Scalable Sparse Bayesian Nonparametric and Matrix Tri-factorization Models for Text Segmentation, Topical analysis and Entity Resolution in Dyadic data.

Ranganath B.N.[1] and Shalabh Bhatnagar[1]

[1]Indian Institute of Science, Bangalore, India

Block Exchangeable model (BEM):Segmentation

- Segmentation of collections of sequence data
  - Laptop review dataset in to facets
  - News dataset in to stories.
- Occurence of facets or the stories
  - Order independent
  - Persistence according to popularity
- Exchangeability in the context of segmentation
  - Specifies the type of permutations of assignments to the random variables that do not affect joint probability under a model.
  - Complete or Group Exchangeability
    - All permutations of assignments are equiprobable
    - Segmentation is arbitrary
    - LDA and Hierarchical Dirichlet processes (HDP): No Persistence aspect
    - Parameters linear in number of documents
  - Markov Exchangeability
    - All permutations with same transition count b/w states are equiprobable.
    - No Order independence though distinction between permutations.
    - HMM-LDA, HDP-HMM and Sticky HDP-HMM: Inference process expensive
    - Quadratic number of transition parameters

BEM:Solution

- Block Exchangeability (BE)
  - All permutations preserving the same number of self-transitions and number of nonself transitions for each state along with the ending state are equiprobable.
  - Amenable to segmentation, Preserves Order Independence.
  - Sparser number of transition parameters (linear).
  - Scalable Inference
  - BE superclass of CE and subclass of ME.
- Core part of BEM

$$G \sim DP(\delta, H), P_i^* \sim Beta(a, b) \quad for \quad i = 1 : \infty$$
$$\pi_{ij} = (1 - P_i^*)G(\phi_j) + P_i^*\delta(i, j), Z_{ij} \sim (\pi_{(Z_{i(j-1)})})$$

- Inference
  - Intractable due to coupling of P^* and G.
  - Decoupled with introduction of Persistence indicator variables C
    - $C_{ij} = 0/1$, continuity of state/ new state
    - Sample Z_{ij} only when $C_{ij} = 1$

BEM:Experiments on News dataset

Model	*fC*1	*IT*	*Perplexity*	*Avg*$_s$
Sticky HDP-HMM	1.0	5.4 sec	3141	**0.3**
BEM	0.12	**0.3** sec	**1103**	0.32

Sparse Matrix Trifactorization (SMTF):Motivation

- Topical analysis of dyadic data
  - Product review data, movie data and bibliographic data.
  - Genres, research and technology areas, product categories are topics in movies, research domain and the product review data.
  - Dyadic association
  - Movie scripts, product literature, research papers and the corresponding words.
  - Consumers, viewers, researchers and the movie scripts, product literature, research papers.
- Sparse relationships between topics and the domain entities
  - Topics and documents
  - Topics and users
- Existing models
  - Probabilistic Author topic model (ATM)
    - Estimates three associations of interest to us
    - Do not address sparsity
  - Collective matrix factorization (CMF)
    - Uses coupled sparse bi-factorization approach.
    - Factorizes a binary author-document association matrix.

SMTF:Formulation

$$1, \arg\min_{\Phi,\Theta,\mathbb{A}} \frac{1}{2}\|D - \Phi\Theta\mathbb{A}\|_F^2 + \lambda_1 \sum_{j=1}^{n} \|(\Theta\mathbb{A})_j\|_1 + \lambda_2 \sum_{j=1}^{m} \|\Theta_j\|_1 + \lambda_3 \sum_{j=1}^{t} \|\Phi_j\|_1$$
$$\text{s.t. } supp(\mathbb{A}) \subseteq supp(A), \Phi \geq 0, \Theta \geq 0, \mathbb{A} \geq 0,$$
$$2, \arg\min_{\Phi,\Theta,Q,\mathbb{A}} \frac{1}{2}\|D - \Phi Q\|_F^2 + \frac{1}{2}\|Q - \Theta\mathbb{A}\|_F^2 + \lambda_Q \sum_{j=1}^{n} \|Q_j\|_1 + \lambda_\Theta \sum_{j=1}^{m} \|\Theta_j\|_1$$
$$+\lambda_\Phi \sum_{j=1}^{t} \|\Phi_j\|_1 s.t. supp(\mathbb{A}) \subseteq supp(A), \quad \Phi \geq 0, \Theta \geq 0, Q \geq 0, \mathbb{A} \geq 0,$$
$$2.1, Solving \quad \Phi : \quad \{\arg\min_{(\Phi^T)_i \geq 0} \frac{1}{2}\|(D^T)_i - Q^T(\Phi^T)_i\|_F^2 + \lambda_\Phi\|(\Phi^T)_i\|_1\}_{i=1:v}$$
$$2.2, Solving \quad \Theta : \quad \{\arg\min_{(\Theta^T)_i \geq 0} \frac{1}{2}\|(Q^T)_i - \mathbb{A}^T(\Theta^T)_i\|_F^2 + \lambda_\Theta\|(\Theta^T)_i\|_1\}_{i=1:t},$$
$$2.3, Solving \quad Q : \quad \{\arg\min_{Q_i \geq 0} \frac{1}{2}\|[D; \Theta\mathbb{A}]_i - [\Phi; eye(t, t)]Q_i\|_F^2 + \lambda_Q\|Q_i\|_1\}_{i=1:n}$$
$$2.4, Solving \quad \mathbb{A} : \quad \{\arg\min_{\mathbb{A}_j \geq 0} \frac{1}{2}\|Q_j - \Theta\mathbb{A}_j\|_F^2, \mathbb{A}_{ij} = 0 \; \forall i \text{ s.t. } \mathbb{A}_{ij} \notin supp(A)\}_{j=1:n}$$

- Subproblems are solved in an alternating minimization framework using Projected FISTA.
- Proof for convergence for Projected FISTA.

SMTF:Results

Table: Performance comparison between SMTF, CMF and ATM

	DBLP				REV			
Model	TE	DTa-F1	DTa-ARI	ATa-CCD	TE	DTa-F1	DTa-ARI	ATa-CCD
$SMTF_1$	**131240**	0.33	0.19	**423**	**43129**	0.53	0.46	**454**
$SMTF_2$	150870	0.28	0.13	**423**	48376	0.51	0.44	508
CMF	154580	0.25	0.01	428	48588	0.16	0.01	884
ATM	-	**0.35**	**0.24**	518	-	**0.60**	**0.55**	612

Sparse Entity Resolution Model (SERM):Motivation

- Entity resolution in Dyadic data
  - Identify correct author entity for each of the aliases in all the documents
  - Grouping over the documents
    - Documents share author entities, topics
- Sparse relationships
  - Smaller number of author entities and topics per group.
  - Smaller number of aliases for author entities.
- Existing models
  - LDA for Entity resolution model (LDA-ER)
    - Do not utilize textual information for disambiguation of the identical aliases.
    - Do not address sparsity issue.
  - Grouped Author topic model (GATM)
    - Utilizes textual information.
    - Uses HDP nonparametric prior over author entities and topics for the groups.
    - Do not address sparsity issue.

SERM:Model

- Structurally similar model as that of GATM
  - Stick breaking prior of DP: nonparametric in the number of groups
  - Nonparametric sparsity promoting prior- Indian Buffet process (IBP) over author entities for groups.
  - Parametric IBP over topics for the groups leading to scalable solution.
  - k-NN mechanism for selecting smaller number of potential author aliases for the author entities leading to scalability
  - Noise model as that in LDA-ER, GATM employed for generating the aliases.

SERM:Results

Table: Best B-CUBED results for SERM and GATM

	Citeseer		Rexa	
Model	time	F1	time	F1
SERM	**3.2**	**86.06**	**1.3**	**77.65**
GATM	21.6	82.21	14	61.49