Multi-criteria regularization for Probabilistic Latent Semantic Analysis

Konstantin Vorontsov
CC RAS • Yandex • MIPT • HSE • MSU

Anna Potapenko
MSU • Yandex School of Data Analysis

• International Conference on Computational Linguistics • Dialogue 2014 (June 4–8, Bekasovo)
1 Probabilistic Topic Modeling
- Introduction
- Overview of Topic Models
- Probabilistic Latent Semantic Analysis

2 Additive Regularization for Topic Modeling
- Theory of regularized EM-algorithm
- Regularization for interpretability
- Experiments

3 Discussion
- More regularizers
- ARTM vs. Bayesian approach
- Conclusions
Topic Modeling and related research areas

Google Scholar citation counts

- Matrix Factorization
- NNMF, Nonnegative Matrix Factorization
- Topic Model
- PLSA, Probabilistic Latent Semantic Analysis
- LDA, Latent Dirichlet Allocation
- Text Categorization
- Text Classification
As more information becomes available, it becomes more difficult to find and discover what we need.

We need new tools to help us organize, search, and understand these vast amounts of information.

---

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

- Discover the hidden themes that pervade the collection.
- Annotate the documents according to those themes.
- Use annotations to organize, summarize, search, form predictions.

### Examples of topics

<table>
<thead>
<tr>
<th>human</th>
<th>evolution</th>
<th>disease</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>evolutionary</td>
<td>host</td>
<td>models</td>
</tr>
<tr>
<td>dna</td>
<td>species</td>
<td>bacteria</td>
<td>information</td>
</tr>
<tr>
<td>genetic</td>
<td>organisms</td>
<td>diseases</td>
<td>data</td>
</tr>
<tr>
<td>genes</td>
<td>life</td>
<td>resistance</td>
<td>computers</td>
</tr>
<tr>
<td>sequence</td>
<td>origin</td>
<td>bacterial</td>
<td>system</td>
</tr>
<tr>
<td>gene</td>
<td>biology</td>
<td>new</td>
<td>network</td>
</tr>
<tr>
<td>molecular</td>
<td>groups</td>
<td>strains</td>
<td>systems</td>
</tr>
<tr>
<td>sequencing</td>
<td>phylogenetic</td>
<td>control</td>
<td>model</td>
</tr>
<tr>
<td>map</td>
<td>living</td>
<td>infectious</td>
<td>parallel</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>malaria</td>
<td>methods</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>parasite</td>
<td>networks</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>parasites</td>
<td>software</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>united</td>
<td>new</td>
</tr>
<tr>
<td>sequences</td>
<td>common</td>
<td>tuberculosis</td>
<td>simulations</td>
</tr>
</tbody>
</table>

---

What is “topic”?

- *Topic* is a special terminology of a particular domain area.
- *Topic* is a set of coherent terms (words or phrases) that often occur together in documents.
- *Topic* is a probability distribution over terms: $p(w|t)$ — frequency of word $w$ in topic $t$

Each document consists of terms:

$p(w|d)$ — (known) frequency of term $w$ in document $d$.

Each document consists of topics, i.e. has its own semantic profile:

$p(t|d)$ — (unknown) frequency of topic $t$ in document $d$.

When writing term $w$ in document $d$ author thinks about topic $t$.

*Topic model* tries to uncover latent topics of a text collection.
What is “Probabilistic Topic Model”?

*Topic model* explains terms *w* in documents *d* by topics *t*:

\[ p(w|d) = \sum p(w|t)p(t|d) \]

**Discussion**

Konstantin Vorontsov (voron@yandex-team.ru)

Multi-criteria regularization for PLSA
Goals and applications of Topic Modeling

Goals:
- Uncover a hidden thematic structure of the text collection
- Find a compressed semantic representation of each document

Applications:
- Information retrieval for long-text queries
- Semantic search in large scientific document collections
- Revealing research trends and research fronts
- Expert search
- Categorization, classification, summarization, segmentation of texts, images, video, signals
- News aggregation
- Recommender systems
- etc...
A classical paradigm of search:

Query: set of words
Result: ranked list of documents that contain these words

From searching words to searching senses:

Query: document (or long fragment, or set of documents)
Result:
- the map of the domain area, research front visualization
- ranked list of documents for query topics,
- ranked list of terms to explain each topic,
- ranked list of authors, cites, named entities, etc.,
- internal semantic structure of any document.
Myths about Probabilistic Topic Modeling

- “bag of words” is a necessary assumption
- latent topics often do not make sense
- probabilistic models and linguistic models are incompatible
- topic modeling = LDA
- topic modeling = PLSA or LDA


Topic Modeling Bibliography:

http://mimno.infosci.cornell.edu/topics.html
Combining temporal and \( n \)-gram topic models

\[ \text{TOT - Mexican War} \]

\[ \text{Our Model - Mexican War} \]

<table>
<thead>
<tr>
<th>1. mexico</th>
<th>8. territory</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. texas</td>
<td>9. army</td>
</tr>
<tr>
<td>3. war</td>
<td>10. peace</td>
</tr>
<tr>
<td>4. mexican</td>
<td>11. act</td>
</tr>
<tr>
<td>5. united</td>
<td>12. policy</td>
</tr>
<tr>
<td>6. country</td>
<td>13. foreign</td>
</tr>
<tr>
<td>7. government</td>
<td>14. citizens</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1. east bank</th>
<th>8. military</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. american coins</td>
<td>9. general herrer</td>
</tr>
<tr>
<td>3. mexican flag</td>
<td>10. foreign coin</td>
</tr>
<tr>
<td>4. separate independent</td>
<td>11. military usurper</td>
</tr>
<tr>
<td>5. american commonwealth</td>
<td>12. mexican treasury</td>
</tr>
<tr>
<td>6. mexican population</td>
<td>13. invaded texas</td>
</tr>
<tr>
<td>7. texan troops</td>
<td>14. veteran troops</td>
</tr>
</tbody>
</table>

Combining temporal and \textit{n}-gram topic models

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{chart.png}
\end{figure}

\begin{tabular}{|c|c|}
\hline
1. government & 8. spanish \\
2. cuba & 9. island \\
3. islands & 10. act \\
4. international & 11. commission \\
5. powers & 12. officers \\
6. gold & 13. spain \\
7. action & 14. rico \\
\hline
\end{tabular}

\begin{tabular}{|c|c|}
\hline
1. panama canal & 8. united states senate \\
2. isthmian canal & 9. french canal company \\
3. isthmus panama & 10. caribbean sea \\
4. republic panama & 11. panama canal bonds \\
5. united states government & 12. panama \\
6. united states & 13. american control \\
7. state panama & 14. canal \\
\hline
\end{tabular}

Information about citations or links between documents helps to build more accurate topic model

Topic model helps to reveal most influential cites

**Goal:** Automatic Hierarchical Text Categorization

**Problem:** How to define a relation “topic $t \rightarrow$ subtopic $s$”?

**Intuition:** Distribution $p(w|s)$ is nested into $p(w|t)$

Nevertheless...

- “Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue.”

- “The evaluation of topic models is also an open issue.”

---

**Bottom-up strategy for building topic hierarchy**

Idea: to use topics as “words” of the upper level

---

Multilingual Topic Models

Jason Chuang, Christopher D. Manning, Jeffrey Heer.
**Topic Modeling mainsteam**

- LDA — Latent Dirichlet Allocation
- Maths: Bayesian Inference, Graphical Models:

David Blei. Probabilistic topic models

**Topic Modeling Bibliography:**
http://mimno.infosci.cornell.edu/topics.html
Given a document collection:

$n_{dw}$ — how many times term $w$ appears in document $d$

Find topic model $p(w|d) = \sum_t \phi_{wt} \theta_{td}$ with parameters $\phi$, $\theta$:

$\phi_{wt} = p(w|t)$ — probabilities of terms $w$ for each topic $t$

$\theta_{td} = p(t|d)$ — probabilities of topics $t$ for each document $d$

The problem of log-likelihood maximization:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} \rightarrow \max_{\phi, \theta},$$

under non-negativeness and normalization constrains

$\phi_{wt} \geq 0; \quad \sum_w \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1$
EM-algorithm

EM-algorithm alternates E-step and M-step until convergence. E-step gives latent topic probabilities from Bayes’ rule:

\[ p(t|d, w) = \frac{\phi_{wt} \theta_{td}}{\sum_s \phi_{ws} \theta_{sd}}; \]

\[ n_{dwt} = n_{dw} p(t|d, w) \] — the number of triples \((d, w, t)\) in \(D\).

M-step gives frequency estimates of conditional probabilities:

\[ \phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, \quad \theta_{td} = \frac{n_{td}}{n_d} \equiv \frac{\sum_w n_{dwt}}{\sum_{w,t} n_{dwt}}, \]

Short notation via proportionality sign \(\propto\):

\[ \phi_{wt} \propto n_{wt}; \quad \theta_{td} \propto n_{td}; \]
The efficient implementation of EM-algorithm

The idea is to incorporate E-step into M-step. No 3D-arrays!

**Input:** collection $D$, num. of topics $|T|$, num. of iterations $i_{\text{max}}$;
**Output:** distributions $\phi$, $\theta$;

1. initialize $\phi_{wt}$, $\theta_{td}$ for all $d \in D$, $w \in W$, $t \in T$;
2. for all iterations $i = 1, \ldots, i_{\text{max}}$
   3. $n_{wt}, n_{td}, n_{t}, n_{d} := 0$ for all $d \in D$, $w \in W$, $t \in T$;
   4. for all documents $d \in D$ and terms $w \in d$
      5. $p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s} \phi_{ws}\theta_{sd}}$ for all $t \in T$;
      6. $n_{wt}, n_{td}, n_{t}, n_{d} += n_{dw}p_{tdw}$ for all $t \in T$;
   7. $\phi_{wt} := n_{wt}/n_{t}$ for all $w \in W$, $t \in T$;
   8. $\theta_{td} := n_{td}/n_{d}$ for all $d \in D$, $t \in T$;

Usually $i_{\text{max}} = 20..50$ iterations are sufficient. Time is $O(n|T|i_{\text{max}})$.
The ill-posed problem:
likelihood maximization has infinitely many solutions.

Regularization of the ill-posed problem:

Let us maximize likelihood with regularizers $R_i(\phi, \theta)$, $i = 1, \ldots, n$

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^n \tau_i R_i(\phi, \theta) \rightarrow \max_{\phi, \theta}$$

under non-negativeness and normalization restrictions

$$\phi_{wt} \geq 0; \quad \sum_w \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1$$

where $\tau_i > 0$ are regularization coefficients.
ARTM: EM-algorithm with regularized M-step

M-step for PLSA:
\[ \phi_{wt} \propto n_{wt}; \quad \theta_{td} \propto n_{td}; \]

M-step for regularized PLSA:
\[ \phi_{wt} \propto \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \quad \theta_{td} \propto \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+; \]

where \((x)_+ = \max(x, 0)\) is a positive cutoff.

Classical Topic Models are particular cases of ARTM:
PLSA: \( R(\phi, \theta) = 0 \)
LDA: \( R(\phi, \theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td} \)

ARTM is a simple and powerful tool for capturing additional criteria, including linguistical requirements and resources.
**Assumptions: what topics would be well-interpretable?**

*Specific topics* $S$ contain domain-specific terms $p(w|t)$ are sparse and different (weakly correlated)

*Background topics* $B$ contain common lexis words $p(w|t)$ are not sparse

$\phi_{wt}$ terms $\times$ topics $\theta_{td}$ topics $\times$ documents
The non-sparsity assumption for background topics $t \in B$: $
abla_{wt}$ are similar to a given distribution $\beta_w$; $\theta_{td}$ are similar to a given distribution $\alpha_t$.

$$\sum_{t \in B} KL_w(\beta_w \| \phi_{wt}) \to \min_{\phi}; \quad \sum_{d \in D} KL_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

We minimize the sum of these KL-divergences to get a regularizer:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \to \max.$$

The regularized M-step applied for all $t \in B$ coincides with LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \quad \theta_{td} \propto n_{td} + \alpha_0 \alpha_t,$$

which is new non-Bayesian interpretation of LDA [Blei 2003].
The sparsity assumption for domain-specific topics $t \in S$: distributions $\phi_{wt}$, $\theta_{td}$ contain many zero probabilities.

We maximize the sum of KL-divergences $KL(\beta \| \phi_t)$ and $KL(\alpha \| \theta_d)$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \rightarrow \max.$$ 

The regularized M-step gives “anti-LDA”, for all $t \in S$:

$$\phi_{wt} \propto (n_{wt} - \beta_0 \beta_w)_+, \quad \theta_{td} \propto (n_{td} - \alpha_0 \alpha_t)_+. $$

The dissimilarity assumption: domain-specific topics \( t \in S \) must be as distant as possible.

We maximize covariances between column vectors \( \phi_t \):

\[
R(\Phi) = -\frac{\tau}{2} \sum_{t,s \in S} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \text{max}.
\]

The regularized M-step makes columns of \( \Phi \) more distant:

\[
\phi_{wt} \propto \left( n_{wt} - \tau \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right)_+.
\]

---

Assumption: infrequent topics are not well-interpretable.

We maximize KL-divergence \( KL(\frac{1}{|T|} \| p(t)) \) to make distribution over topics \( p(t) = \sum_d p(d)\theta_{td} \) sparse:

\[
R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.
\]

The regularized M-step formula results in \( \Theta \) rows sparsing:

\[
\theta_{td} \propto \left( n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.
\]

Effect:
if \( n_t \) is small then all values in the \( t \)-th row may turn into zeros.
Assumption: if topic is well-interpretable then its top words are coherent i.e. frequently appear nearby in the documents.

\[ C_{uw} = \hat{p}(w|u) = \frac{N_{uw}}{N_u} \quad \text{— coherence of a word pair } u, w \in W, \]

\[ N_u, N_{uw} \text{ are document frequency of word } w \text{ and word pair } u, w. \]

Bring together \( \phi_{wt} \) and its coherent words estimate \( \hat{p}(w|t) \):

\[ \hat{p}(w|t) = \sum_u \hat{p}(w|u)p(u|t) = \frac{1}{n_t} \sum_u C_{uw} n_{ut}; \]

\[ R(\Phi, \Theta) = \tau \sum_{t \in T} \sum_{w \in W} \hat{p}(w|t) \ln \phi_{wt} \to \max. \]

The regularized M-step gives a kind of smoothing:

\[ \phi_{wt} \propto n_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} n_{ut}. \]

Assumption: experts have provided us with topic labeling data:
- each document $d \in D_0 \subseteq D$ belongs to a subset of topics $T_d \subseteq T$;
- each topic $t \in T_0 \subseteq T$ contains a subset of words $W_t \subseteq W$.

$\phi^0_{wt} \sim \text{uniform distribution over subset of terms } W_t$
$\theta^0_{td} \sim \text{uniform distribution over subset of topics } T_d$

We minimize the sum of KL-divergences $\text{KL}(\phi^0_t \parallel \phi_t)$ and $\text{KL}(\theta^0_t \parallel \theta_t)$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \rightarrow \max.$$  

The regularized M-step results in LDA-like smoothing:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi^0_{wt} \quad \theta_{td} \propto n_{td} + \alpha_0 \theta^0_{td}$$

The goal of the experiment:
Can we improve interpretability without loss of the likelihood?

The set of regularizers:
- smoothing background topics
- sparsing domain-specific topics
- decorrelation of domain-specific topics
- topic selection

Dataset: NIPS (Neural Information Processing System)
- 1566 papers from NIPS conference;
- collection length $\approx 2.3 \cdot 10^6$,
- vocabulary size $\approx 1.3 \cdot 10^4$. 
Multi-criteria optimization requires multiple quality measures.

- Hold-out *perplexity*: \( P = \exp\left(- \frac{1}{n} \mathcal{L} \right) \)
- *Sparsity* — the number of zero elements \( \phi_{wt} \) and \( \theta_{td} \)
- Interpretability measures for each topic \( t \):
  - topic *coherence* [Newman, 2010]
  - topic *kernel size*: \( |W_t| \), kernel \( W_t = \{ w : p(t|w) > 0.25 \} \)
  - topic *purity*: \( \sum_{w \in W_t} p(w|t) \)
  - topic *contrast*: \( \frac{1}{|W_t|} \sum_{w \in W_t} p(t|w) \)

- Model degeneracy:
  - number of non-zero topics: \( |T| \)
  - the fraction of background words: \( \frac{1}{n} \sum_{d,w} \sum_{t \in B} p(t|d,w) \)
ARTM — Additive Regularization of Topic Model

M-step formula for combined regularization:

\[
\phi_{wt} \propto \left( n_{wt} + \tau_1 \beta_w[t \in B] - \tau_2 \beta_w[t \in S] - \tau_3 \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right) +
\]

\[
\theta_{td} \propto \left( n_{td} + \tau_4 \alpha_t[t \in B] - \tau_5 \alpha_t[t \in S] - \tau_6 \frac{n_d}{n_t} \theta_{td} \right) +
\]

A new problem arises: how to choose the regularization path \( \tau = (\tau_1, \ldots, \tau_6) \) as a function of the iteration number?
Quality measures as functions of the iteration number
(grey lines — PLSA, black lines — ARTM)
Quality measures as functions of the iteration number (grey lines — PLSA, black lines — ARTM)
### Topics (kernel words are bold)

<table>
<thead>
<tr>
<th>PLSA:</th>
<th>ARTM:</th>
<th>background:</th>
<th>background:</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>margin</td>
<td>training</td>
<td>model</td>
</tr>
<tr>
<td>algorithm</td>
<td>Kalman</td>
<td>set</td>
<td>data</td>
</tr>
<tr>
<td>set</td>
<td>boosting</td>
<td>class</td>
<td>models</td>
</tr>
<tr>
<td>training</td>
<td>AdaBoost</td>
<td>learning</td>
<td>parameters</td>
</tr>
<tr>
<td>networks</td>
<td>cancer</td>
<td>classification</td>
<td>noise</td>
</tr>
<tr>
<td>Kalman</td>
<td>bagging</td>
<td>algorithm</td>
<td>neural</td>
</tr>
<tr>
<td>neural</td>
<td>committee</td>
<td>data</td>
<td>mixture</td>
</tr>
<tr>
<td>error</td>
<td>datasets</td>
<td>number</td>
<td>prediction</td>
</tr>
<tr>
<td>performance</td>
<td>combining</td>
<td>classifier</td>
<td>set</td>
</tr>
<tr>
<td>margin</td>
<td>Breiman</td>
<td>performance</td>
<td>gaussian</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>resampling</td>
<td>examples</td>
<td>likelihood</td>
</tr>
<tr>
<td>models</td>
<td>breast</td>
<td>decision</td>
<td>networks</td>
</tr>
<tr>
<td>algorithms</td>
<td>UCI</td>
<td>results</td>
<td>test</td>
</tr>
<tr>
<td>committee</td>
<td>EKF</td>
<td>probability</td>
<td>figure</td>
</tr>
<tr>
<td>methods</td>
<td>Freund</td>
<td>problem</td>
<td>training</td>
</tr>
<tr>
<td>cancer</td>
<td>voting</td>
<td>tree</td>
<td>performance</td>
</tr>
<tr>
<td>test</td>
<td>margins</td>
<td>error</td>
<td>network</td>
</tr>
<tr>
<td>linear</td>
<td>sequential</td>
<td>based</td>
<td>number</td>
</tr>
<tr>
<td>datasets</td>
<td>misclassification</td>
<td>classes</td>
<td>input</td>
</tr>
<tr>
<td>figure</td>
<td>Schapire</td>
<td>algorithms</td>
<td>results</td>
</tr>
</tbody>
</table>

---

Konstantin Vorontsov (voron@yandex-team.ru)
## Topics (kernel words are bold)

<table>
<thead>
<tr>
<th>PLSA:</th>
<th>ARTM:</th>
</tr>
</thead>
<tbody>
<tr>
<td>music</td>
<td>estimator</td>
</tr>
<tr>
<td>rules</td>
<td>music</td>
</tr>
<tr>
<td>note</td>
<td>musical</td>
</tr>
<tr>
<td>representation</td>
<td>notes</td>
</tr>
<tr>
<td>neural</td>
<td>Mozer</td>
</tr>
<tr>
<td>events</td>
<td>melody</td>
</tr>
<tr>
<td>net</td>
<td>composition</td>
</tr>
<tr>
<td>set</td>
<td>Bach</td>
</tr>
<tr>
<td>time</td>
<td>chorales</td>
</tr>
<tr>
<td>musical</td>
<td>melodic</td>
</tr>
<tr>
<td>figure</td>
<td>jackknife</td>
</tr>
<tr>
<td>network</td>
<td>cooperative</td>
</tr>
<tr>
<td>notes</td>
<td>subnet</td>
</tr>
<tr>
<td>input</td>
<td>GEM</td>
</tr>
<tr>
<td>melody</td>
<td>melodies</td>
</tr>
<tr>
<td>structure</td>
<td>ICL</td>
</tr>
<tr>
<td>harmony</td>
<td>tonal</td>
</tr>
<tr>
<td>tau</td>
<td>accent</td>
</tr>
<tr>
<td>pitch</td>
<td>augmented</td>
</tr>
<tr>
<td>temporal</td>
<td>piece</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PLSA:</th>
<th>ARTM:</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>compression</td>
</tr>
<tr>
<td>matrix</td>
<td>quantization</td>
</tr>
<tr>
<td>algorithm</td>
<td>singular</td>
</tr>
<tr>
<td>vector</td>
<td>pyramid</td>
</tr>
<tr>
<td>network</td>
<td>Sanger</td>
</tr>
<tr>
<td>output</td>
<td>decomposition</td>
</tr>
<tr>
<td>vectors</td>
<td>generalized</td>
</tr>
<tr>
<td>figure</td>
<td>SVD</td>
</tr>
<tr>
<td>input</td>
<td>SONN</td>
</tr>
<tr>
<td>singular</td>
<td>MDL</td>
</tr>
<tr>
<td>data</td>
<td>primitives</td>
</tr>
<tr>
<td>mapping</td>
<td>Tenorio</td>
</tr>
<tr>
<td>image</td>
<td>plasma</td>
</tr>
<tr>
<td>dimensional</td>
<td>image</td>
</tr>
<tr>
<td>nonlinear</td>
<td>dimensional</td>
</tr>
<tr>
<td>algorithms</td>
<td>matrices</td>
</tr>
<tr>
<td>inverse</td>
<td>compressed</td>
</tr>
<tr>
<td>error</td>
<td>reconstruction</td>
</tr>
<tr>
<td>optimal</td>
<td>book</td>
</tr>
<tr>
<td>compression</td>
<td>memorized</td>
</tr>
<tr>
<td>optimal compression</td>
<td>Laplace</td>
</tr>
<tr>
<td></td>
<td>eigen</td>
</tr>
</tbody>
</table>
### Topics (kernel words are bold)

<table>
<thead>
<tr>
<th>PLSA:</th>
<th>ARTM:</th>
</tr>
</thead>
<tbody>
<tr>
<td>face</td>
<td>face</td>
</tr>
<tr>
<td>images</td>
<td>faces</td>
</tr>
<tr>
<td>faces</td>
<td>facial</td>
</tr>
<tr>
<td>recognition</td>
<td>Cottrell</td>
</tr>
<tr>
<td>set</td>
<td>Pentland</td>
</tr>
<tr>
<td>image</td>
<td>gesture</td>
</tr>
<tr>
<td>based</td>
<td>lane</td>
</tr>
<tr>
<td>HME</td>
<td>emotion</td>
</tr>
<tr>
<td>facial</td>
<td>person</td>
</tr>
<tr>
<td>representation</td>
<td>steering</td>
</tr>
<tr>
<td>view</td>
<td>appearance</td>
</tr>
<tr>
<td>figure</td>
<td>BALUJA</td>
</tr>
<tr>
<td>model</td>
<td>setpoint</td>
</tr>
<tr>
<td>experts</td>
<td>camera</td>
</tr>
<tr>
<td>network</td>
<td>tracking</td>
</tr>
<tr>
<td>human</td>
<td>pose</td>
</tr>
<tr>
<td>expert</td>
<td>Pomerleau</td>
</tr>
<tr>
<td>space</td>
<td>mouth</td>
</tr>
<tr>
<td>examples</td>
<td>Darrell</td>
</tr>
<tr>
<td>system</td>
<td>lighting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PLSA:</th>
<th>ARTM:</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>MLP</td>
</tr>
<tr>
<td>set</td>
<td>query</td>
</tr>
<tr>
<td>queries</td>
<td>set</td>
</tr>
<tr>
<td>data</td>
<td>queries</td>
</tr>
<tr>
<td>algorithm</td>
<td>queries</td>
</tr>
<tr>
<td>learning</td>
<td>documents</td>
</tr>
<tr>
<td>documents</td>
<td>relevant</td>
</tr>
<tr>
<td>number</td>
<td>retrieval</td>
</tr>
<tr>
<td>performance</td>
<td>relevant</td>
</tr>
<tr>
<td>words</td>
<td>relevant</td>
</tr>
<tr>
<td>relevant</td>
<td>ranking</td>
</tr>
<tr>
<td>classification</td>
<td>ordering</td>
</tr>
<tr>
<td>algorithms</td>
<td>recursive</td>
</tr>
<tr>
<td></td>
<td>text</td>
</tr>
<tr>
<td></td>
<td>axis</td>
</tr>
</tbody>
</table>
ARTM provides a multi-objective model improvement:

- **sparsity** augments from 0 to 95%–98%
- **coherence** augments from 0.1 to 0.3
- **purity** augments from 0.15 to 0.8
- **contrast** augments from 0.4 to 0.6
- **kernel size** augments from 0 to 150 terms
- almost without any loss of the **perplexity**

**Recommendations for choosing regularization path:**

- **turn on sparsing** gradually after first 10-20 iterations
- **turn on topic selection** after turning on sparsing
- **turn off sparsing** as soon as kernel size begins to decrease
- **turn on background smoothing** from the beginning
- **turn on decorrelation** as much as possible from the beginning
- make topic **selection** and **decorrelation** at different iterations
Experiment 2: PressRelease collection. Temporal model

The goal of the experiment:
Can ARTM help to build temporal topic model?

The set of regularizers:
- smoothing background topics
- sparsifying domain-specific topics
- sparsifying $p(t|\text{time})$ distributions
- penalizing noisy variations of $p(t|\text{time})$

Dataset:
- 20,000 press releases of 4 countries, 2001–2013;
Experimental work *Nikita Doykov*, MSU, 2014
ARTM for temporal Topic Modeling

Experimental work of Nikita Doykov, MSU, 2014
Variety of regularizers for ARTM

Understood and implemented:
1. smoothing
2. sparsing
3. topic decorrelation
4. topic selection

Understood but not implemented yet:
5. semi-supervised learning
6. coherence maximization
7. using links or cites between documents
8. using document categories or classes
9. using time-stamped data
10. ...
ARTM vs. Bayesian approach

David Blei. Probabilistic topic models

Topic Modeling Bibliography:
http://mimno.infosci.cornell.edu/topics.html
ARTM vs. Bayesian approach

**Bayesian Inference for Probabilistic Topic Modeling**

1. Fully probabilistic generative model of data
2. Dirichlet distribution plays a central role in the theory
3. Complicated maths for combined and multi-objective models
4. High barrier to entry into PTMs research field

**Additive Regularization for Topic Modeling**

1. Semi-probabilistic approach
2. No Dirichlet prior, no integration, no graphical models
3. Simple maths for combined and multi-objective models
4. Very short way from an idea to the algorithm
Further research work

- More linguistically motivated regularization
- Regularization for $p(t|d, w)$ beyond “bag-of-words” assumption
- ARTM + Lexical Chains
- ARTM for $n$-gram models and Term Extraction
- ARTM for multi-lingual and cross-lingual search
- ARTM for building topic hierarchies
- BigARTM — starting open source project for Large-Scale Parallel Distributed Multi-Objective Topic Modeling
- Convergence of the regularized EM-algorithm
- Choosing a regularization path
- Applications


User:Vokov

Вероятностные тематические модели
(курс лекций, К. В. Воронцов)

Тематическое моделирование