Topic Modeling for Segment-based Documents

Giovanni Ponti[1] Andrea Tagarelli[2] George Karypis[3]

[1] ENEA - Portici Research Center, Italy

[2] Dept. DEIS, University of Calabria, Italy

[3] Dept. CSE, University of Minnesota, USA

20th Italian Symposium on Advanced Database Systems
June 24-27, 2012 - Venice, Italy

Outline

Statistical Topic Modeling

Main assumption: text data represented as a mixture of probability distributions over terms

Generative models for documents

- A probabilistic process to express the document features as being generated by a number of latent variables
- Word occurrences modeled by a latent variable
- Each word can be assigned to more class variables, more topics can describe a document

Topic modeling vs. Vector-space text modeling

(Latent) Semantic aspects underlying correlations between words $\Longrightarrow$ document topical structure

Multi-Topic Documents

Naturally comprised of topically-coherent blocks (segments)

- Each of the segments can discuss a theme
- Each theme can be considered as a mixture of topics $\Longrightarrow$ better representation of topical dependence
- Each word may refer to different topics based on the document portions it belongs to

GMs for Multi-Topic Documents

Classic solutions: no full identification of topic correlations

- BOW assumption negatively affects the generative process: every word associated to only one topic across the document

Modeling Topically Segmented Documents

Our Proposal

Introduce a variable modeling the within-document segments $\Longrightarrow$ document as mixture of the topic distributions in the segments

Key Ideas

Contextualize the word-to-topic assignments to segments

- Word generation should depend on topics as well as segments
- Latent topic variable directly associated to segments (rather than to the whole document)

Topic Modeling

PLSA [Hofmann, 2001]

- Probabilistic version of LSA conceived to better handling problems of term *polysemy*
- Generative model for a single text document
- Utilizes a latent variable for statistical topic model, in order to express the *mixture* of distributions within a document

LDA [Blei et al., 2003]

- Generative model for a corpus of text documents
- Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words
- 3-level scheme (corpus, documents, and terms)
- Dirichlet distribution is used to assign documents to topics

Text Segmentation

Subdivision of a text into smaller units (e.g., paragraphs) each discussing a single main topic

Tools: linguistic criteria and statistical similarity measures

TextTiling

- Baseline method for TS (block-similarity-based approach)
- Subdivides a text into multi-paragraph, contiguous, disjoint blocks
- Assumption: terms discussing a subtopic tend to co-occur locally $\Rightarrow$ topic switch detected by the ending/beginning of co-occurrence of a given set of terms
- Segment boundaries are inferred from min values in the sequence of cosine-sim values for all pairs of adjacent blocks

Combining TM and TS

Topic segments tend to be lexically cohesive and a switch to a topic corresponds to a shift in the term distribution

A few proposals:

- On cascade: e.g., topic-based TS using PLSA (CIKM, 2002), TS with LDA-based Fisher kernel (ACL-HLT, 2008)
- Integrated: e.g., hierarchical Bayesian extension of LDA upon text segmented (CAI, 2008)

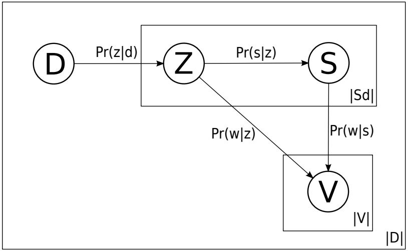
STM [Du et al., 2010]

- Two-param Poisson-Dirichlet process using Gibbs sampler in a hierarchical model
- Extends LDA by adding a level to model the document segments

Notations

- A collection of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$
- A set of words $\mathcal{V} = \{w_1, \ldots, w_M\}$ (vocabulary of $\mathcal{D}$)
- Each document $d \in \mathcal{D}$ is a sequence of n_d words
- A set of (hidden) topics $\mathcal{Z} = \{z_1, \ldots, z_T\}$
  - $\mathcal{Z}$ represents a latent variable model that associates topics (unobserved class variables) with word occurrences (observed data)
- A set of segments $\mathcal{S} = \{S_1, \ldots, S_N\}$
  - Each document $d \in \mathcal{D}$ is assumed to be provided as a set S_d of contiguous, non-overlapping segments
  - No assumption on how segments were detected

SGM overview



1. Select a document d from $\mathcal{D} \Rightarrow \Pr(d)$
2. For each segment $s \in S_d$:
    1. Choose a topic z for the document $d \Rightarrow \Pr(z|d)$
    2. Associate topic-to-segment probability to the segment s for the selected topic $z \Rightarrow \Pr(s|z)$
    3. For each word w in the segment s:
        - Choose a word w from the current topic and segment $\Rightarrow \Pr(w|z, s)$

Probability model

Translation into a joint probability model for triadic data (triad: document, segment, word)

$$\Pr(d, s, w) = \Pr(d) \sum_{z \in \mathcal{Z}} \Pr(z|d) \Pr(s|z) \Pr(w|z, s)$$

Model parameter estimation

E-step:

$$\Pr(z|d,s,w) = \frac{\Pr(z,d,s,w)}{\Pr(d,s,w)} = \frac{\Pr(z|d)\Pr(s|z)\Pr(w|z,s)}{\sum_{z\in\mathcal{Z}}\Pr(z|d)\Pr(s|z)\Pr(w|z,s)}$$

M-step:

$$\mathbf{E}[\mathcal{L}] = \sum_{d\in\mathcal{D}}\sum_{s\in S_d}\sum_{w\in\mathcal{V}} n(d,s,w) \times \sum_{z\in\mathcal{Z}}\Pr(z|d,s,w)\log(\Pr(d,s,w))$$

Update formulas:

$$\begin{aligned}
\Pr(z|d) &\propto \sum_{s\in S_d}\sum_{w\in\mathcal{V}} n(d,s,w)\Pr(z|d,s,w) \\
\Pr(s|z) &\propto \sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{V}} n(d,s,w)\Pr(z|d,s,w) \\
\Pr(w|z,s) &\propto \sum_{d\in\mathcal{D}} n(d,s,w)\Pr(z|d,s,w)
\end{aligned}$$

Cluster Analysis

Goal

To assess the impact of SGM-based representation of documents on document clustering

- Documents are represented as pmfs over a topic-feature space
  - identified by a mixture model of the topic distributions for each document
  - lower-dimensional than the corresponding term-feature space
- Information-theoretic notion of distance/similarity between pmfs
- Clusters should contain documents that share the same/similar topic assignment (mixtures) $\Longrightarrow$ possibly overlapping topic-sets

Distance for document pmfs

Hellinger distance

Given a discrete random variable defined on a sample space $X = \{x_1, \ldots, x_R\}, x_r \in \Re, \forall r \in [1..R]$ and two pmfs p, q for that variable

$$HL(p, q) = \sqrt{1 - BC(p, q)}$$

where $BC(p, q) = \sum_{i=1}^{R} \sqrt{p(x_i)\ q(x_i)}$ is the Bhattacharyya coefficient for p and q

- Bhattacharyya coefficient represents the cosine between two vectors for p and q, which are composed by the square root of the probabilities of the mixtures that shape p and q
- It does not require symmetrization, since it is already symmetric (unlike Kullback-Leibler divergence)
- Hellinger distance is a metric

AHC Algorithm for document pmfs

Input: a set of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$ modeled as pmfs, (optionally) a desired number K of clusters

Output: a set of partitions $\mathbf{C}$

$\mathcal{C} \leftarrow \{C_1, \ldots, C_N\}$ such that $C_i = \{d_i\}, \forall i \in [1..N]$
$\mathcal{P}_{C_i} \leftarrow d_i, \forall i \in [1..N]$, as initial cluster prototypes
$\mathbf{C} \leftarrow \{\mathcal{C}\}$
repeat
    let C_i, C_j be the pair of clusters in $\mathcal{C}$ such that $\frac{1}{2}(HL(\mathcal{P}_{C_i \cup C_j}, \mathcal{P}_{C_i}) + HL(\mathcal{P}_{C_i \cup C_j}, \mathcal{P}_{C_j}))$ is minimum
    $C' \leftarrow \{C_i \cup C_j\}$
    $updatePrototype(C')$
    $\mathcal{C} \leftarrow \{C \mid C \in \mathcal{C}, C \neq C_i, C \neq C_j\} \cup \{C'\}$
    $\mathbf{C} \leftarrow \mathbf{C} \cup \{\mathcal{C}\}$
until $|\mathcal{C}| = 1$ (alternatively, if required, $|\mathcal{C}| = K$)

Evaluation framework

- Three multi-topic datasets

dataset	*size (#docs)*	*#words*	*#topic-labels*	*avg #topic-labels per doc*	*#topic-sets*	*avg #docs per topic-set*
IEEE	4,691	129,076	12	4.56	76	61.72
PubMed	3,687	85,771	15	3.20	33	111.73
RCV1	6,588	37,688	23	3.50	49	134.45

- Competing generative models: PLSA, LDA, Ext-PLSA
- Text Segmentation algorithm: **TextTiling**
- Reference partitions for each dataset → **topic-sets** generation
- Assessment criteria: **F-measure**, **Entropy**, **NMI**

Extracting topic-sets

Topic-set (θ): subset of topics in $\mathcal{Z}$ entirely covered by a user-specified portion of $\mathcal{D}$

Overlapping topic-label sets $\Longrightarrow$ multi-topic hard clustering of documents

Given: $\mathcal{D} = \{d_1, \ldots, d_7\}$ and a set of topic-labels $\mathcal{Z} = \{z_1, \ldots, z_5\}$ in $\mathcal{D}$

External document labeling information:

$d_1 \leftarrow \{z_3, z_5\}$ $d_2 \leftarrow \{z_1, z_4\}$ $d_3 \leftarrow \{z_1, z_2, z_5\}$ $d_4 \leftarrow \{z_1, z_4\}$
$d_5 \leftarrow \{z_3, z_5\}$ $d_6 \leftarrow \{z_1, z_4\}$ $d_7 \leftarrow \{z_1, z_2, z_5\}$

3 topic-sets detected:

$\theta_1 = \{z_3, z_5\}$ $\theta_2 = \{z_1, z_4\}$ $\theta_3 = \{z_1, z_2, z_5\}$

$\Longrightarrow$ 3-class partition of $\mathcal{D}$ (i.e., a hard document clustering):

$\{\{d_1, d_5\}, \{d_2, d_4, d_6\}, \{d_3, d_7\}\}$

Extracting topic-sets

Example

Topic-sets generation

$$
\begin{array}{c|ccccc}
 & z_1 & z_2 & z_3 & z_4 & z_5 \\
\hline
d_1 & & & \times & & \times \\
d_2 & \times & & & \times & \\
d_3 & \times & \times & & & \times \\
d_4 & \times & & & \times & \\
d_5 & & & \times & & \times \\
d_6 & \times & & & \times & \\
d_7 & \times & \times & & & \times
\end{array}
\Rightarrow
\begin{array}{lcl}
\theta_1 = \{z_3, z_5\} & \rightarrow & \{d_1, d_5\} \\
\theta_2 = \{z_1, z_4\} & \rightarrow & \{d_2, d_4, d_6\} \\
\theta_3 = \{z_1, z_2, z_5\} & \rightarrow & \{d_3, d_7\}
\end{array}
$$

Extracting segments by TextTiling algorithm

- Two main parameters: block size, #words in a token sequence (window size)
  - interrelated, data-dependent
  - suggested values: 6÷10 for text-unit size, 20 for token-sequence size
- Our setup:
  - Selected values: 3÷15 for text-unit size, 20±10 for token-sequence size
  - Selected configurations: SGM^{min}, SGM^{avg}, and SGM^{max}, for each dataset

ENEA-GRID and CRESCO HPC System

ENEA-GRID

ENEA-GRID provides a unified and homogenous environment for ENEA computational resources located in 6 calculus centers connected via GARR network.
It offers:

- More than 40Tflops of integrated computational power
- Multiplatform systems, i.e., Linux x86_64 ($\sim$5000 cores for CRESCO systems), AIX SP5 ($\sim$256 CPU), and special systems (e.g., GPUs)
- Unified access to remote resources via SSH, NX, and FARO web portal
- A distributed file system (AFS) and a parallel high-performance one (GPFS)
- Cloud services, Virtual Labs, and resource monitoring systems

Experiments have been carried out on **CRESCO HPC System**, located in ENEA Portici Research Center

ENEA
Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile

CRESCO Project

Clustering results

(On IEEE)

segmentation setting	*#segments*	*F*	*E*	*NMI*
SGM^{avg}	155,828	0.64	0.58	0.49
SGM^{min}	89,539	0.59	0.62	0.45
SGM^{max}	179,491	0.58	0.60	0.47

- Higher effectiveness achieved by SGM^{avg}
- More segments would seem to be preferable to smaller segmentations but

More segments $\Rightarrow$ more subtopics discovered $\Rightarrow$ tendency to overfit data

Clustering results (2)

	F				*E*				*NMI*			
	PLSA	Ext-PLSA	LDA	**SGM**	PLSA	Ext-PLSA	LDA	**SGM**	PLSA	Ext-PLSA	LDA	**SGM**
IEEE	0.53	0.56	0.46	0.64	0.70	0.73	0.62	0.58	0.37	0.32	0.44	0.49
PubMed	0.48	0.50	0.43	0.58	0.57	0.54	0.49	0.42	0.50	0.52	0.58	0.64
RCV1	0.49	0.54	0.42	0.56	0.57	0.59	0.51	0.48	0.49	0.46	0.54	0.59
avg score	*0.50*	*0.53*	*0.44*	*0.59*	*0.61*	*0.62*	*0.54*	*0.49*	*0.45*	*0.43*	*0.52*	*0.57*
avg gain	*+0.09*	*+0.06*	*+0.16*	—	*+0.12*	*+0.13*	*+0.05*	—	*+0.12*	*+0.14*	*+0.05*	—

SGM-based clustering always better than all other methods

- *F* improvements from 0.06 (vs. Ext-PLSA) to 0.16 (vs. LDA) — major gains for relatively longer documents (e.g., IEEE and PubMed)
- *E* improvements from 0.05 (vs. LDA) to 0.13 (vs. Ext-PLSA)
- *NMI* improvements from 0.05 (vs. LDA) to 0.14 (vs. Ext-PLSA)

About the competing methods: LDA (resp. Ext-PLSA) better in terms of NMI and Entropy (resp. F-measure)

- hint: LDA clustering solutions tend to be less coarse than those obtained by PLSA and Ext-PLSA

Clustering results (3) - Comparison with VSM methods

	SGM-based clustering			VSM-based clustering		
dataset	*F*	*E*	*NMI*	*F*	*E*	*NMI*
IEEE	0.64	0.58	0.49	0.21	0.84	0.21
PubMed	0.58	0.42	0.64	0.31	0.79	0.28
RCV1	0.56	0.48	0.59	0.39	0.63	0.45
avg score	*0.61*	*0.49*	*0.57*	*0.30*	*0.75*	*0.31*

Baseline method

- CLUTO's Bisecting K-Means performed on *tf.idf*-weighted term-segment matrix
- Upon CLUTO solution, hard-clustering of documents derived (MV basis)

SGM-based clustering always outperformed VSM-based clustering: (on average) +0.31 *F*, +0.26 *E*, and 0.26 *NMI*

- SGM produces hard-clustering that corresponds to a finer mapping docs-to-topic-sets (better for multi-topic docs)
- Baseline solutions likely to be biased by topics frequent in most of the segments within the same doc

Conclusion

- Exploiting a given segmentation of (multi-topic) documents to identify finer-grained topic distributions in the document generative process
- A new model variable is introduced for the within-document segments

SGM Document Clustering performance

Improvements up to $+10\%$ than LDA and PLSA methods in terms of F-measure, Entropy, and NMI

Future work

- Better investigation of how TS impacts on the performance of SGM-based document clustering, on specific domains
- Other evaluations: qualitative (how do final clusters look?), scalability tests

- Comparison with LDA carried out on segments as documents
- Comparison with other approaches that discard any independence assumption between words, e.g., n-gram, HMM-models

Thanks!

Contact

Giovanni Ponti, Ph.D.
ENEA - Portici Research Center

giovanni.ponti@enea.it
http://www.afs.enea.it/gponti

CRESCO Project
http://www.cresco.enea.it