Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts

Vu Nguyen[1], **Dinh Phung**[1], Long Nguyen[2], S. Venkatesh[1] and Hung Bui[3]

[1] Centre for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia
[2] Department of Statistics, University of Michigan, Ann Arbor, USA
[3] Laboratory for Natural Language Understanding, Nuance Communications, Sunnyvale, USA

Grouped Data, Content and Context

- Data content naturally present themselves in groups
  - From now refer to groups as documents

To Wong Binghao [redacted] My Best Friend.

« previous entry | next entry »
Mar. 15th, 2010 | 11:44 am
mood: grateful
music: Glee Cast: Sweet Caroline

Recent Entries
Friends
Archive
User Info

im trying to think of a way to begin this without coming across as an obessed stalker but.. what the heck, i dont really give a damn what people think anymore.

I MISS YOU WONG BINGHAO.
you've been in there for about 2 weeks alr, & i miss you dreadfully.
i dont even think you'll be reading this, but just wanted to get it off my chest haha.

Maximum entropy discrimination

Tommi Jaakkola[1]
tommi@ai.mit.edu
Marina Meila[1]
mmp@ai.mit.edu
Tony Jebara[2]
jebara@media.mit.edu
[1] MIT AI Lab, 545 Technology Square, Cambridge, MA 02139
[2] MIT Media Lab, 20 Ames Street, Cambridge, MA 02139

August 18, 1999

Abstract

We present a general framework for discriminative estimation based on the maximum entropy principle and its extensions. All calculations involve distributions over structures and/or parameters rather than specific settings and reduce to relative entropy projections. This holds even when the data is not separable within the chosen parametric class, in the context of anomaly detection rather than classification, or when the labels in the training set are uncertain or incomplete. Support vector machines are naturally subsumed under this class and we provide several extensions. We are also able to estimate exactly and efficiently discriminative distributions over tree structures of class-conditional models within this framework. Preliminary experimental results are indicative of the potential in these techniques.

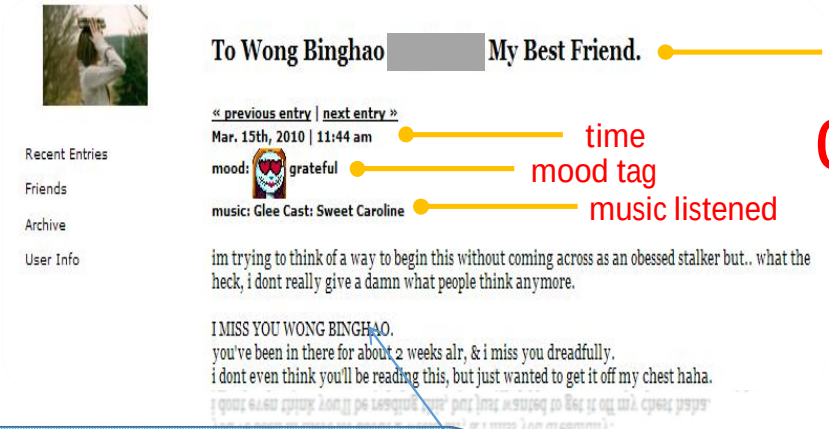
content

hawaii
maui
hdr

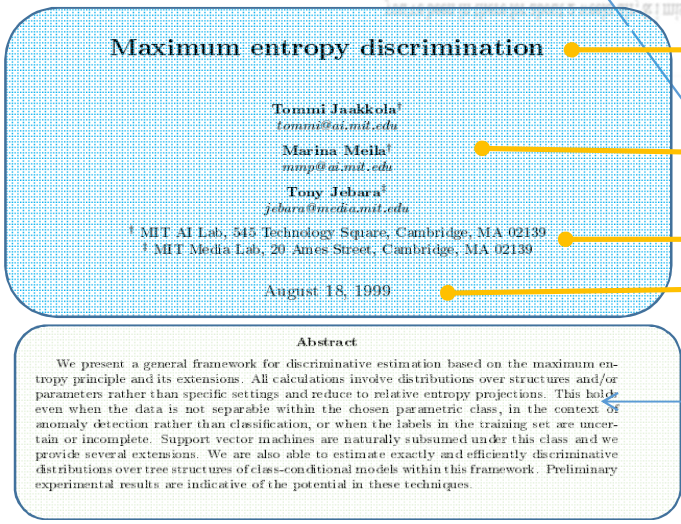
tree
building
person
woman bending
woman standing
tree
bench
window
roof
sidewalk
road
sky
cloud

Grouped Data, Content and Context

- Data content naturally present themselves in groups
  - From now refer to groups as documents

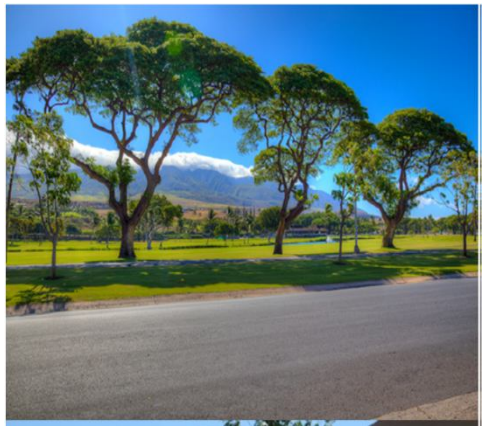


context



content

image tags



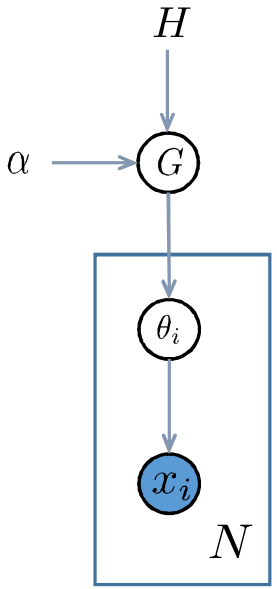
hawaii
maui
hdr

tree
building
person
woman bending
woman standing
tree
bench
window
roof
sidewalk
road
sky
cloud

Grouped Data, Content and Context

- Goal: jointly discover clusters contents and contexts (e.g., words and spatial locations.
- Multiple advantages:
  - Context-aware topic modelling of contents
  - Context clusters sharing content topics
  - Infer context given content and vice-versa
- Currently, no principled way for jointly model both contents and document contexts.

Context clustering: DP Mixtures



- Infinite mixture model.
- Random measure

$$G \sim \mathrm{DP}(\alpha H) \qquad \theta_i \overset{\text{iid}}{\sim} G \qquad x_i \sim F\left(.|\theta_i\right)$$

- Random partition (Chinese Restaurant Process)

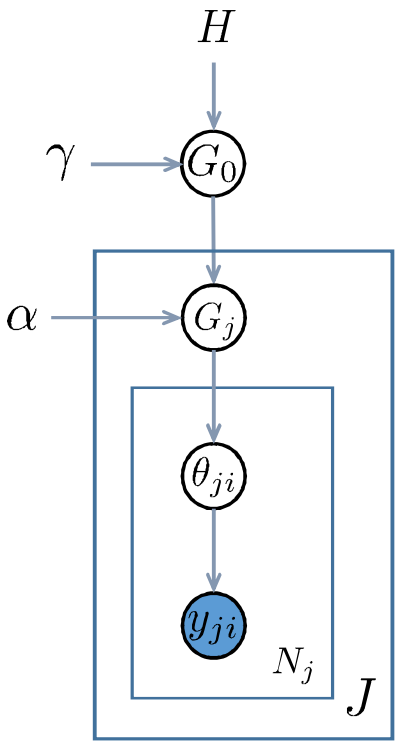
$$\mathbf{z} \sim \mathrm{CRP}_{\overline{1,J}}(\alpha), \forall c \in \mathbf{z}, \theta_c \overset{\text{iid}}{\sim} H$$

$$\mathrm{CRP}_S(\mathbf{z} \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+|S|)} \alpha^{|\mathbf{z}|} \prod_{c \in \mathbf{z}} \Gamma\left(|c|\right)$$

- Stick-breaking

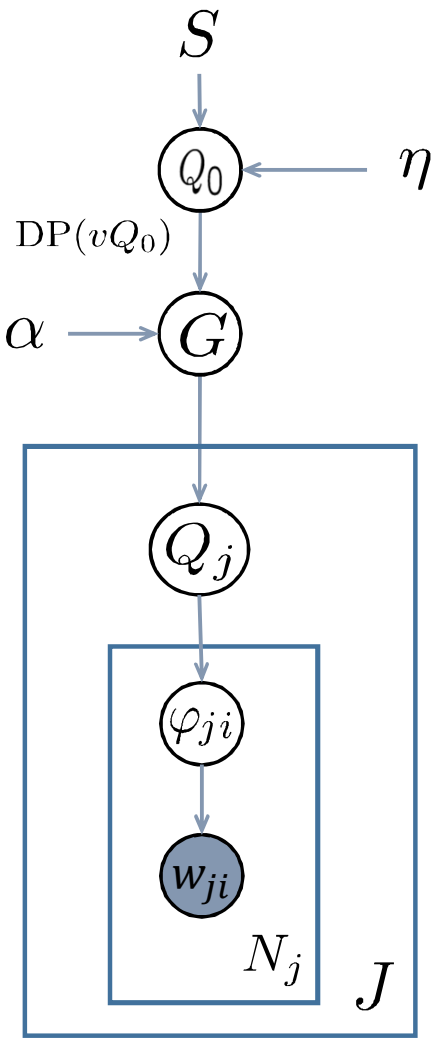
$$\pi \sim \mathrm{GEM}(\alpha) \quad z_i \overset{\text{iid}}{\sim} \pi \quad \psi_k \overset{\text{iid}}{\sim} H \quad \theta_i = \psi_{z_i}$$

Content Topic Modelling: HDP



- Cluster contents/words into topics, shared across documents.
- Do not cluster documents, i.e., $P(G_j = G_{j'}) = 0$
- Cannot exploit context during topic modelling.
- Document clustering may be achieved in a cascaded fashion:
  - Find content topics
  - Find topic-mixing coefficients for each document
  - Clustering documents based on these coefficients

Document Clustering: Nested DP



- Use DP as a base measure of another DP:

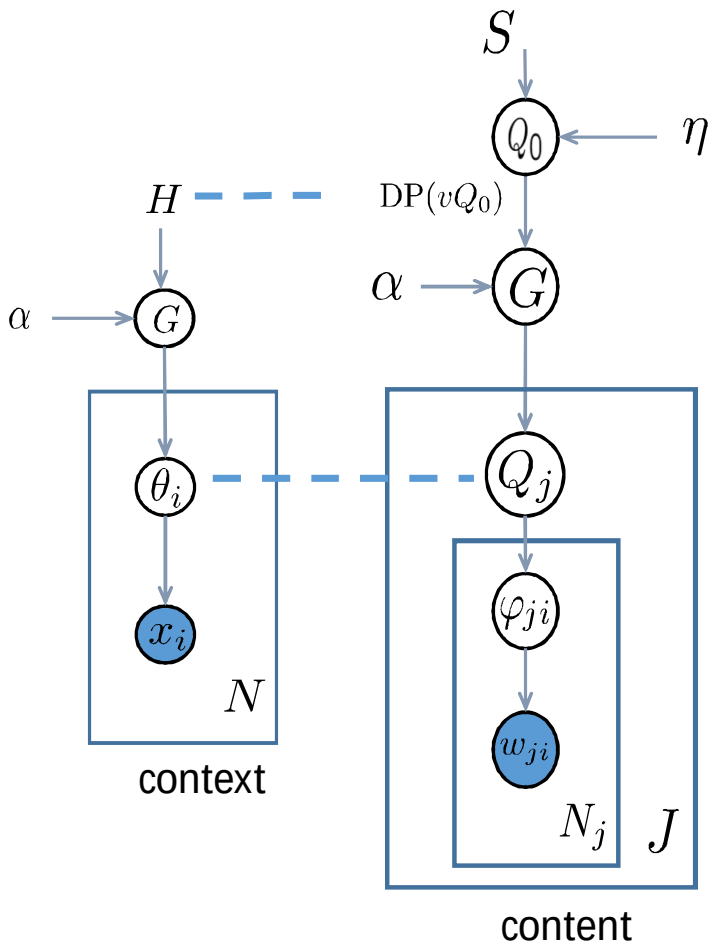
$$G \sim \mathrm{DP}(\alpha \mathrm{DP}(vQ_0))$$

- G is discrete, but its atoms are measures: they are content-generating distributions.
- Drawing from discrete G effectively clusters documents: $P(Q_j = Q_{j'}) = \frac{1}{\alpha+1}$
- Documents in the same cluster have the same content-generating mixture distributions Q_j
- Impose a DP prior for Q_0 enable Q_j (s) in different documents to share topics.

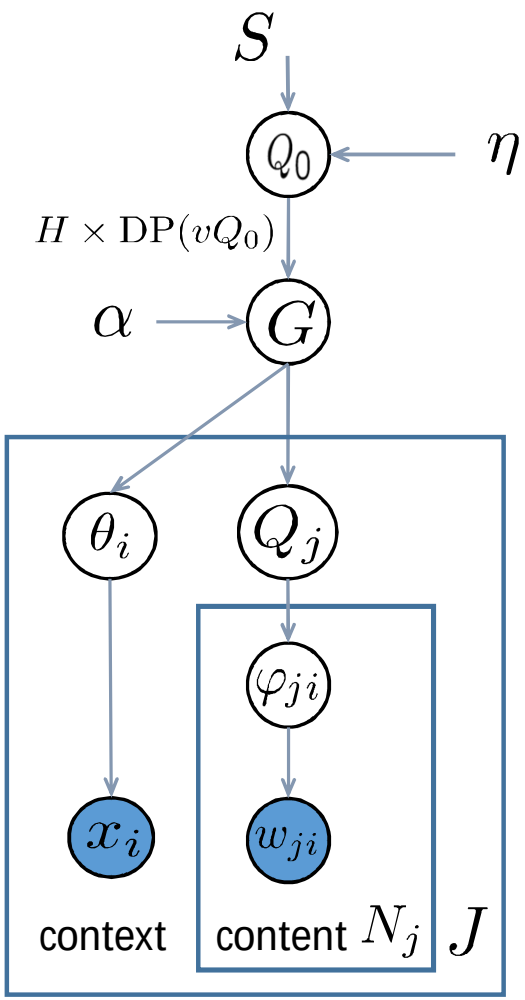
Joint Content and Context Model: MC^2

- Pairing context (document-level) with content (word-level) is unnatural since they lie on different levels.
- One possibility: treat context as index for distributions over contents
  - But, *raw contextual data cannot be used as index* (e.g., noisy tags, continuous location coordinates)
- Our idea: introduce distributions over contexts
  - *Context cluster* acts as an index into a distribution of contents.
  - This allows context (time/space) to influence both topics and document clusters.
- How to make this concrete?

Joint Content and Context Model: MC^2

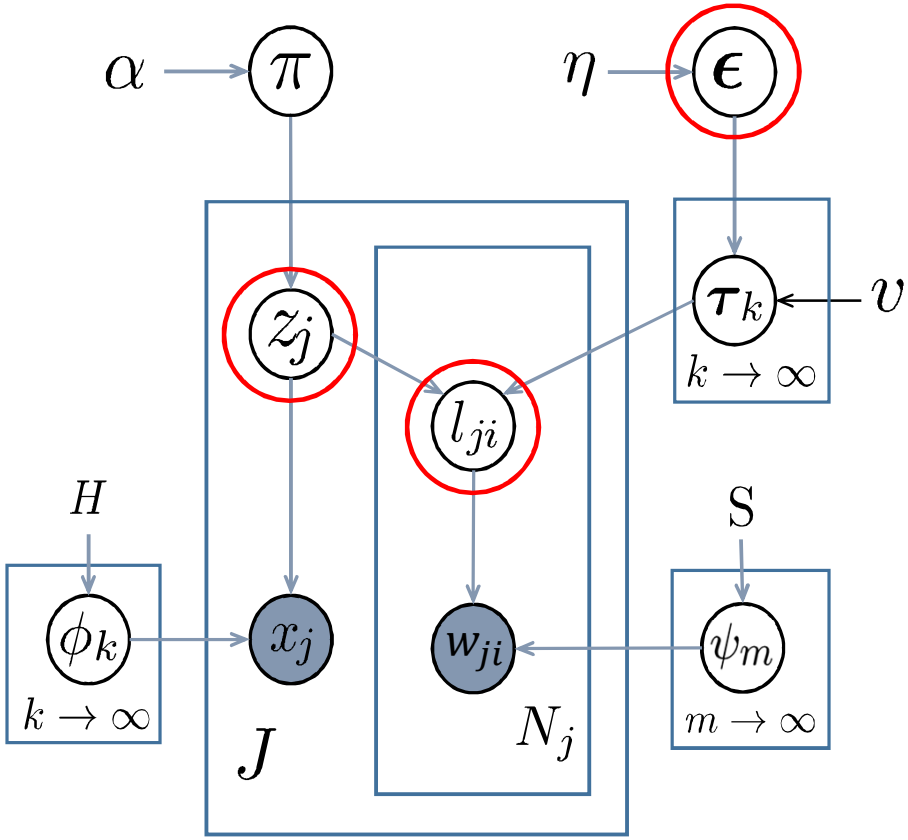


Joint Content and Context Model: MC^2



- Form a product of context-generating base-measure and content-generating DP: $H \times DP(vQ_0)$
- Use this as a base-measure in the nested DP framework.
- Marginalizing content yields a DP mixture over context.
- Marginalizing context yields an nDP mixture of contents
- See paper for proof.

Model representation



Stick-Breaking View

Variables sampled during collapsed Gibbs

Application I: document modelling

- PNAS dataset
  - 79,800 documents (only titles and timestamps)
  - Vocabulary size is 36,782
  - Content observation is word
  - Context observation is timestamp (last 90 years, 1915 – 2005)
- NIPS dataset
  - 1740 documents
  - Vocabulary size is 13,649
  - Content observation is word
  - Three types of context information:
    - Timestamp (1987 – 1999)
    - Author information (2037 unique authors)
    - Titles

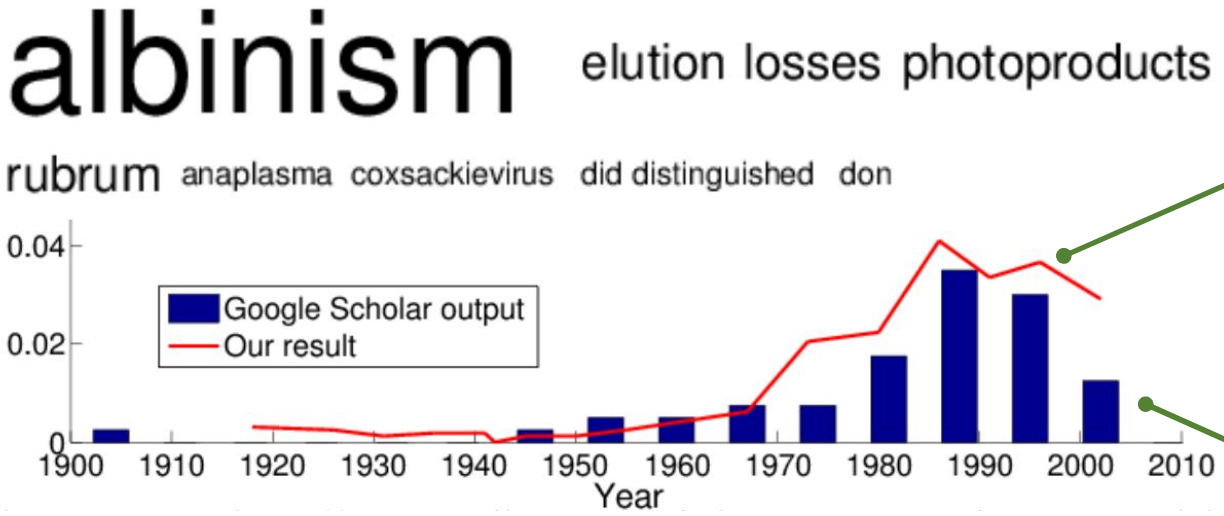
Perplexity Evaluation

Method	Perplexity (*on words only*)				Feature used
	PNAS	(K,M)	NIPS	(K,M)	
HDP (Teh et al., 2006b)	3027.5	$(-, 86)$	1922.1	$(-, 108)$	words
npTOT (Dubey et al., 2012; Phung et al., 2012)	2491.5	$(-, 145)$	1855.33	$(-, 94)$	words+timestamp
MC² without context	1742.6	$(40, 126)$	1583.2	$(19, 61)$	words
MC² with titles	–	–	1393.4	$(32, 80)$	words+title
MC² with authors	–	–	1246.3	$(8, 55)$	words+authors
MC² with timestamp	**895.3**	$(12, 117)$	**984.7**	$(15, 95)$	words+timestamp

Note: missing results are due to title and author information not available in PNAS dataset).
(K,M): (# document clusters, #word topics).

Time as Context

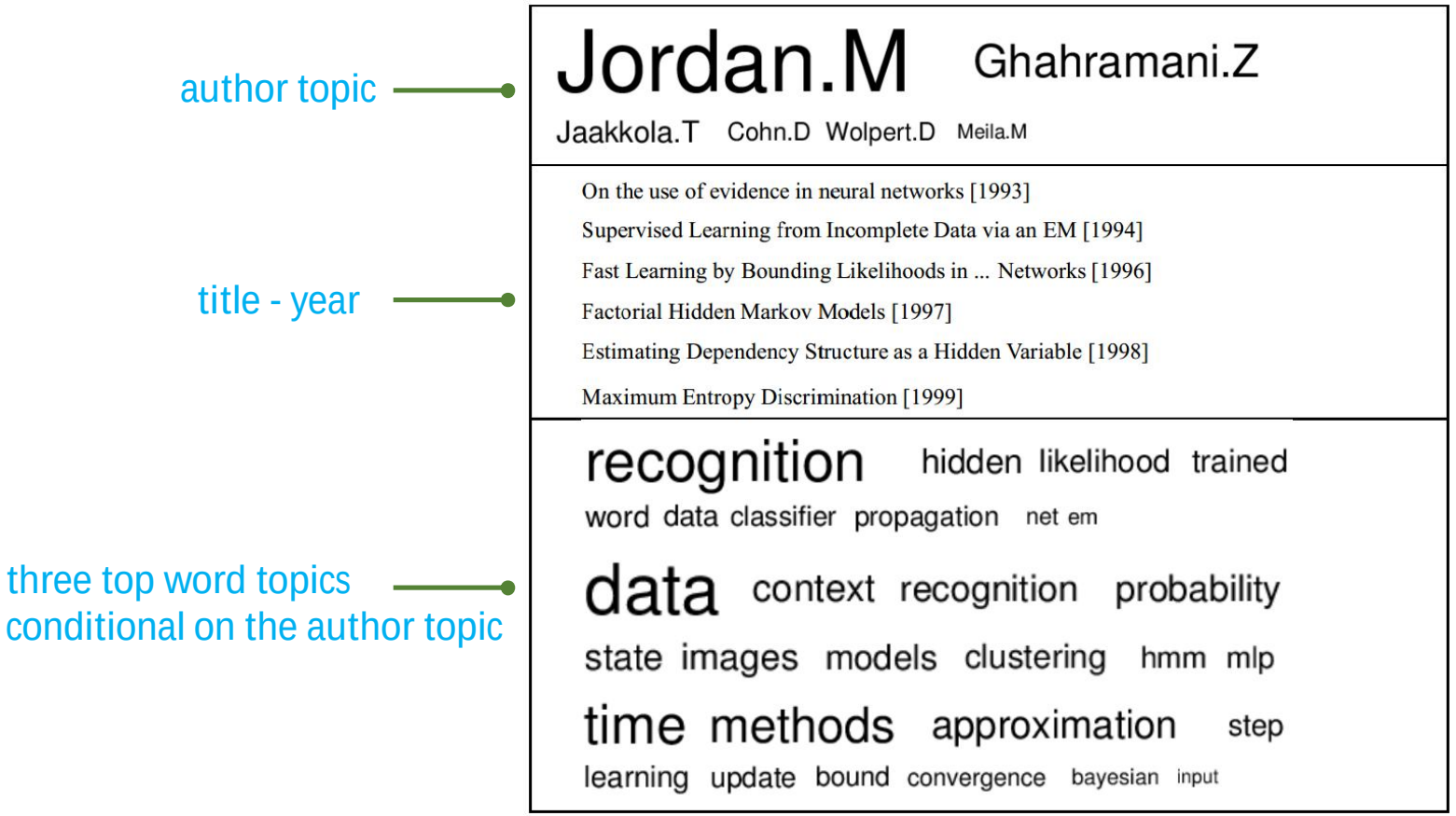
word topic



conditional distribution of timestamp context on word topic

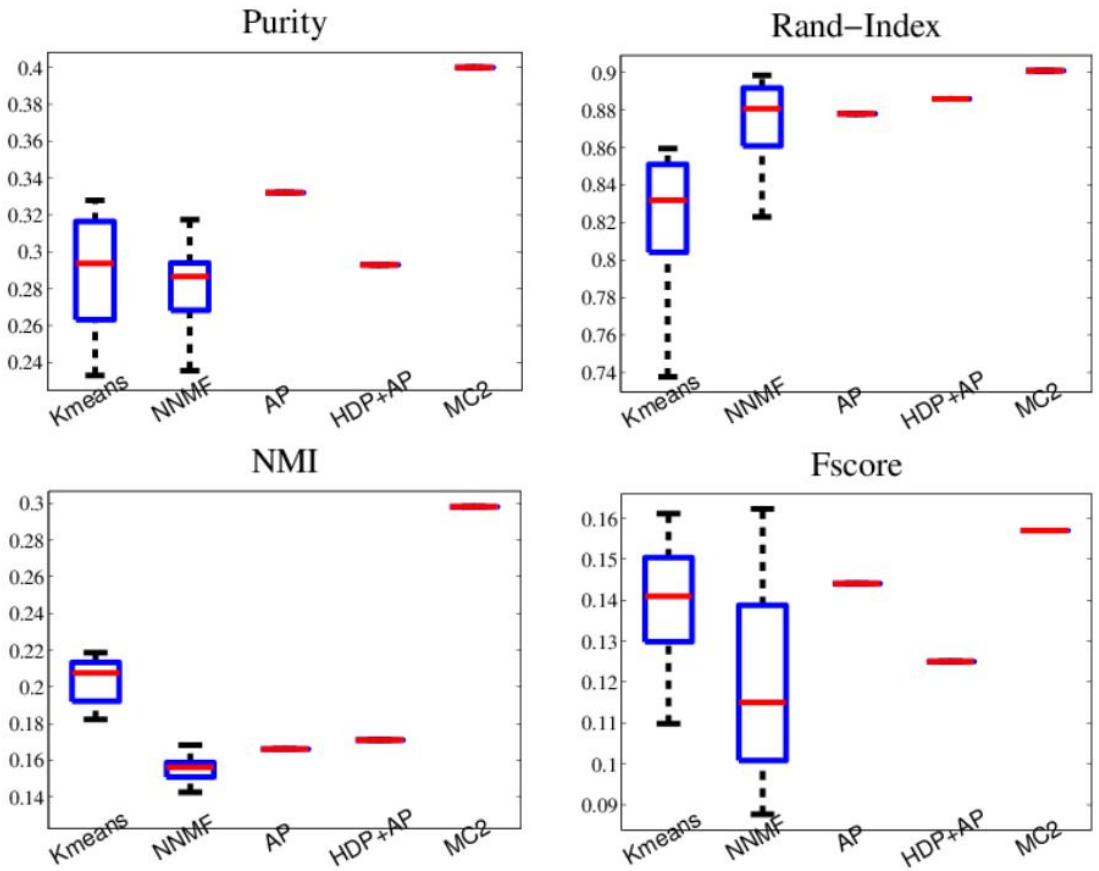
Manual count statistic of keyword “albinism” in Google Scholar

Author as Context



Application II: Image Clustering

- NUSWIDE 13-animal dataset
  - 13 classes (2054 images)
  - Content: SIFT (500 dimension)
  - Context: tags (1000 dimension)



Influence of Context

- What happens when the amount of context varies?
  - Easy to implement when some contexts are missing
  - Vary the percentage of context missing from 100 to 0

Missing(%)	Purity	NMI	RI	F-score
0%	0.407	0.298	0.901	0.157
25%	0.338	0.245	0.892	0.149
50%	0.32	0.236	0.883	0.137
75%	0.313	0.187	0.860	0.112
100%	0.306	0.188	0.867	0.119

- Big jump from 0 to 50%, but after that seems to be negligible, suggesting a good cut-off for computational saving.

Conclusion

- Jointly cluster documents and discover content topics while exploiting context
  - Principled framework to leverage contextual information
  - Demonstrate the importance of context even during topic modelling of contents.
  - Applicable to many types of contexts (time, location, tags, ages, patient's medical information).
- Fully nonparametric Bayesian
  - Automatically infers dimensions of latent structures (i.e., #document clusters, #topics).
  - An elegant framework that combines DP and nDP with nice marginalization property.
- Future work
  - Readily to generalize to arbitrary grouping levels with nested contexts.
  - Multilevel supervised learning: multilevel regression and classification.