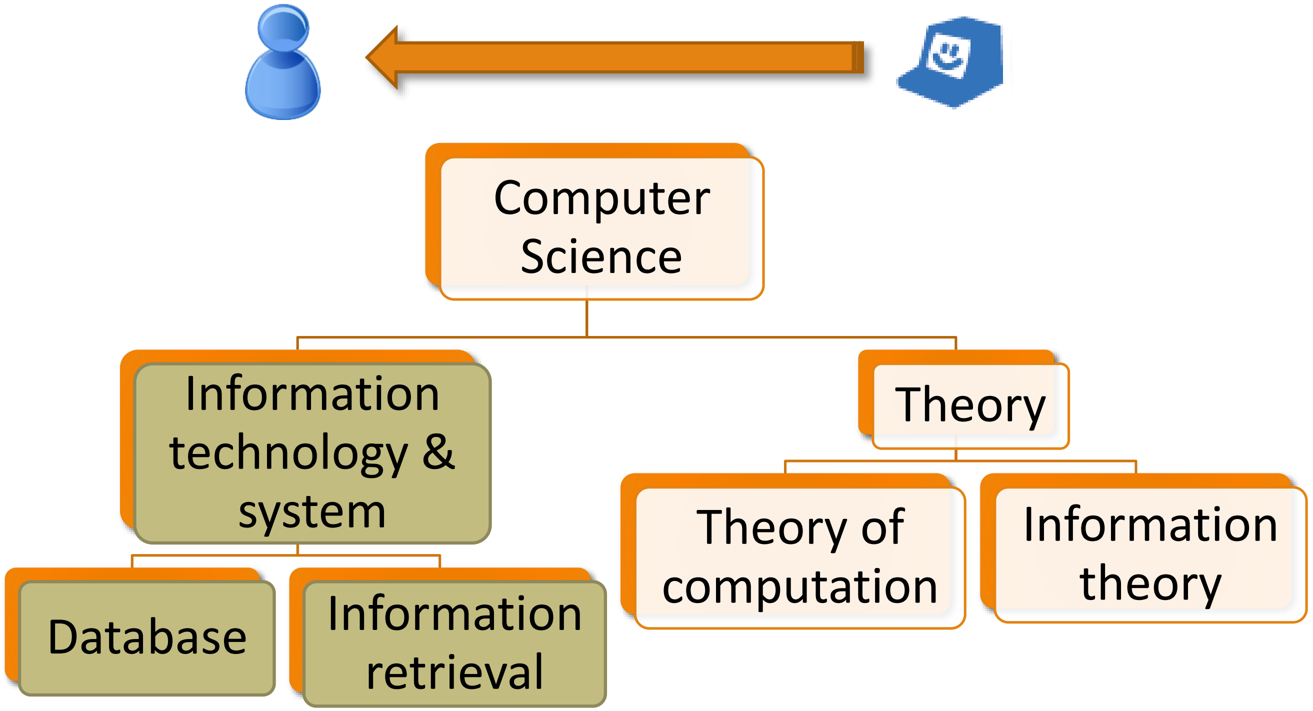
Information technology & system

Theory

Theory of computation

Information theory

- ❑ The operator returns a modified hierarchy



Computer Science

Information technology & system

Theory

Database

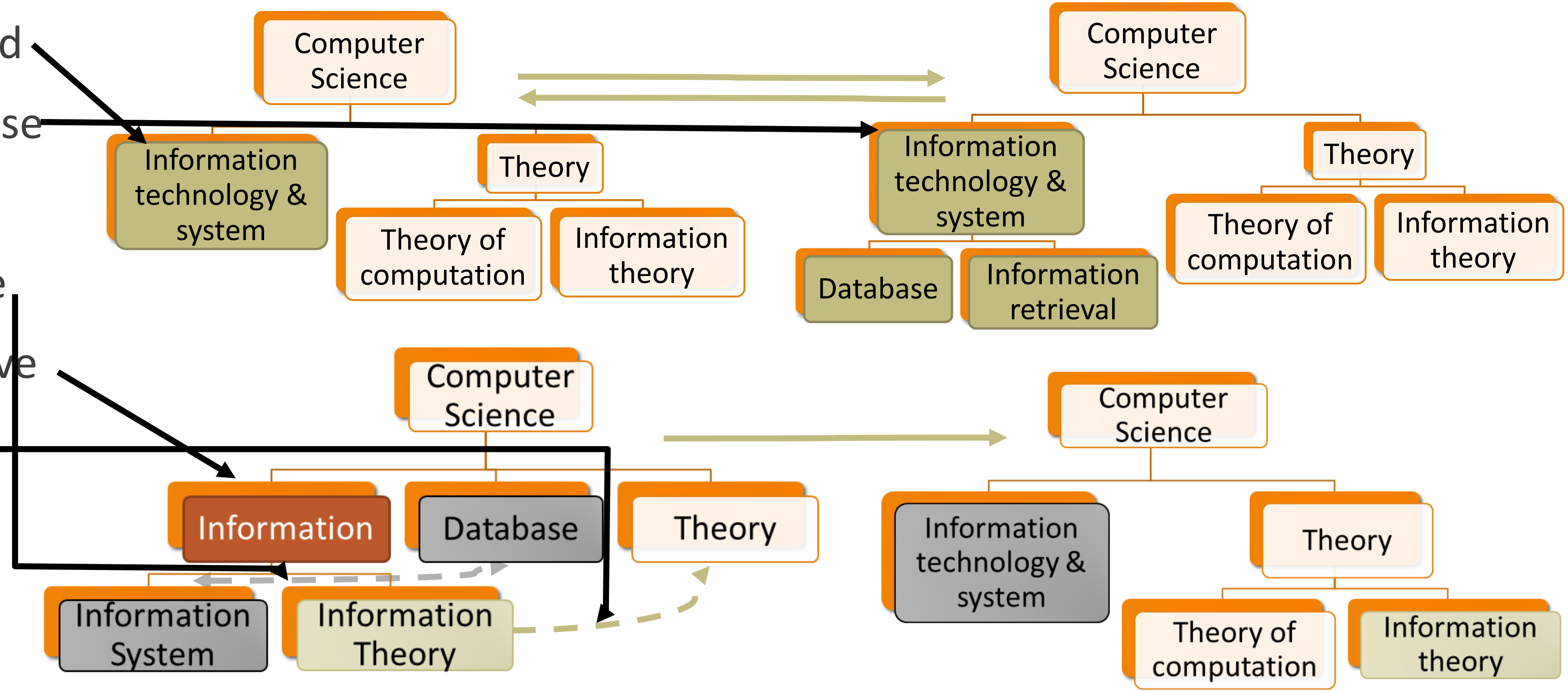
Information retrieval

Theory of computation

Information theory

Operators

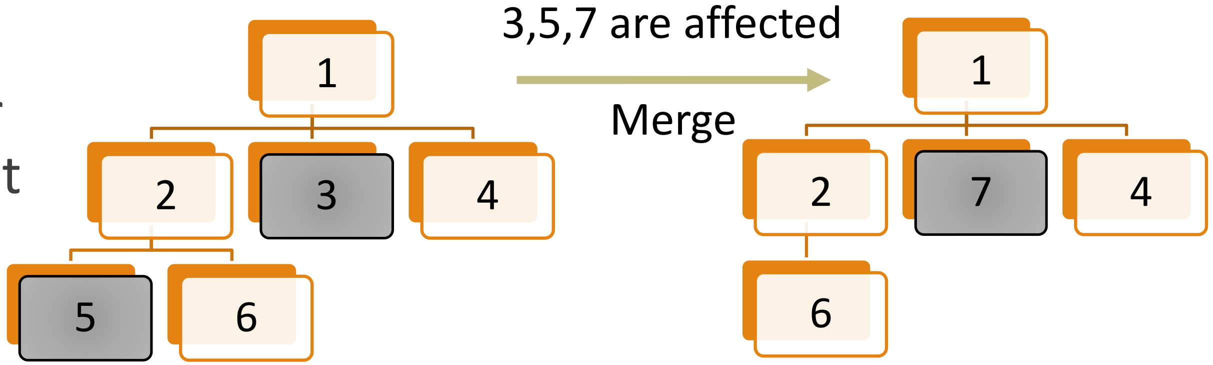
- ❑ Expand
- ❑ Collapse
- ❑ Split
- ❑ Merge
- ❑ Remove
- ❑ Move



Challenge: Consistency & Efficiency

- ❑ Single-run consistency

  The generative models before and after an operator should be equivalent except for a few affected nodes



- ❑ Multi-run consistency

  Returned hierarchy should be (nearly) identical with identical input

- ❑ Efficiency

  Small # data scans

✗ Traditional inference methods produce large variance across runs

✗ Traditional inference methods require hundreds to thousands of iterations to converge (no guarantee)

Our Solution:
Scalable Recursive Tensor Decomposition

- □ Single-run consistency

  The generative models before and after an operator should be equivalent except for a few affected nodes

- □ Multi-run consistency

  Returned hierarchy should be (nearly) identical with identical input

- □ Efficiency

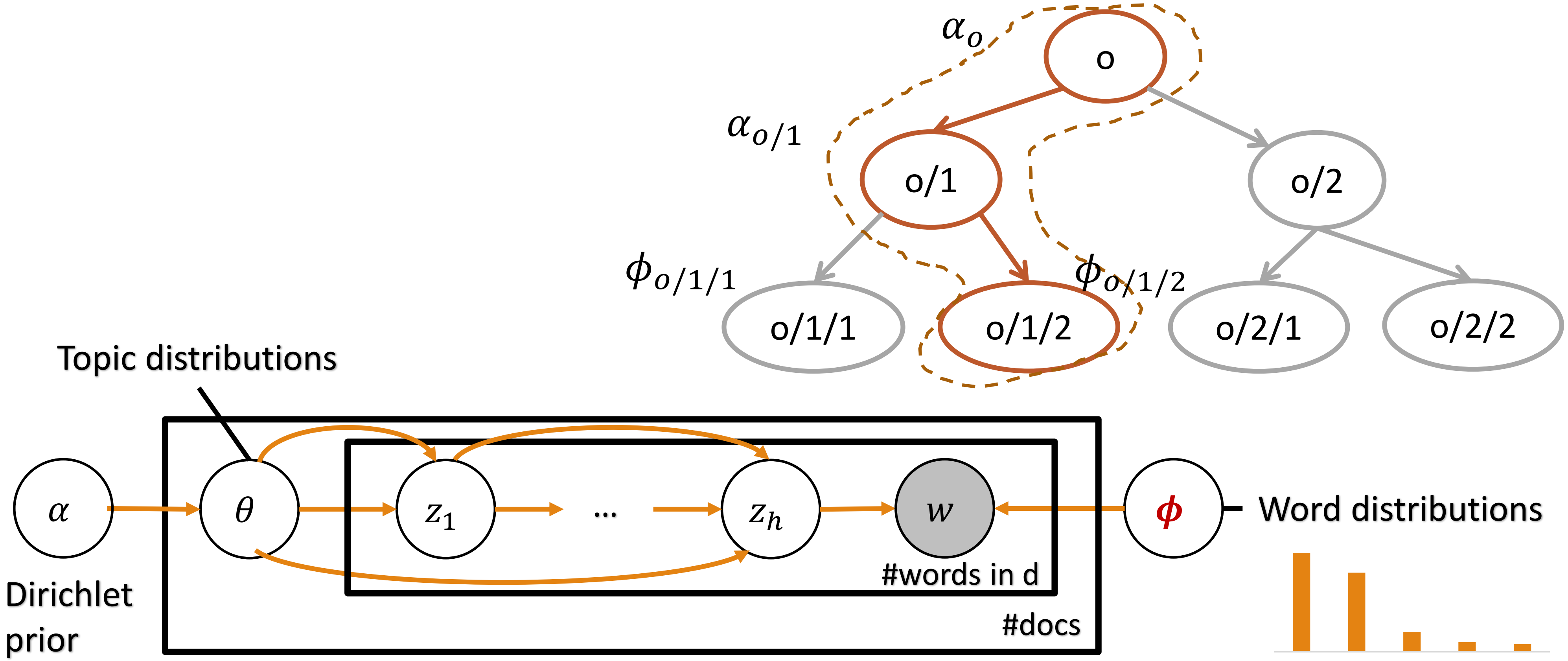
  Small # data scans

- ✓ A new hierarchical topic model that supports consistent manipulation operators

- ✓ A moment-based inference method with theoretical bound of output variance

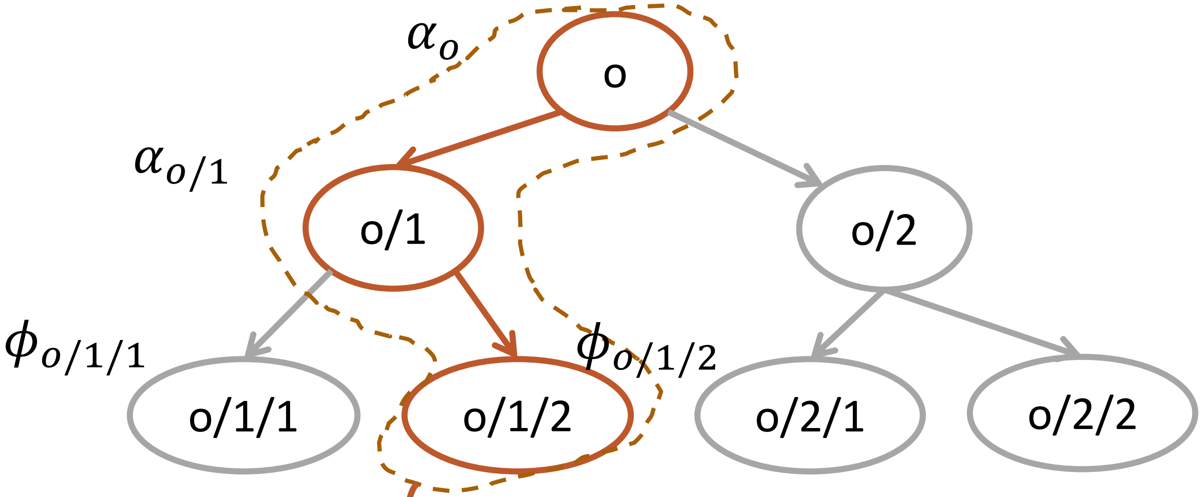
- ✓ Only requires 3 scans of data

Latent Dirichlet Allocation with Topic Tree



Latent Dirichlet Allocation with Topic Tree

LDA with Topic Tree (cont'd)



To generate a token in document d:

1. Sample a topic path $z_1 \rightarrow \cdots \mathrm{z_h}$ according to θ_d
2. Sample a word w according to ϕ_{z_h}

Atomic Operators

- ❑ $EXP(t, k)$

  Discover k subtopics of a leaf topic t

- ❑ $MER(t_1, t_2)$

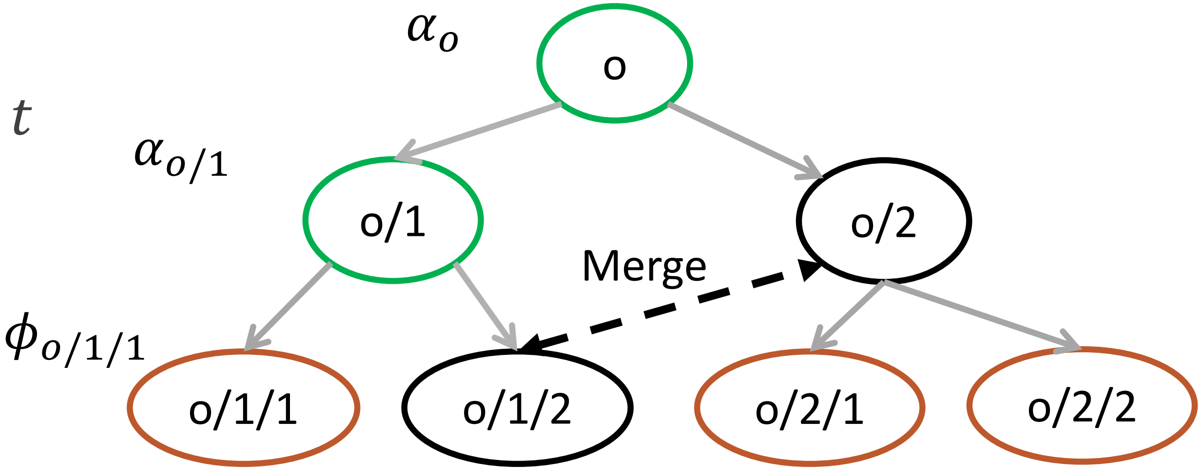
  Merge two topics

- ❑ $MOV(t_1, t_2)$

  Move the subtree rooted at t_1 to be

  u

These 3 atomic operators are sufficient



Consistency condition

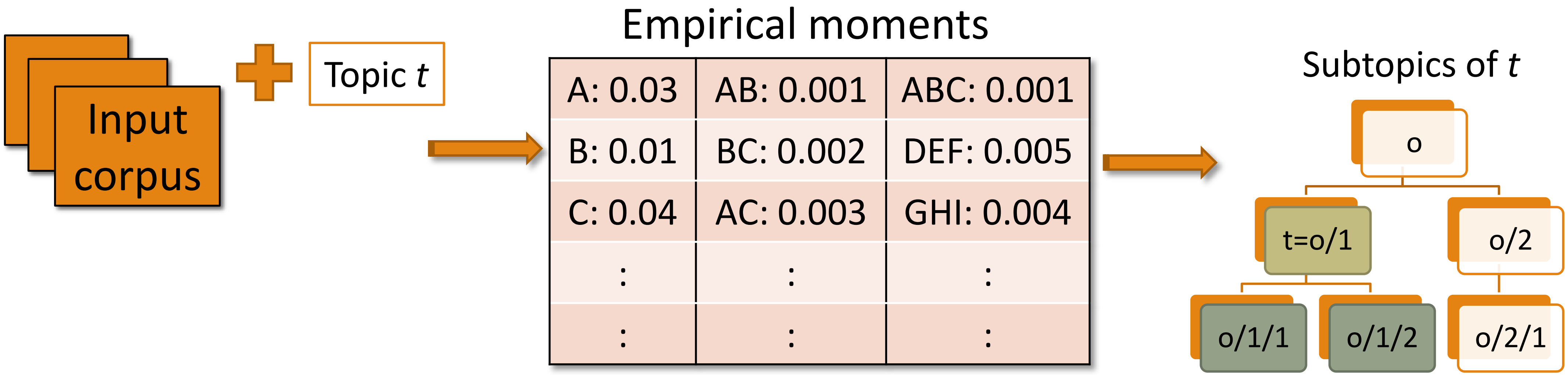
For each unaffected leaf node, α and ϕ remain unchanged

For each internal node t,

$$\alpha = \sum_{c\ is\ t's\ child} \alpha_c$$

Implementation of Operators

❑ Decompose *moments* (expectation of patterns) for EXP



Input corpus

Topic *t*

Empirical moments

A: 0.03	AB: 0.001	ABC: 0.001
B: 0.01	BC: 0.002	DEF: 0.005
C: 0.04	AC: 0.003	GHI: 0.004
:	:	:
:	:	:

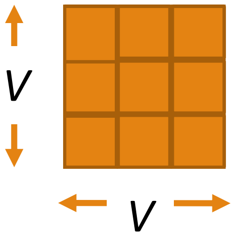
Subtopics of *t*

❑ Leverage the special property of the moments (sparse, low rank, and decoupled decomposition) for scale up

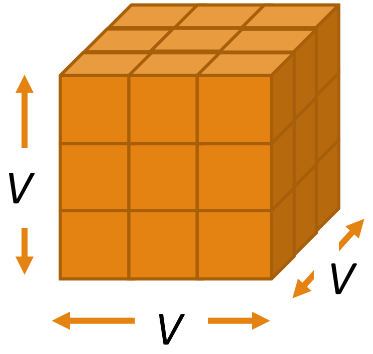
❑ Manipulate moments efficiently for MER and MOV

Tensor Orthogonal Decomposition for *EXP* operator

Theorem. *The patterns up to length 3 are sufficient for* $EXP(t,k)$



$$M_2(t) = \sum_{j=1}^{k} \lambda_j \boldsymbol{\phi}_{t/j} \otimes \boldsymbol{\phi}_{t/j}, M_3 = \sum_{j=1}^{k} \lambda_j \boldsymbol{\phi}_{t/j} \otimes \boldsymbol{\phi}_{t/j} \otimes \boldsymbol{\phi}_{t/j}$$



V: vocabulary size; *k*: subtopic number

computing: 0.03
machinery: 0.01
intelligence: 0.04
:

length 1

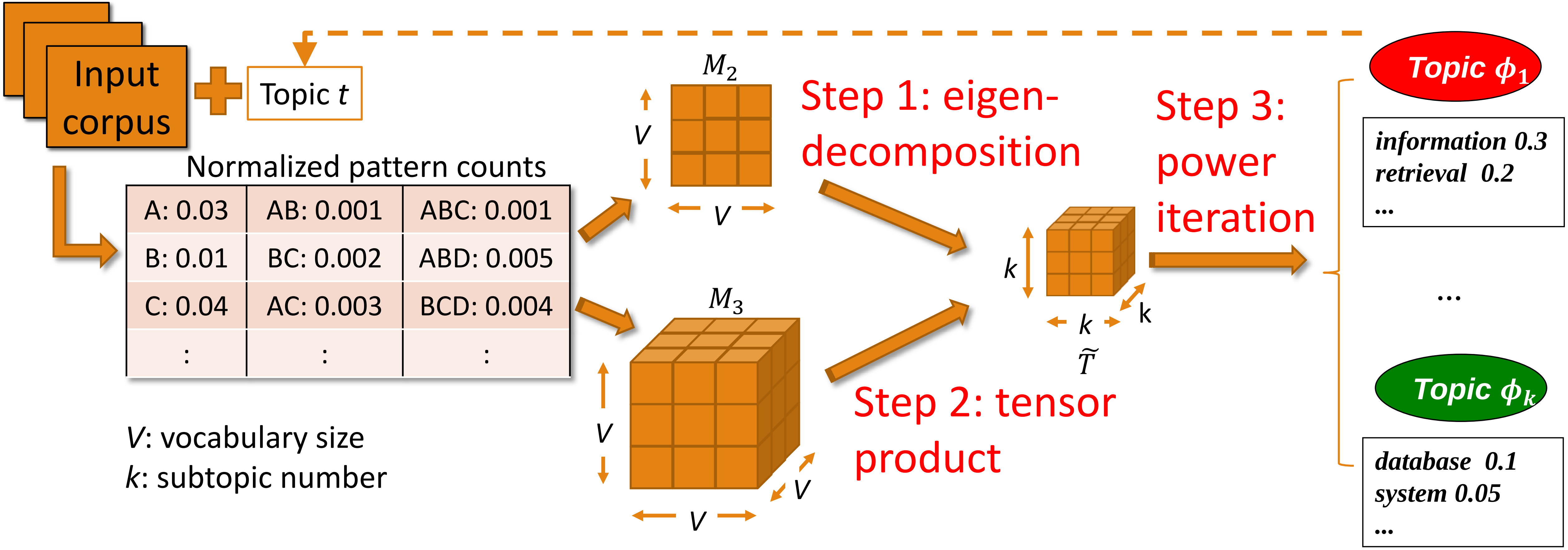
computing machinery: 0.001
computing intelligence: 0.002
machinery intelligence: 0.003
:

length 2 (pair)

computing machinery intelligence: 0.001
support vector machines: 0.005
conditional random fields: 0.004
:

length 3 (triple)

Tensor Orthogonal Decomposition for *EXP* operator



Input corpus

Topic *t*

Normalized pattern counts

A: 0.03	AB: 0.001	ABC: 0.001
B: 0.01	BC: 0.002	ABD: 0.005
C: 0.04	AC: 0.003	BCD: 0.004
:	:	:

V: vocabulary size
k: subtopic number

Step 1: eigen-decomposition

Step 2: tensor product

Step 3: power iteration

Topic ϕ_1

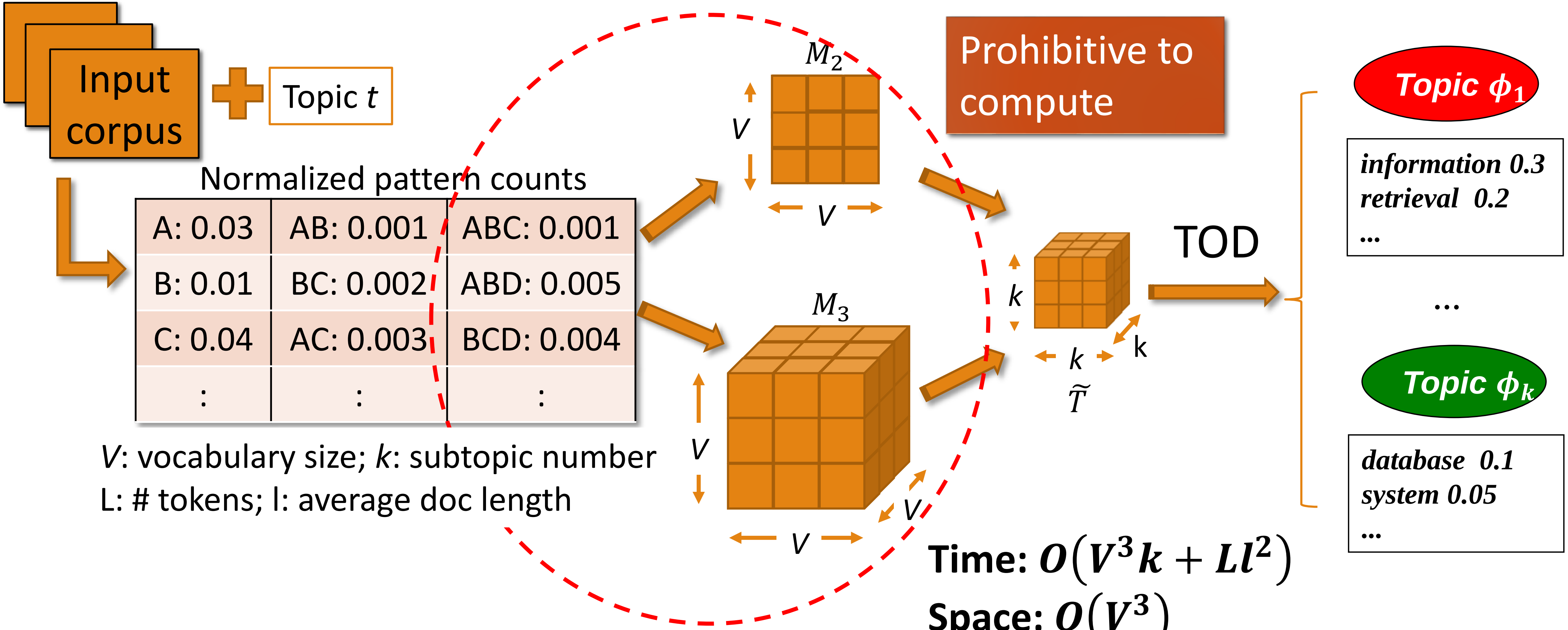
***information* 0.3**
***retrieval* 0.2**
…

…

Topic ϕ_k

***database* 0.1**
***system* 0.05**
…

Tensor Orthogonal Decomposition for *EXP* operator – Not Scalable



Normalized pattern counts

A: 0.03	AB: 0.001	ABC: 0.001
B: 0.01	BC: 0.002	ABD: 0.005
C: 0.04	AC: 0.003	BCD: 0.004
:	:	:

information 0.3
retrieval 0.2
…

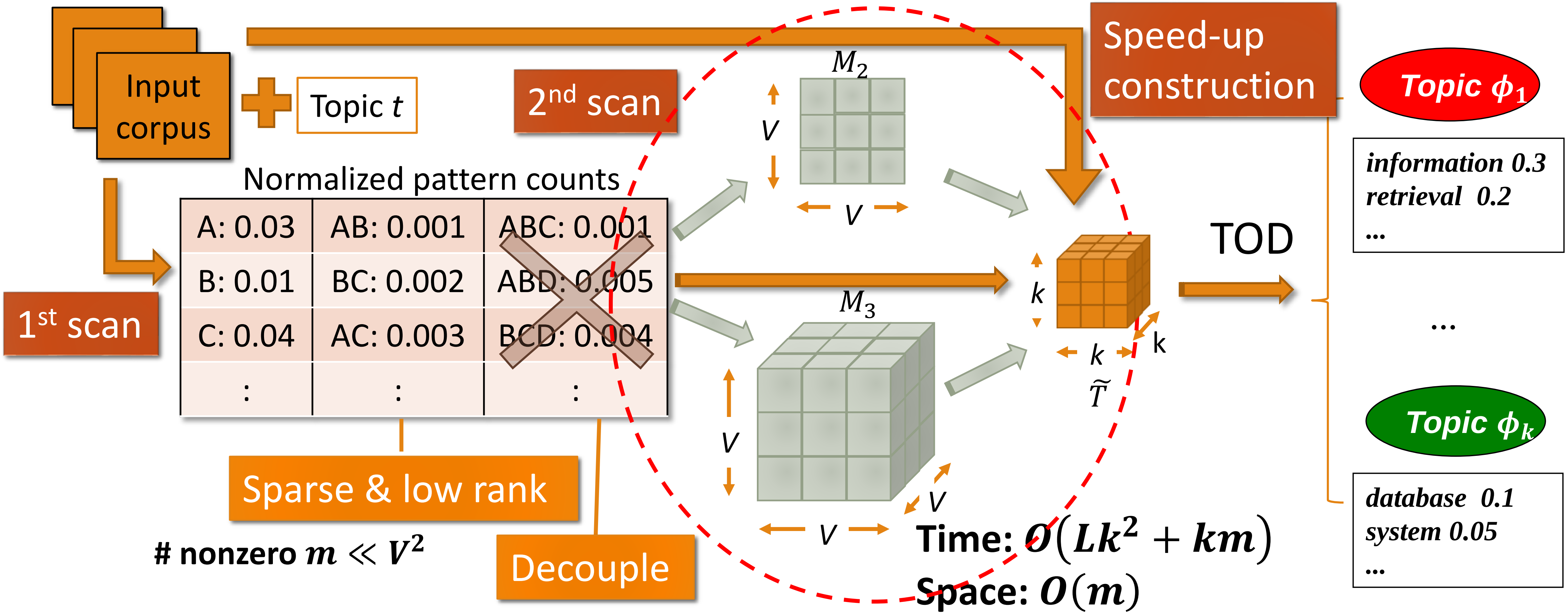
database 0.1
system 0.05
…

V: vocabulary size; *k*: subtopic number
L: # tokens; l: average doc length

Time: $\boldsymbol{O\left(V^3k + Ll^2\right)}$

Space: $\boldsymbol{O\left(V^3\right)}$

Scalable Tensor Orthogonal Decomposition



Normalized pattern counts

A: 0.03	AB: 0.001	~~ABC: 0.001~~
B: 0.01	BC: 0.002	~~ABD: 0.005~~
C: 0.04	AC: 0.003	~~BCD: 0.004~~
:	:	~~:~~

# nonzero $m \ll V^2$

information 0.3
retrieval 0.2
…

database 0.1
system 0.05
…

Time: $O(Lk^2 + km)$

Space: $O(m)$

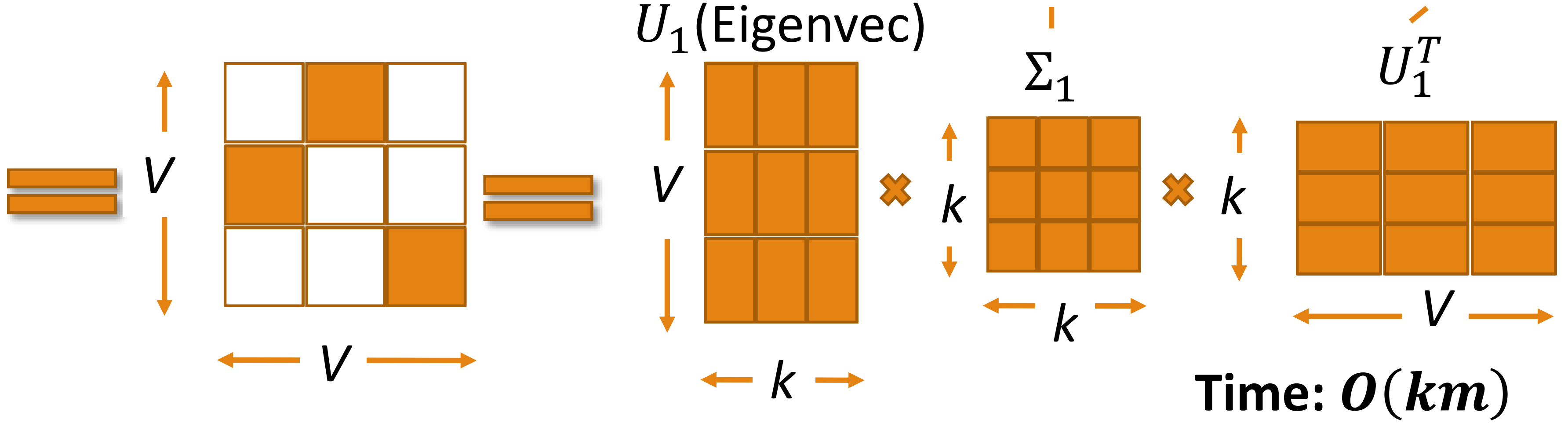
Find Eigen-Decomposition of M_2
Step 1: Eigen-Decomp. of A Sparse Matrix

$$M_2 = E_2 - c_1 E_1 \otimes E_1 = \mathrm{U}_1 \widetilde{M_2} \mathrm{U}_1^{\mathrm{T}} \in \mathbb{R}^{V*V}$$

$$\Sigma_1 - c_1 (U_1^T E_1) \otimes (U_1^T E_1)$$

E_2 (Sparse)

AB: 0.001
BC: 0.002
AC: 0.003
:



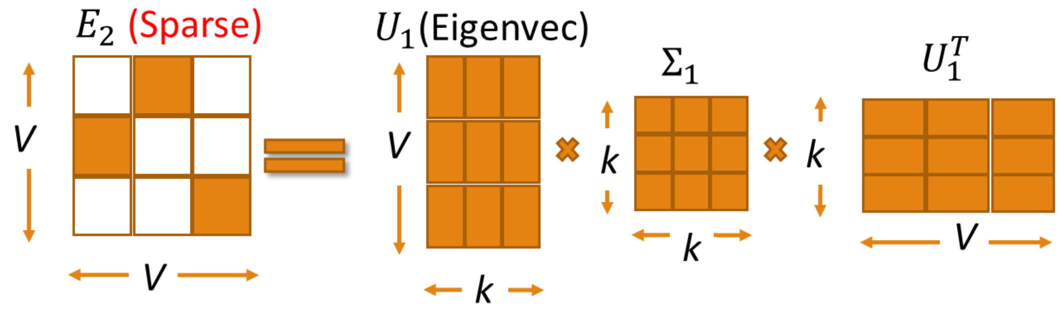
Time: $O(km)$
Space: $O(m)$

Find Eigen-Decomposition of M_2
Step 2: Eigen-Decomp. of A Small Matrix

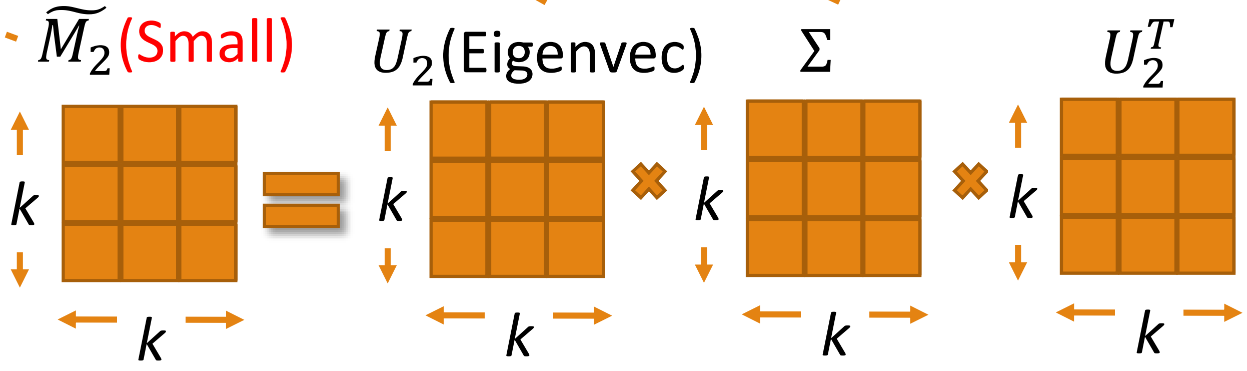
$$M_2 = (U_1 U_2)\Sigma(U_1 U_2)^T = \mathrm{M}\Sigma\mathrm{M}^{\mathrm{T}}$$

Step 1. Eigen-decomposition of E_2

$\Rightarrow \quad (M_2 = U_1 \widetilde{M}_2 U_1^T)$

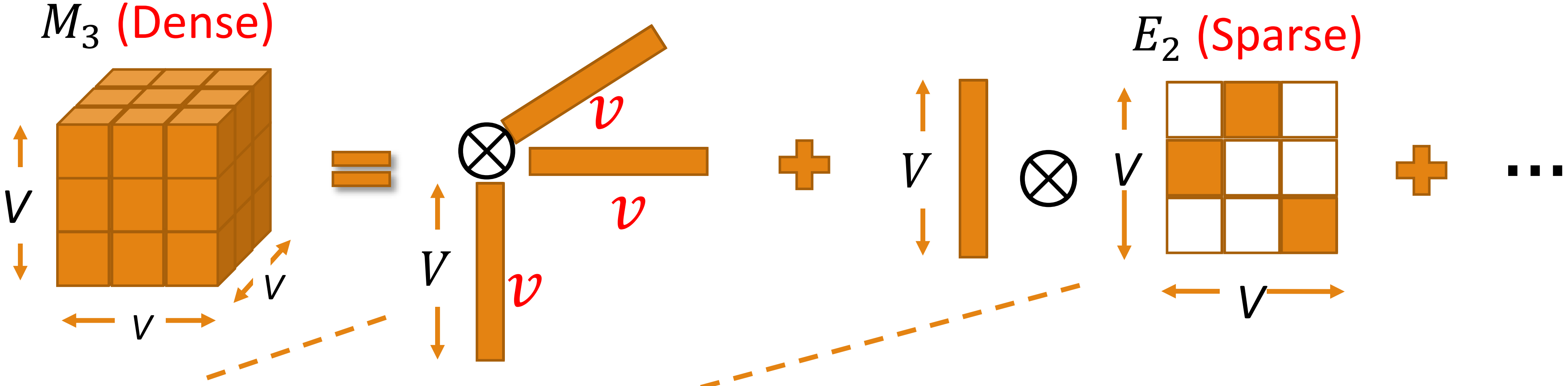


Step 2. Eigen-decomposition of $\widetilde{M}_2$ (**small**)



Construction of Small Tensor Via Decoupling of Third-Order Moment

$$\tilde{T} = M_3(W, W, W) \qquad W = M\Sigma^{-\frac{1}{2}}, W^T M_2 W = I$$



$$(v^{\otimes 3})(W, W, W) = (W^T v)^{\otimes 3}$$

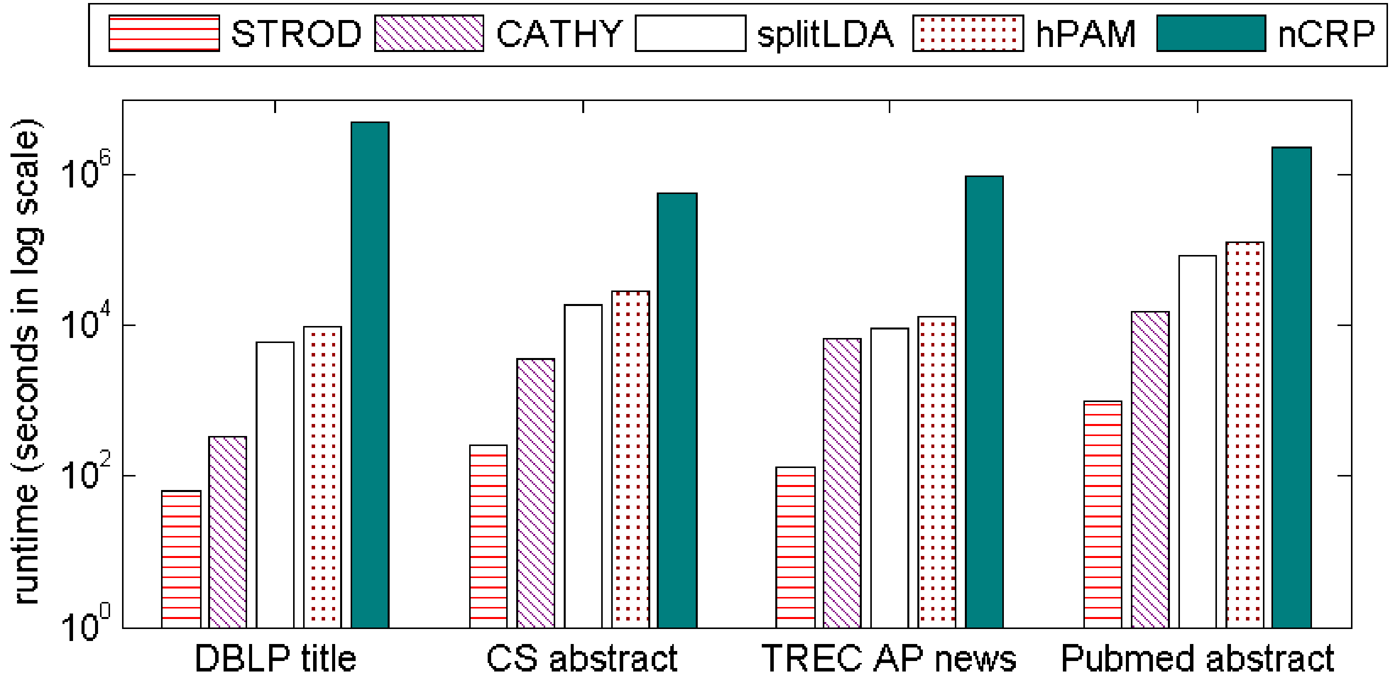
$$(v \otimes E_2)(W, W, W) = W^T v \otimes W^T E_2 W$$

Time: $O(Lk^2)$
Space: $O(Vk)$

Efficiency

- Several orders of magnitude faster
- Three scans vs. thousands of scans

STROD – Scalable tensor orthogonal decomposition
CATHY – EM algorithm for network-based clustering
splitLDA – Recursively apply LDA
hPAM – hierarchical Pachinko Allocation Model
nCRP – nested Chinese Restaurant Process



#Tokens	11M	39M	19M	169M
#Docs	1.9M	529K	106K	1.5M

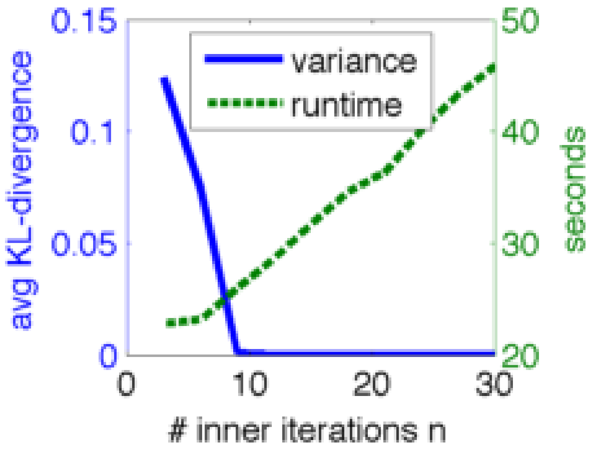
Consistency & Quality

- Variance is almost 0
- Convergence is fast
- Good performance in topic intrusion study

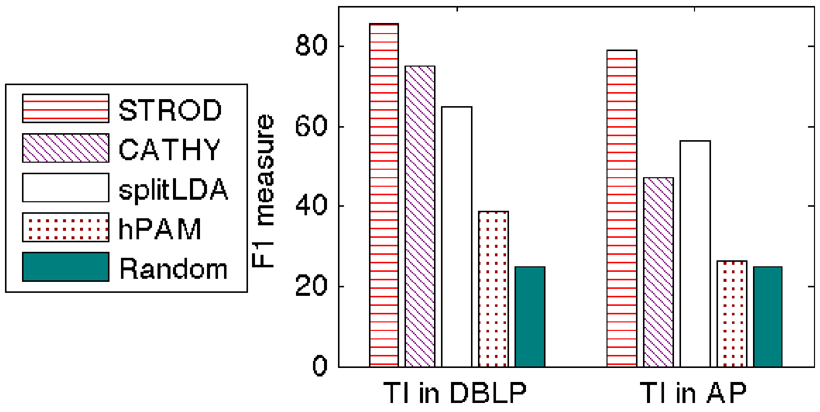
(a) Variance (avg KL-divergence) across 10 runs

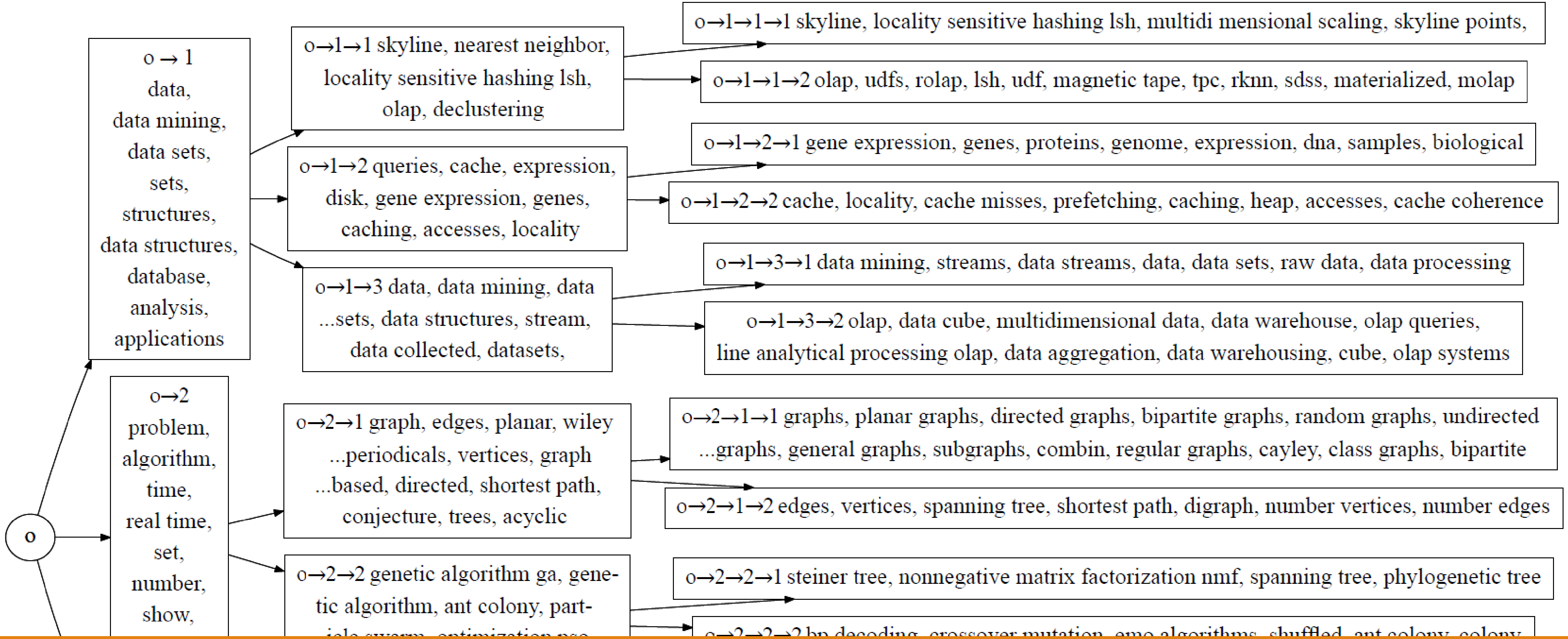
Method	DBLP title	CS abstract	TREC AP news	Pubmed abstract
hPAM	5.58	5.72	5.89	Too slow
splitLDA	3.39	1.60	1.58	Too slow
CATHY	17.3	1.96	1.42	3.12
STROD	**0.611**	**0.000138**	**0.00452**	**0.000527**

(b) Variance w.r.t # iterations



(c) Quality





o

o → 1 data, data mining, data sets, sets, structures, data structures, database, analysis, applications

o→1→1 skyline, nearest neighbor, locality sensitive hashing lsh, olap, declustering

o→1→1→1 skyline, locality sensitive hashing lsh, multidi mensional scaling, skyline points,

o→1→1→2 olap, udfs, rolap, lsh, udf, magnetic tape, tpc, rknn, sdss, materialized, molap

o→1→2 queries, cache, expression, disk, gene expression, genes, caching, accesses, locality

o→1→2→1 gene expression, genes, proteins, genome, expression, dna, samples, biological

o→1→2→2 cache, locality, cache misses, prefetching, caching, heap, accesses, cache coherence

o→1→3 data, data mining, data ...sets, data structures, stream, data collected, datasets,

o→1→3→1 data mining, streams, data streams, data, data sets, raw data, data processing

o→1→3→2 olap, data cube, multidimensional data, data warehouse, olap queries, line analytical processing olap, data aggregation, data warehousing, cube, olap systems

o→2 problem, algorithm, time, real time, set, number, show,

o→2→1 graph, edges, planar, wiley ...periodicals, vertices, graph ...based, directed, shortest path, conjecture, trees, acyclic

o→2→1→1 graphs, planar graphs, directed graphs, bipartite graphs, random graphs, undirected ...graphs, general graphs, subgraphs, combin, regular graphs, cayley, class graphs, bipartite

o→2→1→2 edges, vertices, spanning tree, shortest path, digraph, number vertices, number edges

o→2→2 genetic algorithm ga, genetic algorithm, ant colony, particle swarm optimization pso

o→2→2→1 steiner tree, nonnegative matrix factorization nmf, spanning tree, phylogenetic tree

o→2→2→2 bp decoding, crossover mutation, emo algorithms, shuffled, ant colony, colony

Phrase-Represented Hierarchy Sample

Conclusion

1. Interactive operators help topic hierarchy curators; the challenge is consistency and efficiency

2. A solution based on scalable tensor recursive orthogonal decomposition:
   - A new hierarchical topic model that supports consistent operators
   - One operator requires at most three scans of the whole corpus
   - Fast runtime, Low variance, and high quality

References

1. *[Griffiths et al. 04]* T. Griffiths, M. Jordan, J. Tenenbaum, and D. M. Blei. *Hierarchical topic models and the nested chinese restaurant process*, NIPS'04.
2. *[Li & McCallum 06]* W. Li, A. McCallum. *Pachinko allocation: Dag-structured mixture models of topic correlations*, ICML'06.
3. *[Mimno et al. 07]* D. Mimno, W. Li, A. McCallum. *Mixtures of hierarchical topics with pachinko allocation*, ICML'07.
4. *[Kim et al. 12a]* J. H. Kim, D. Kim, S. Kim, and A. Oh. *Modeling topic hierarchies with the recursive chinese restaurant process*, CIKM'12.
5. *[Anandkumar et al. 12]* A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, M. Telgarsky. *Tensor decompositions for learning latent variable models*, arXiv:1210.7559, 2012.
6. *[Ahmed et al. 13]* A. Ahmed, L. Hong, A. Smola. *Nested chinese restaurant franchise process: Applications to user tracking and document modeling*, ICML'13.
7. [*Pujara & Skomoroch 12]* J. Pujara and P. Skomoroch. *Large-scale hierarchical topic models. In NIPS Workshop on Big Learning, 2012.*

Code and data are available at http://illimine.cs.uiuc.edu/software/strod/