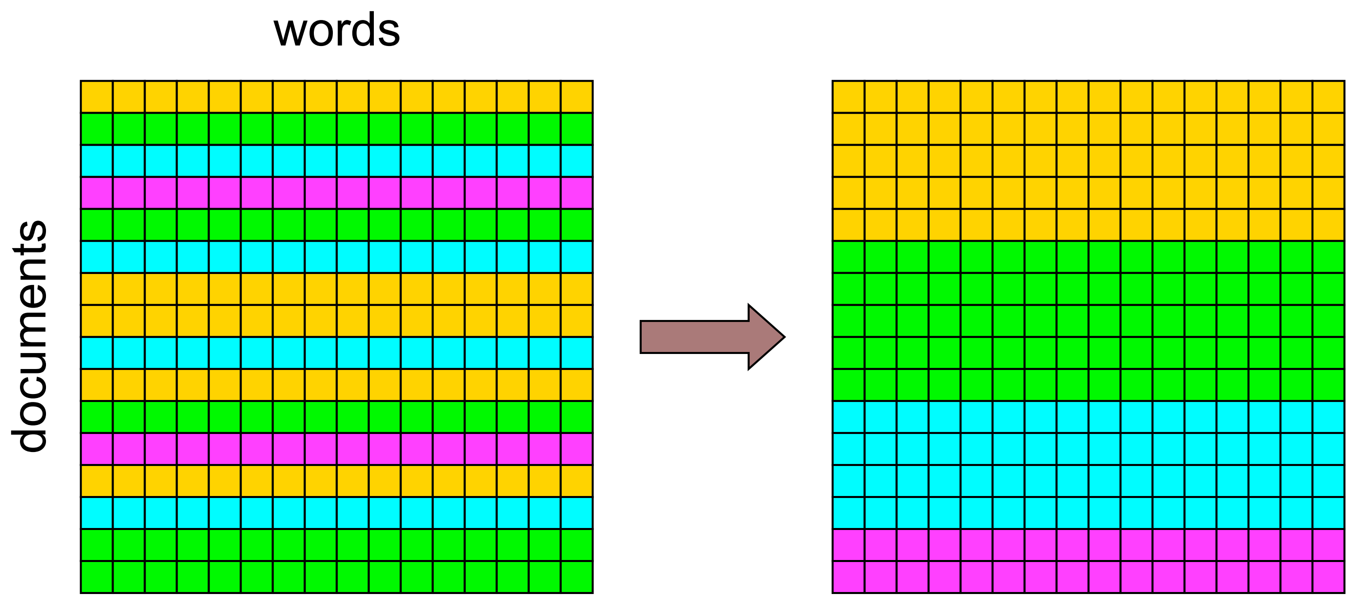
Multi-way Distributional Clustering via Pairwise Interactions

Ron Bekkerman *UMass*
Ran El-Yaniv *Technion*
Andrew McCallum *UMass*

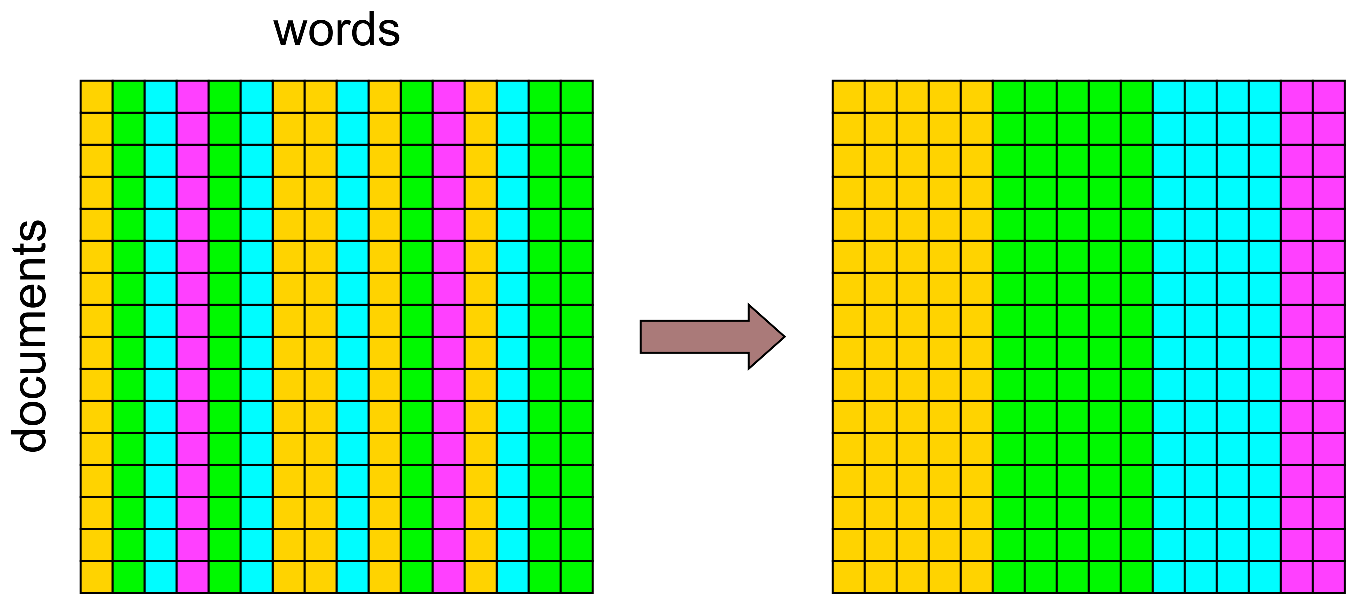
Email: ronb@cs.umass.edu

Contingency table clustering



- Not necessarily documents!
  - Images/features, genes/samples, movies/actors…

Contingency table clustering



- Similarly we can cluster columns (words)

Two-way clustering

Or both!

Two-way clustering

a.k.a. *Double* clustering, *Bi*-clustering, *Co*-clustering, *Coupled* clustering etc.

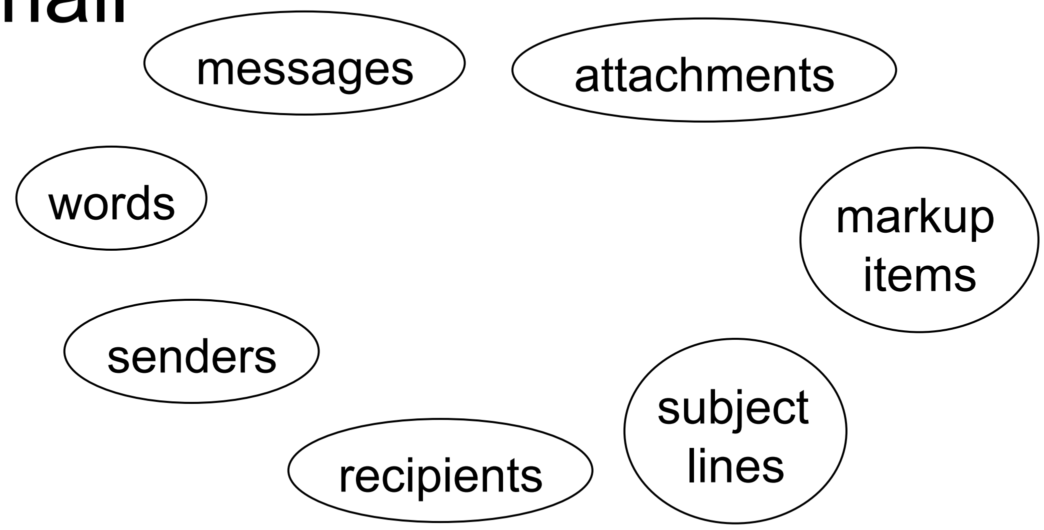
- Main motivation:
  - Can overcome statistical sparseness
- Extensively studied:

Slonim & Tishby, SIGIR-2000
Getz, Levine & Domany PNAS-2000
El-Yaniv & Souroujon NIPS-2001
Dhillon, Mallela & Modha KDD-2003

- All showing impressive improvements

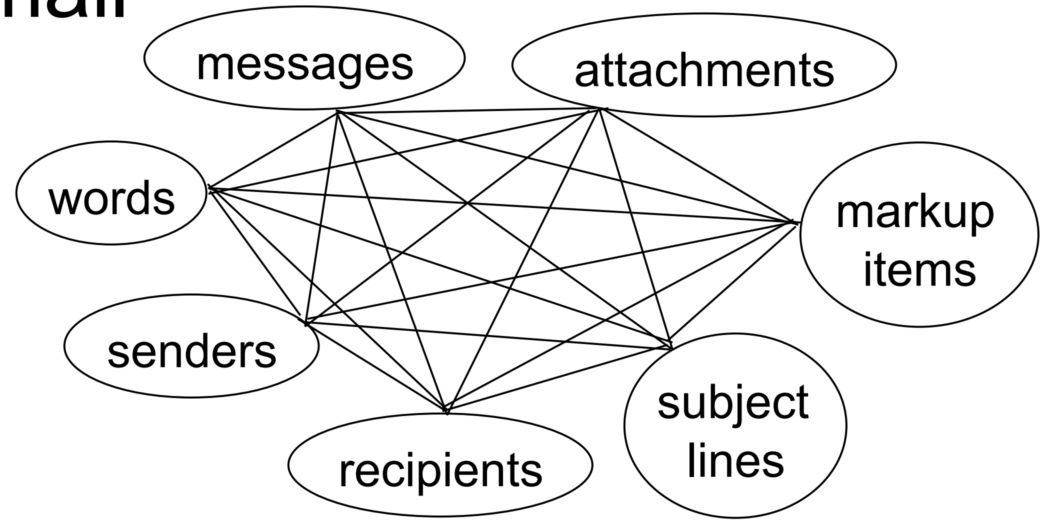
Multiple views

- Various views of data can be observed
- **Example**: email



Multiple views

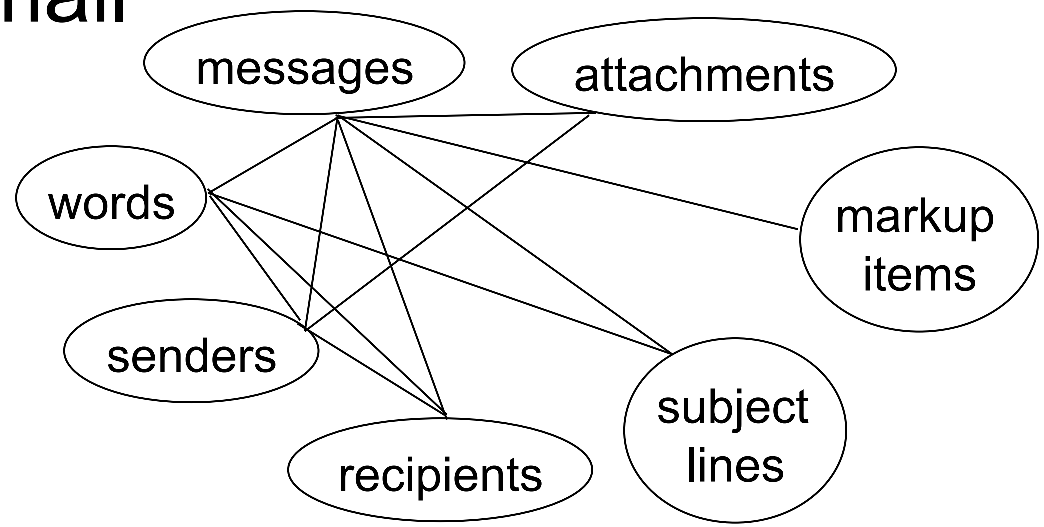
- Various views of data can be observed
- **Example**: email



- Statistical interaction of views

Multiple views

- Various views of data can be observed
- **Example**: email



- Statistical interaction of views
  - Not necessarily all interactions are relevant

Multi-way clustering

- **Motivating question:** can we extend two-way clustering to utilize multiple views?

- **Goal:** construct N “clean” clusterings of N interdependent variables

Our contributions

- **Objective function** for fitting useful multi-way interaction model
- **Novel clustering algorithm** to maximize the objective
- **Striking empirical results**

Earlier attempts

- Multivariate Information Bottleneck (mIB)

Friedman et al. UAI-2001

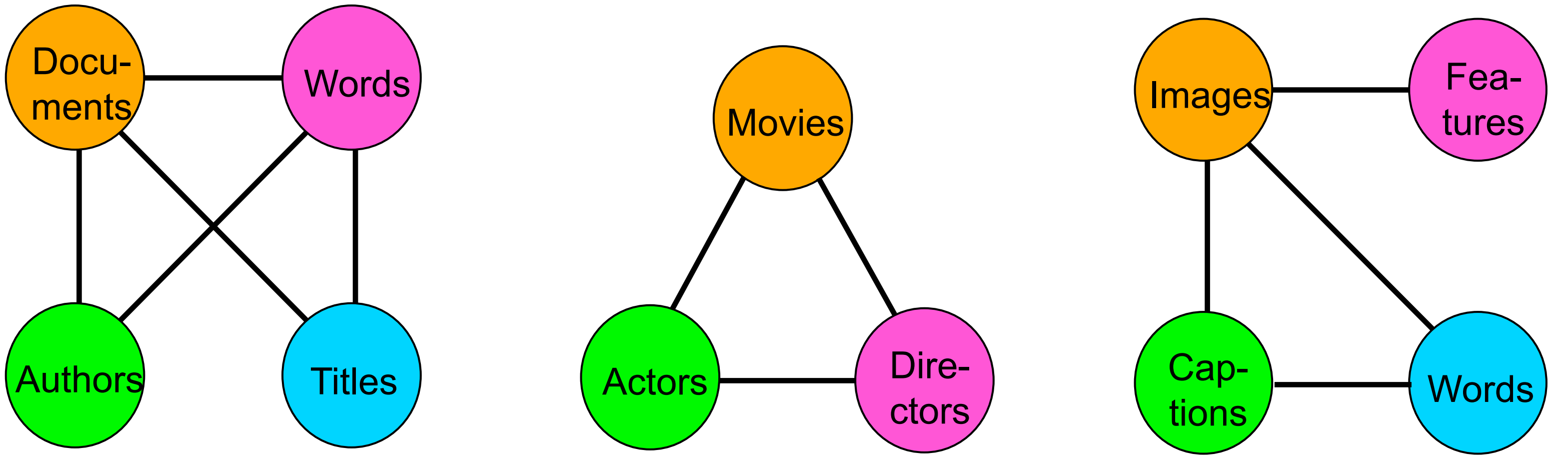
  - Very general approach for dealing with several variables
  - **Objective:** Multi-Information

$$\mathbf{I}(\widetilde{X}_1;\ldots;\widetilde{X}_N) = \sum_{\widetilde{x}_1,\ldots,\widetilde{x}_N} P(\widetilde{X}_1,\ldots,\widetilde{X}_N)\log\frac{P(\widetilde{X}_1,\ldots,\widetilde{X}_N)}{P(\widetilde{X}_1)\ldots P(\widetilde{X}_N)}$$

- Not feasible for practical applications

Our approach

- Consider only pairwise interactions
- Pairwise interaction graph
  - Defines interactions between $N \geq 2$ variables

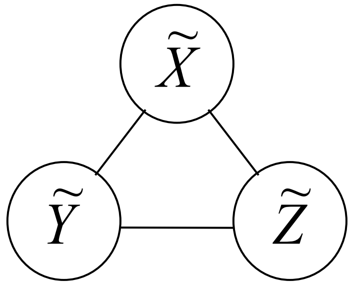


Our objective

- Let $(\widetilde{\boldsymbol{X}}, \boldsymbol{E})$ be pairwise interaction graph
- Extending Dhillon et al.:
- **Objective:** weighted sum of pairwise MI
  - $$\boxed{\max_{\widetilde{X}_1,\ldots,\widetilde{X}_N} \sum_{(\widetilde{X}_i,\widetilde{X}_j)\in \boldsymbol{E}} w_{ij}\, \mathrm{I}(\widetilde{X}_i;\widetilde{X}_j)}$$
  - Subject to $\left|\widetilde{X}_i\right| = K_i,\ \ i = 1,\ldots,N$
- No multi-dimensional probability tables
- Can be easily factorized

Objective factorization

- Consider triangle:



- Objective in this case:
  - $\max_{\tilde{X},\tilde{Y},\tilde{Z}} \; w_1 \, \mathrm{I}(\tilde{X};\tilde{Y}) + w_2 \, \mathrm{I}(\tilde{Y};\tilde{Z}) + w_3 \, \mathrm{I}(\tilde{X};\tilde{Z})$
- …is broken into 3 parts:
  - $\max_{\tilde{X}} \; w_1 \, \mathrm{I}(\tilde{X};\tilde{Y}) + w_3 \, \mathrm{I}(\tilde{X};\tilde{Z})$
  - $\max_{\tilde{Y}} \; w_1 \, \mathrm{I}(\tilde{X};\tilde{Y}) + w_2 \, \mathrm{I}(\tilde{Y};\tilde{Z})$
  - $\max_{\tilde{Z}} \; w_2 \, \mathrm{I}(\tilde{Y};\tilde{Z}) + w_3 \, \mathrm{I}(\tilde{X};\tilde{Z})$

Implementation

- We have tried various schemes:
  - Top-down
  - Bottom-up
  - Flat (K-means, sequential IB)
- Best results obtained with **hybrid**
  - Top-down for some variables
  - Bottom-up for other variables
  - Flat correction routine after each split/merge

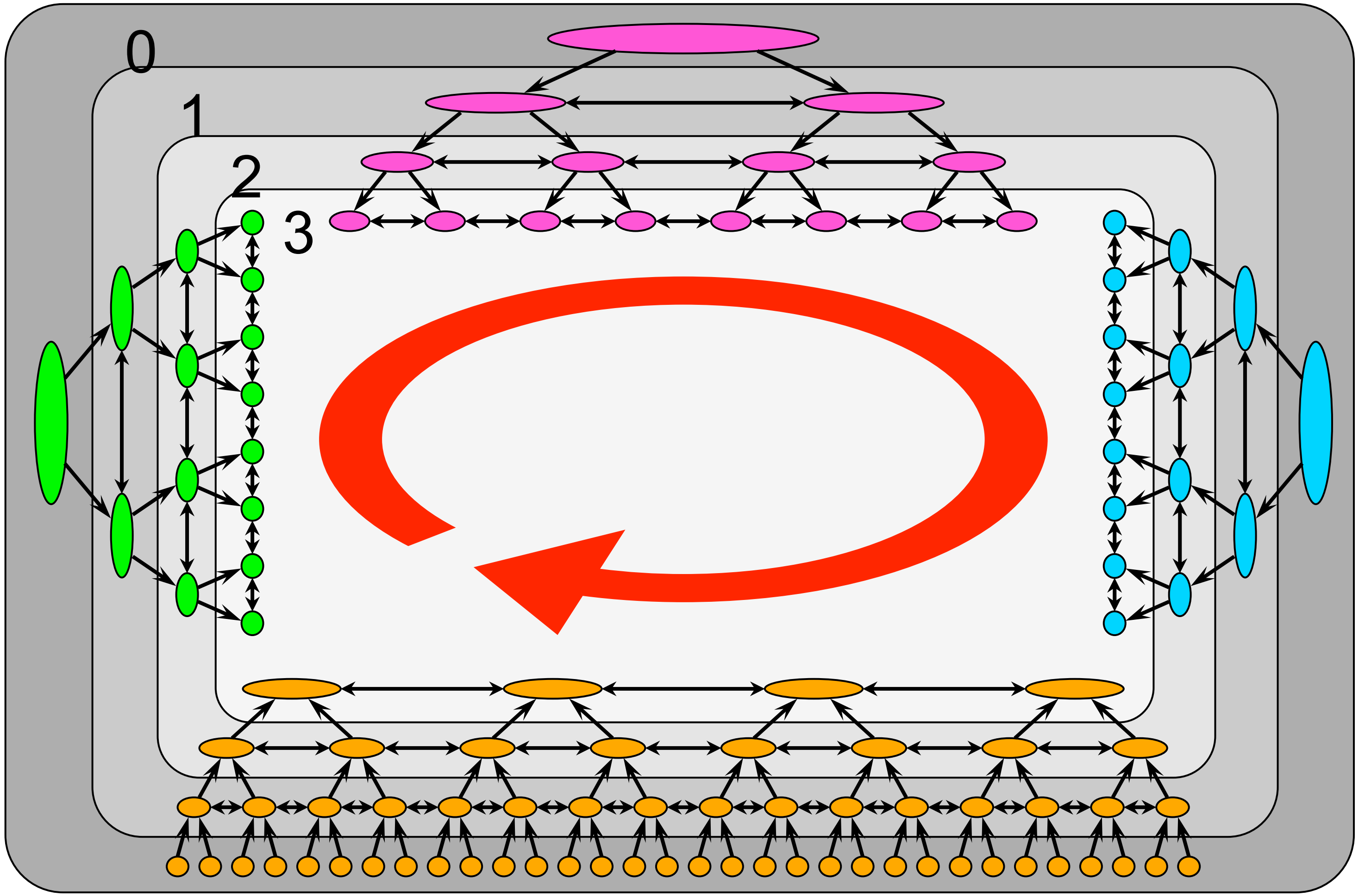
Multi-way Distributional Clustering

MDC

- Initialization
  - If $i \in S^{up}$, put each x_i in a singleton cluster
  - If $i \in S^{down}$, put all x_i in one common cluster
- Main loop
  - If $i \in S^{up}$, merge every two closest clusters
  - If $i \in S^{down}$, split each cluster to two halves
  - Correction loop
    - Pull each x_i out of its cluster
    - Put it into $\tilde{x}_i$ s.t. the objective is maximized

Slonim et al.
SIGIR-2002

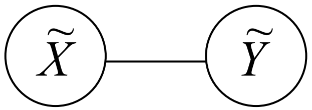
Example: 4-way MDC



Computational complexity

- General case
  - At each iteration of the main loop:
  - Pass over all x_i
  - Pass over all $\tilde{x}_i$
  - Pass over all $\tilde{x}_j, \ \forall j \neq i$

  $O(|X_i|^3)$
- If bottom-up system is only one $\Rightarrow$ $o(|X_i|^3)$
- 2-way case

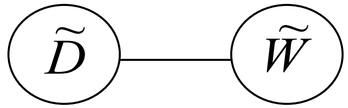


  - At each iteration $|\tilde{x}_i|$ is doubled
  - While $|\tilde{y}_i|$ is halved

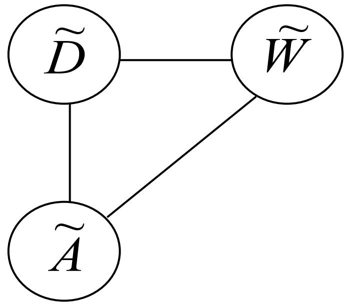
  $O(|X_i|^2)$

Experimental setup

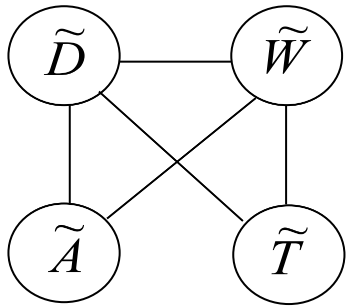
- 2-way MDC
  - ***D***ocuments and ***W***ords



- 3-way MDC
  - ***D***ocuments, ***W***ords and ***A***uthors

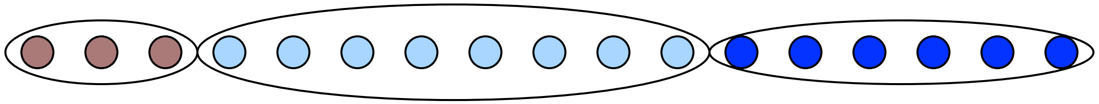
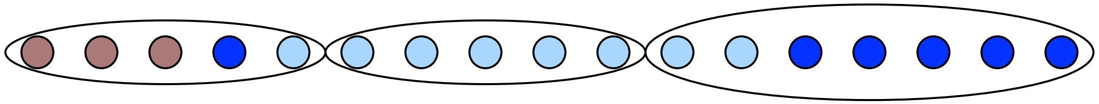


- 4-way MDC
  - ***D***ocuments, ***W***ords, ***A***uthors and documents' ***T***itles



- Documents: bottom-up, the rest: top-down

Evaluation methodology

- Clustering evaluation
  - Is generally unintuitive
  - Is an entire ML research field
- We use the "accuracy" measure
  - Following Slonim et al. and Dhillon et al.
  - Ground truth: 
  - Our results: 
  - $$Acc = \frac{1}{|X|}\sum_{c} \gamma_c$$

Size of dominant class in cluster c

Datasets

- Three CALO email datasets:
  - acheyer: 664 messages, 38 folders
  - mgervasio: 777 messages, 15 folders
  - mgondek: 297 messages, 14 folders
- Two Enron email datasets:
  - kitchen-l: 4015 messages, 47 folders
  - sanders-r: 1188 messages, 30 folders
- The 20 Newsgroups: 19997 messages

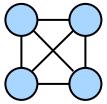
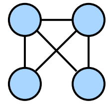
Results

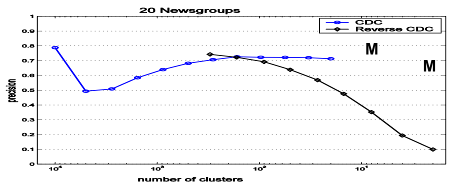
	Slonim et al.	Dhillon et al.	2-way MDC	3-way MDC	4-way MDC	SVM
acheyer	44.7±0.6	47.0±0.2	48.1±0.7	50.5±0.4	52.1±0.8	65.8±2.9
mgervasio	40.2±2.3	36.6±1.6	44.9±1.2	48.6±0.8	54.2±0.6	77.6±1.0
mgondek	62.1±1.4	69.5±1.6	77.1±1.4	80.8±1.2	81.6±1.0	92.6±0.8
kitchen-l	33.2±0.5	33.0±0.3	41.9±0.7	38.5±0.2		73.1±1.2
sanders-r	64.8±0.4	59.3±1.2	67.7±0.3	67.1±0.8		87.6±1.0
20NG	61.0±0.7	57.7±0.2	71.8±0.7			91.3±0.3

Improvement over the baseline

	Slonim et al.	Dhillon et al.	2-way MDC	3-way MDC	4-way MDC	SVM
acheyer		47.0±0.2	+1.1	+3.5	+5.1	
mgervasio	40.2±2.3		+2.7	+8.4	+14.0	
mgondek		69.5±1.6	+7.6	+11.3	+12.1	
kitchen-l	33.2±0.5		+8.7	+5.3		
sanders-r	64.8±0.4		+2.9	+2.3		
20NG	61.0±0.7		+10.8			

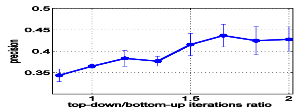
More results

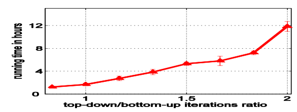
- 4-way: full graph  **<** original graph 
  - Some interactions are unnecessary
- 2-way: agglomerative MDC **≈** original MDC
  - But it is dramatically less efficient
    - Would run 300 times longer on 20NG
- Documents can be clustered top-down
  - And words bottom-up
  - Eventually close results



Even more results

- Scheduling matters

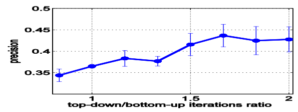


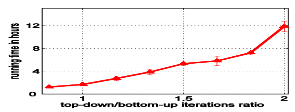


  - More splits usually improve the results
  - MDC runs 7X slower with split/merge ratio = 2
- Social network analysis
  - Goal: to cluster people in a social network
  - Tested on Melinda Gervasio’s email
    - She created 4 groups of correspondents
    - 62.3% accuracy with 4 clusters, 76.6% precision with 8 clusters

Even more results

- Scheduling matters





  - More splits usually improve the results
  - MDC runs 7X slower with split/merge ratio = 2
- Social network analysis
  - Goal: to cluster people in a social network
  - Tested on Melinda Gervasio’s email
    - She created 4 groups of correspondents
    - 62.3% accuracy with 4 clusters, 76.6% precision with 8 clusters

Discussion

- Improvement over Slonim et al.
    - Which is a 1-way clustering algorithm
  - Shows that **multi-modality** helps
- Improvement over Dhillon et al.
    - Which is a 2-way clustering algorithm
  - Shows that **hierarchical setup** helps
- MDC is an efficient method
  - Which allows exploring complex models
    - 3-way, 4-way etc.

Conclusion

- Unsupervised model without generative assumptions
- Exploit multiple views of your data
- Efficient algorithm
- Impressive empirical results ☺

Future work

- Inference of optimal schedule
- Inference on “optimal” number of clusters?
- Extend to semi-supervised setup

Future work

- Inference of optimal schedule
- Inference on "optimal" number of clusters?
- Extend to semi-supervised setup

Thanks

MDC 0.1 can be downloaded from http://www.cs.umass.edu/~ronb/mdc.html