Multi-criteria regularization for Probabilistic Latent Semantic Analysis

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

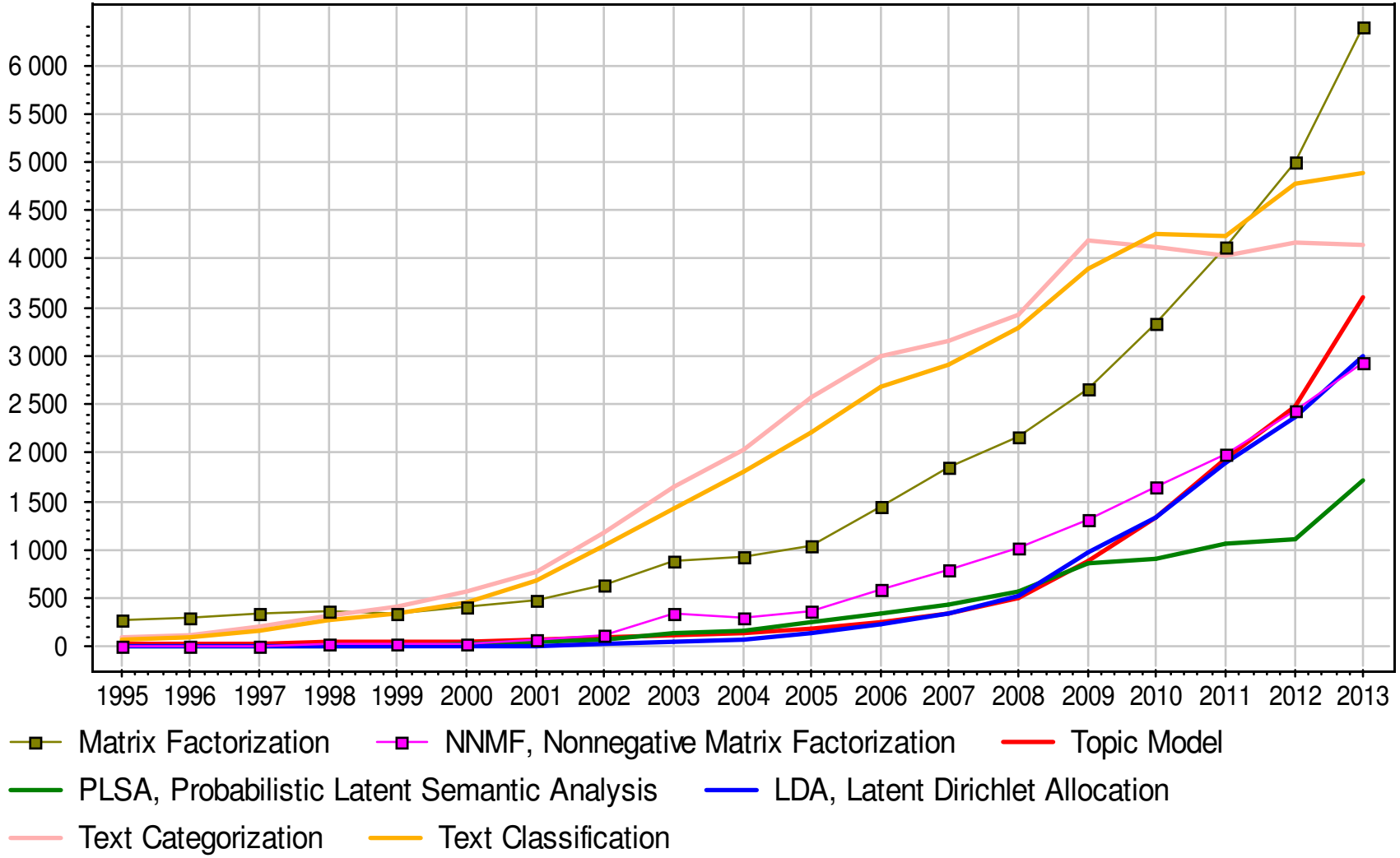
Anna Potapenko
MSU • Yandex School of Data Analysis

• International Conference on Computational Linguistics •
Dialogue 2014 (June 4–8, Bekasovo)

Contents

Topic Modeling and related research areas

Google Scholar citation counts



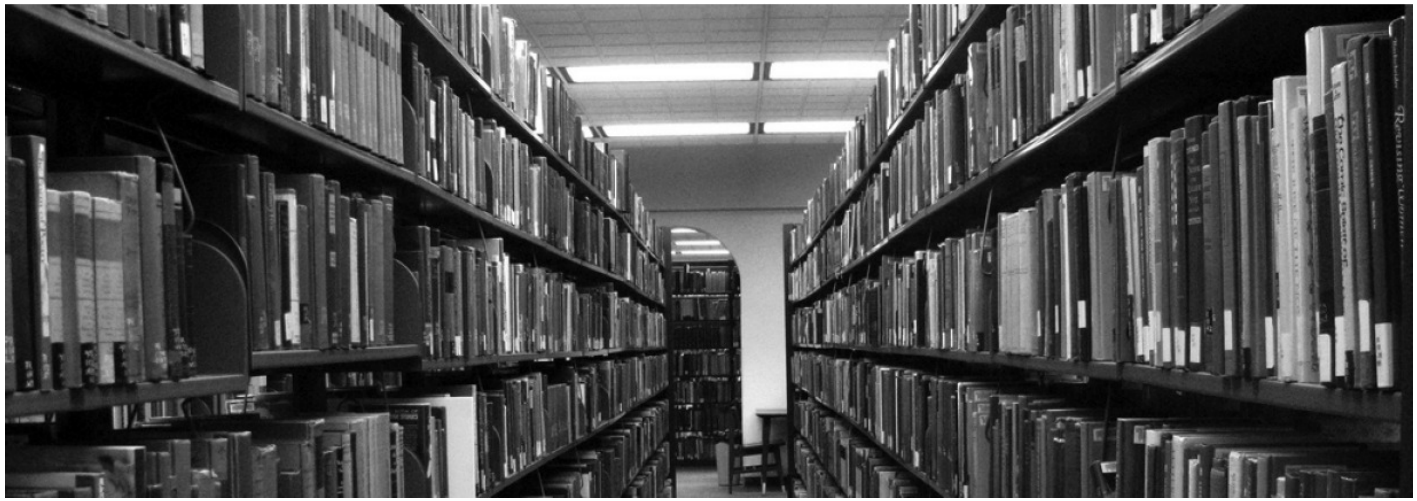
From Information Retrieval to Information Systematization

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.

David Blei. Princeton University. http://www.cs.princeton.edu/~blei

From Information Retrieval to Information Systematization



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.

David Blei. Princeton University. http://www.cs.princeton.edu/~blei

Examples of topics

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

David Blei. Princeton University. http://www.cs.princeton.edu/~blei

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.

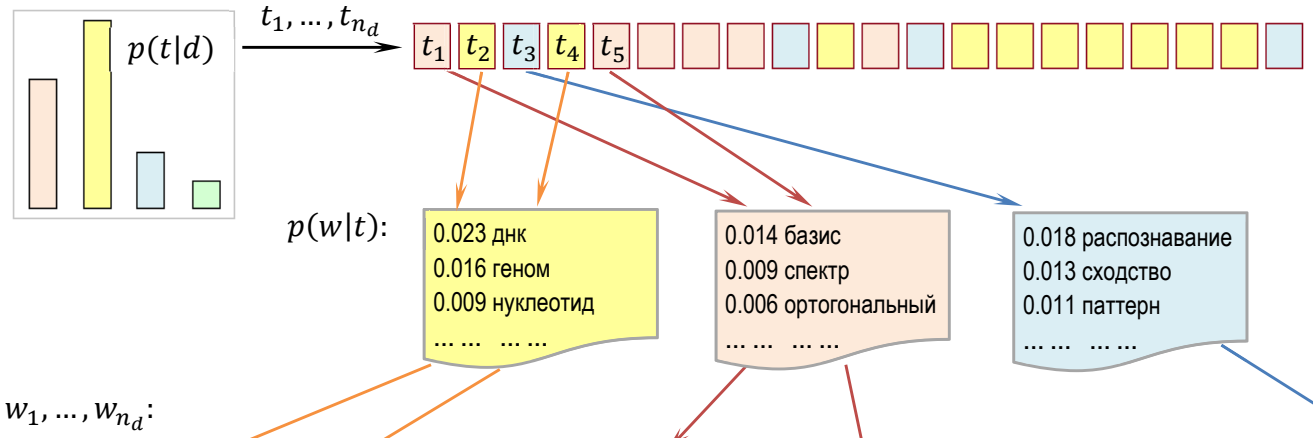
Each document consists of topics, i.e. has its own semantic profile:
$p(t|d)$ — (unknown) frequency of topic t in document d.

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:

$$p(w|d) = \sum_t p(w|t)p(t|d)$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

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...

Topic Modeling for Scientific Information Retrieval

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

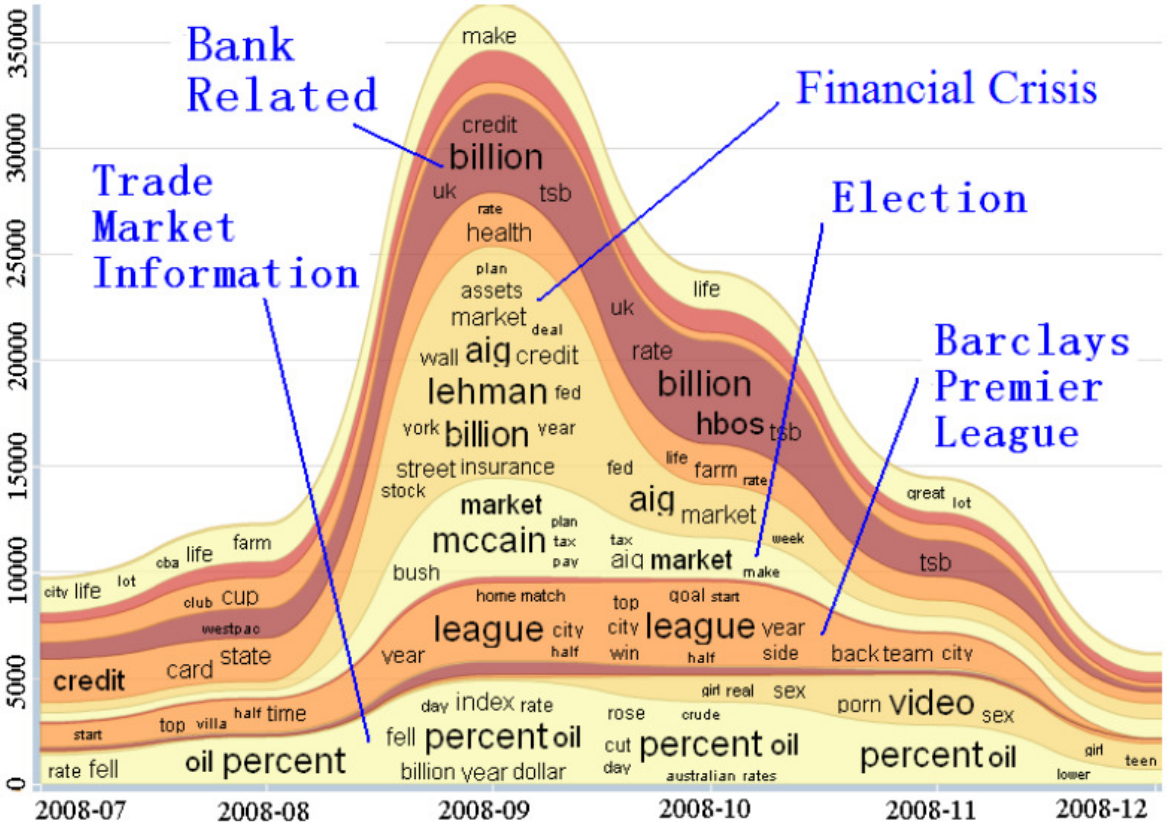
- "bag of words" is a necessary assumption
- latent topics often do not make sense
- probabilistic models and linguistic models are incompatible
- topic modeling = LDA
- topic modeling = PLSA or LDA

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.
Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(русский перевод на www.MachineLearning.ru)

Topic Modeling Bibliography:

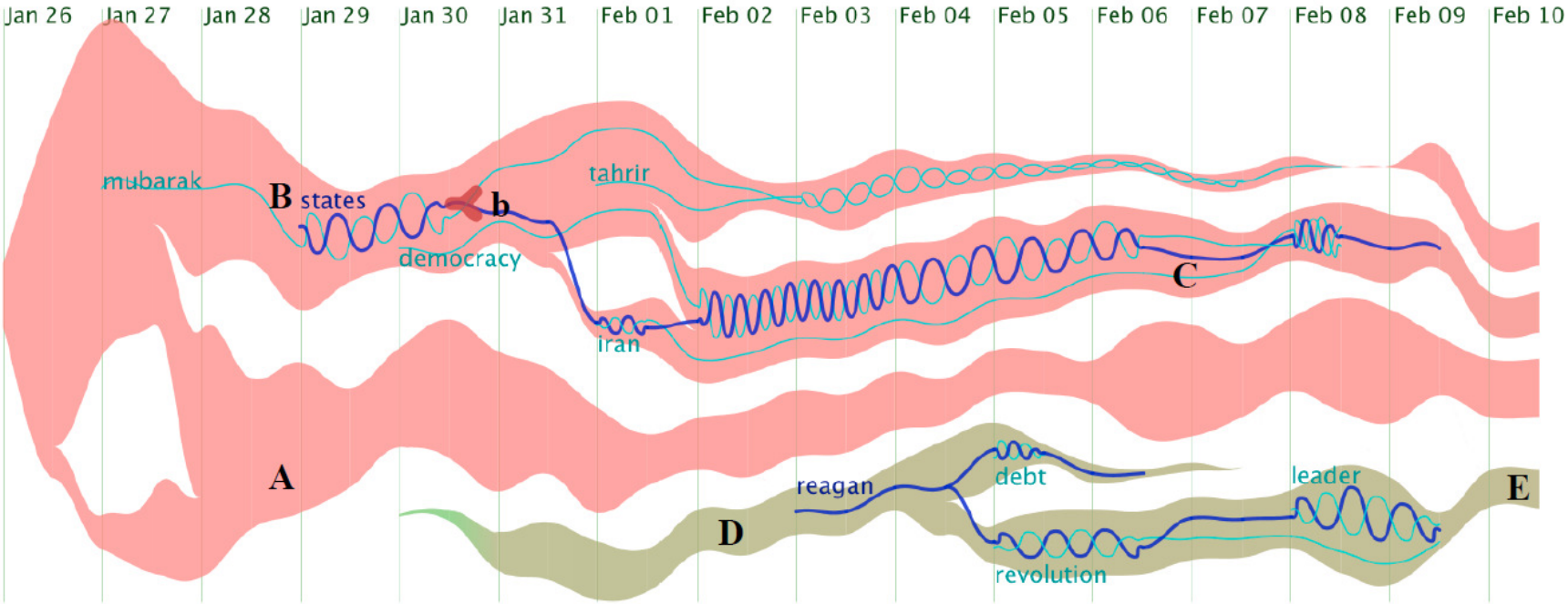
http://mimno.infosci.cornell.edu/topics.html

Temporal Topic Model



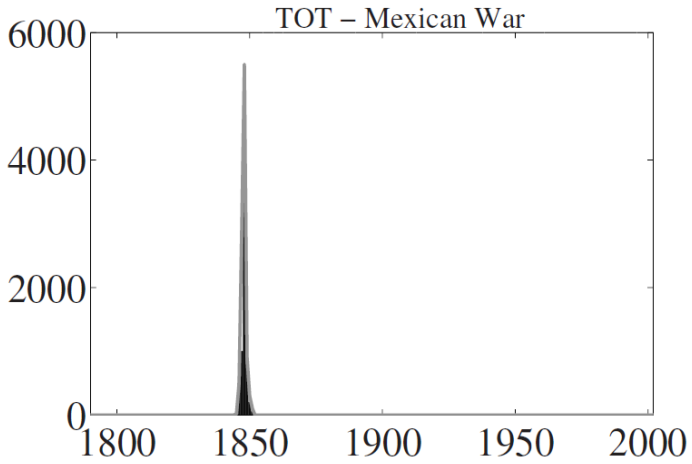
Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25–28, 2010.

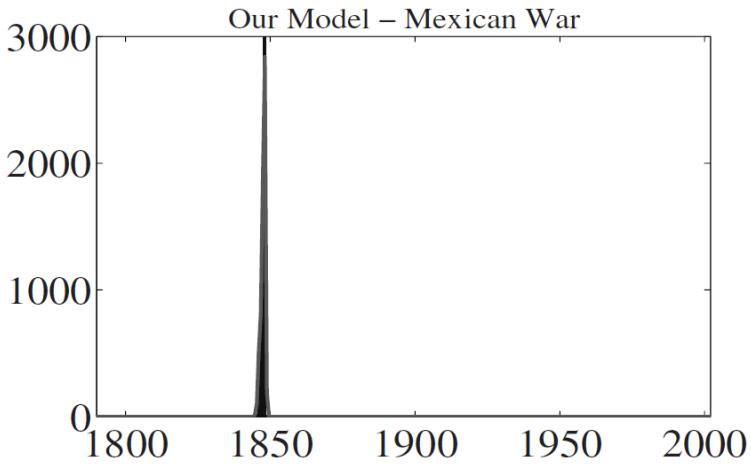
Temporal Topic Model with evolving topics



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions On Visualization And Computer Graphics, Vol. 17, No. 12, December 2011.

Combining temporal and *n*-gram topic models



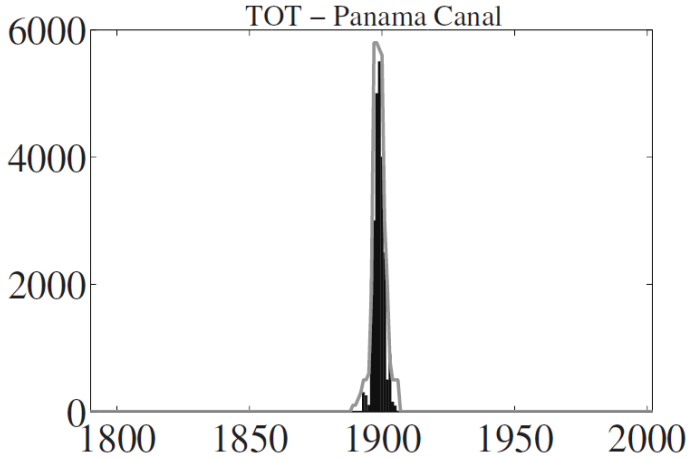


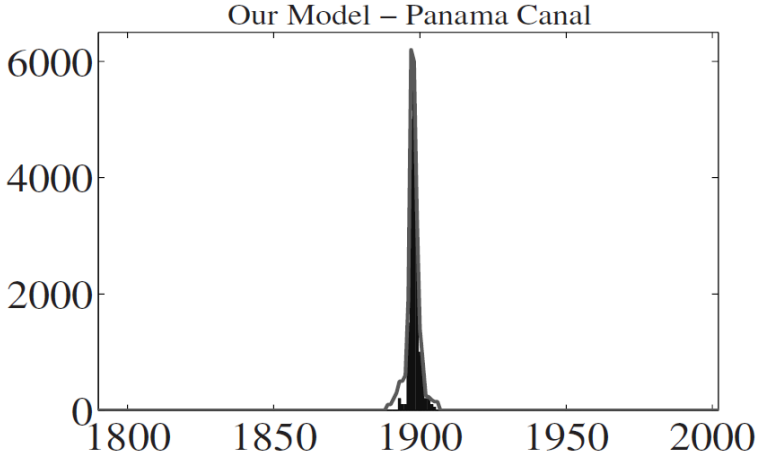
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Combining temporal and *n*-gram topic models





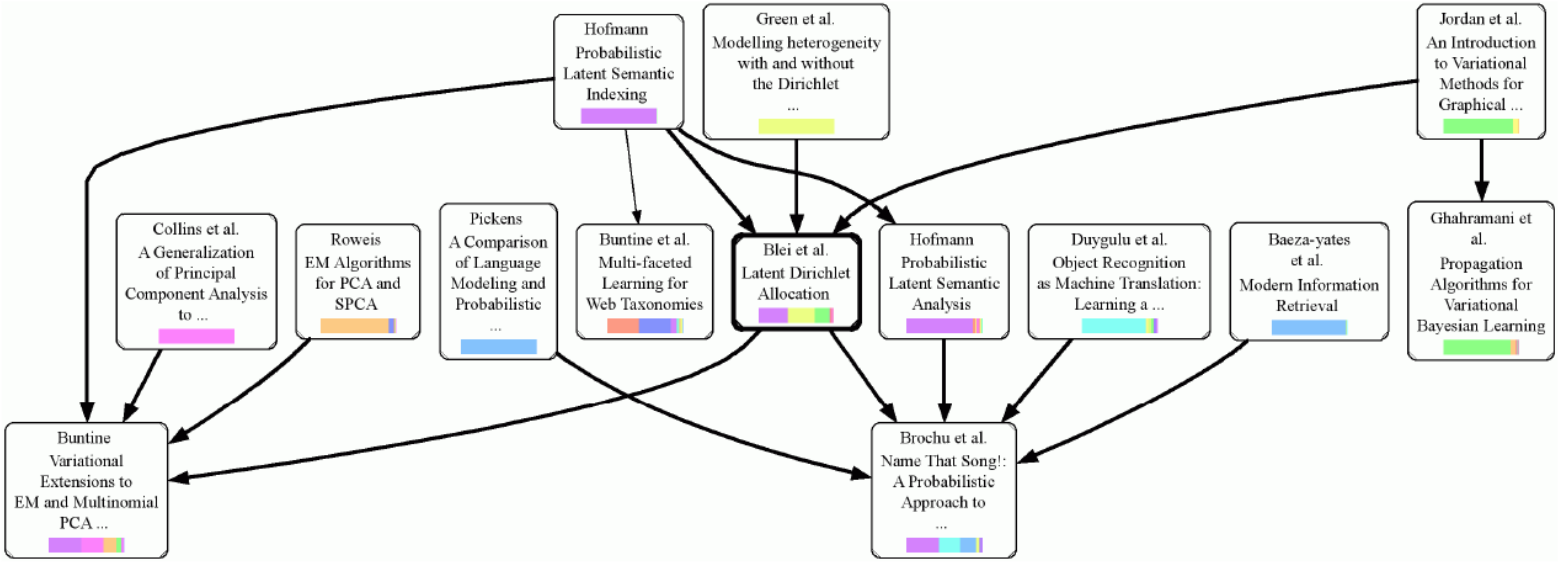
1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

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

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

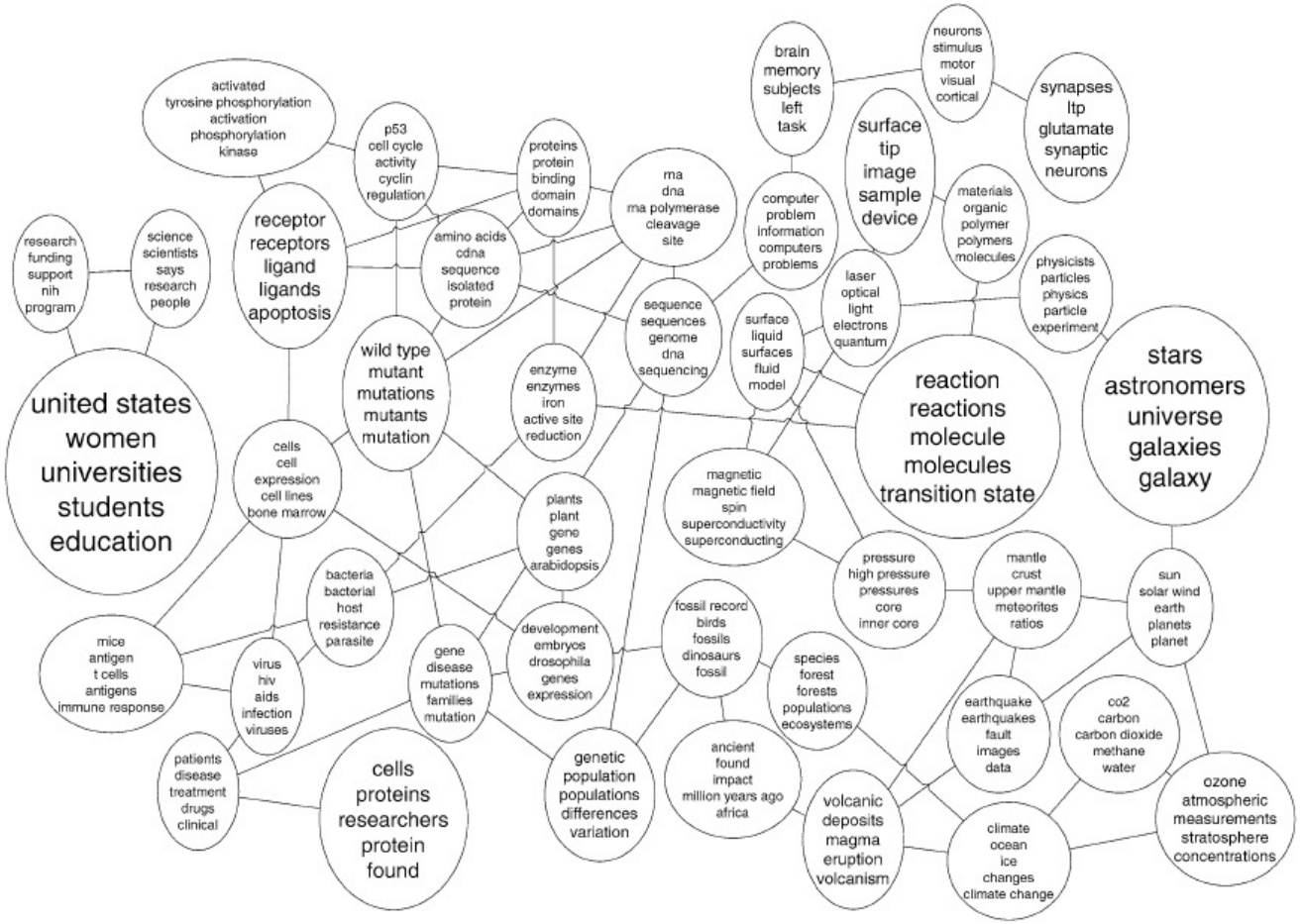
Citation and Links Topic Model

- Information about citations or links between documents helps to build more accurate topic model
- Topic model helps to reveal most influential cites



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences // ICML-2007, Pp. 233–240.

Correlated Topic Model



D. Blei, J. Lafferty. A correlated topic model of *Science* // Annals of Applied Statistics, 2007. Vol. 1, Pp. 17-35.

Hierarchical Topic Model

Goal: Automatic Hierarchical Text Categorization
Problem: How to define a relation "topic $t \to$ subtopic s"?
Intuition: Distribution $p(w|s)$ is nested into $p(w|t)$

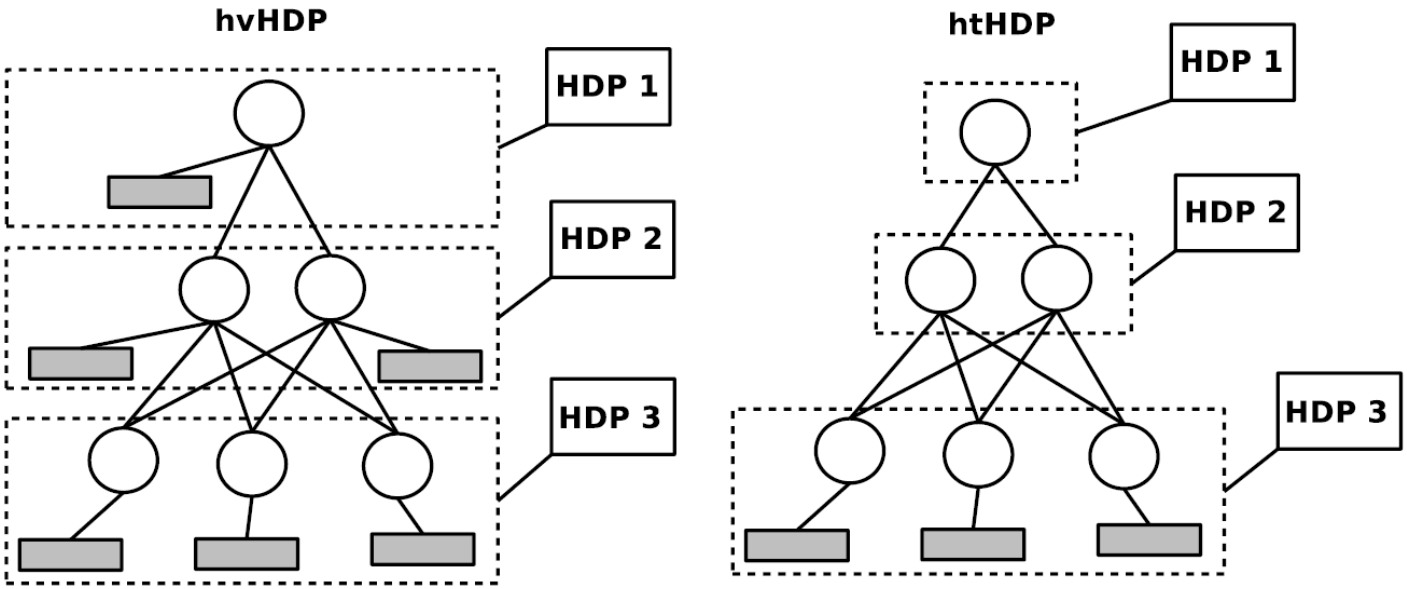
Nevertheless...

- "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."

E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749–2775.

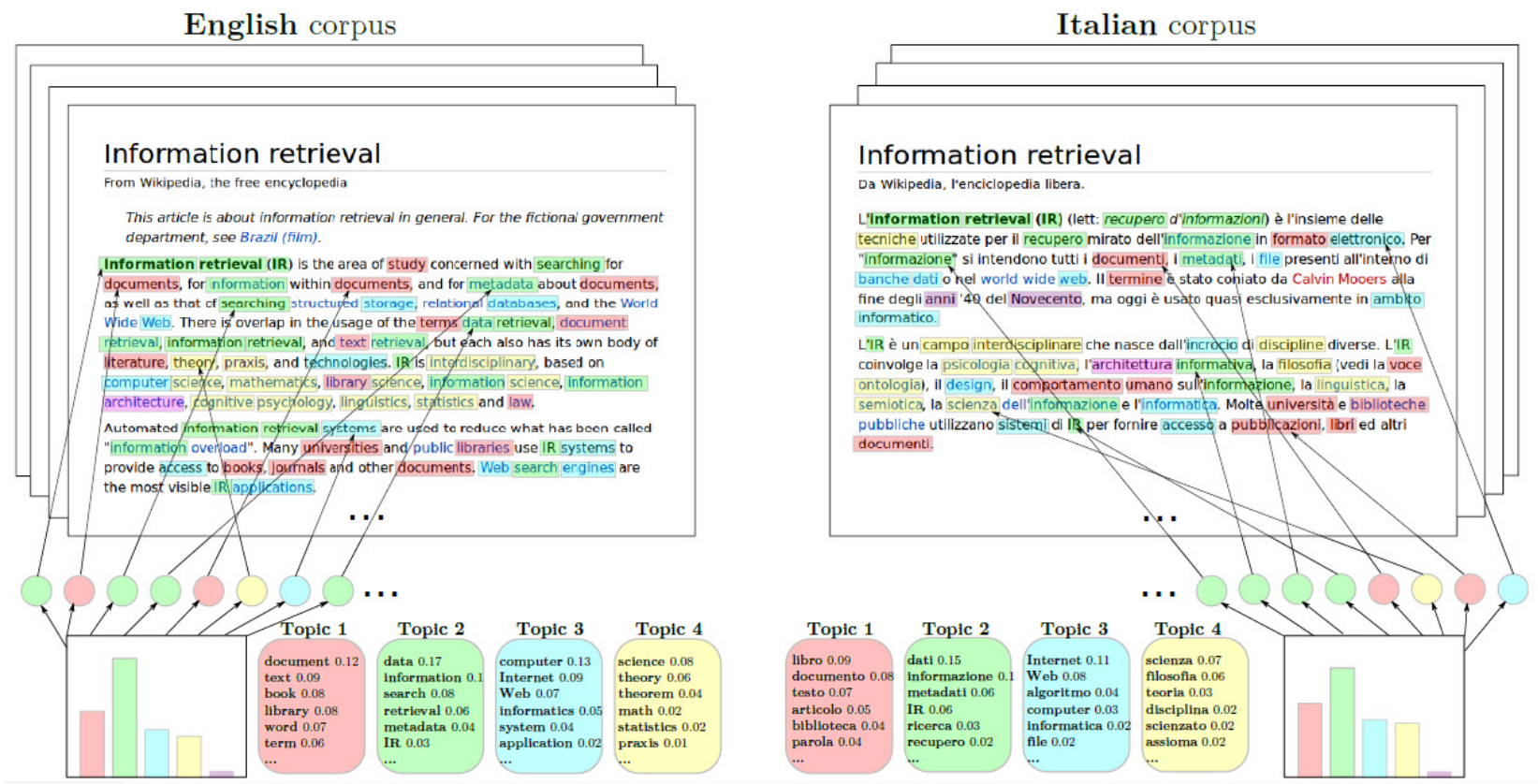
Bottom-up strategy for building topic hierarchy

Idea: to use topics as "words" of the upper level



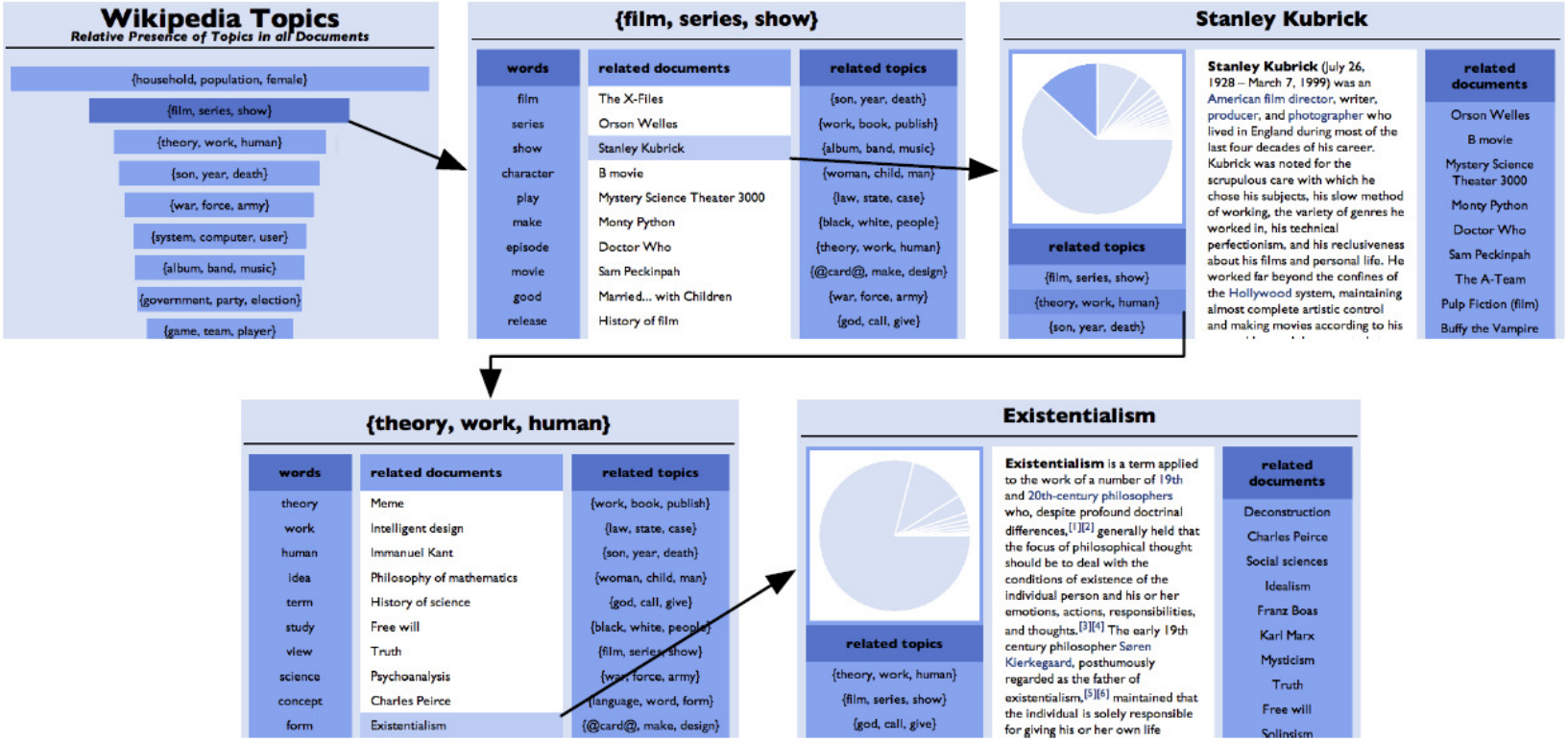
E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749-2775.

Multilingual Topic Models



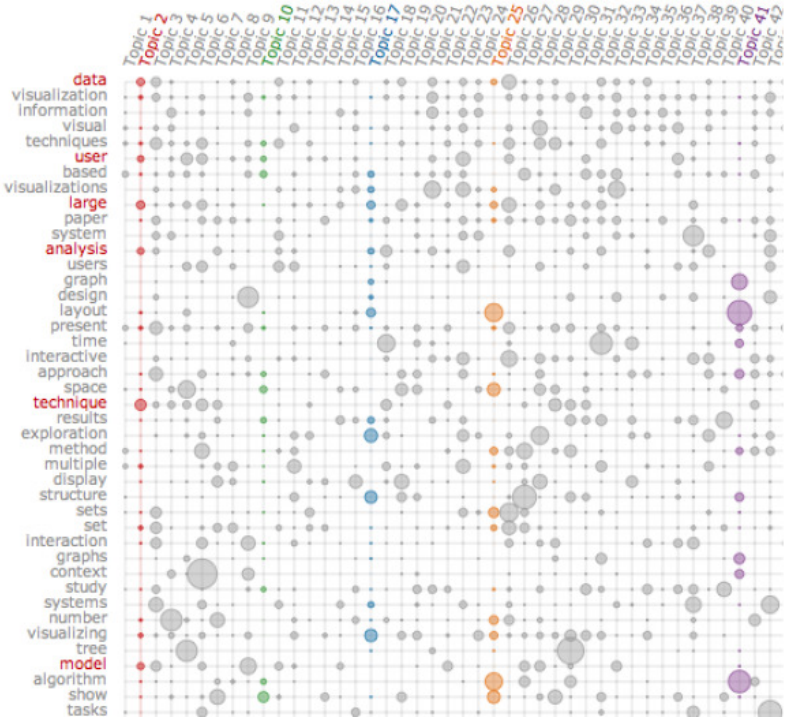
I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

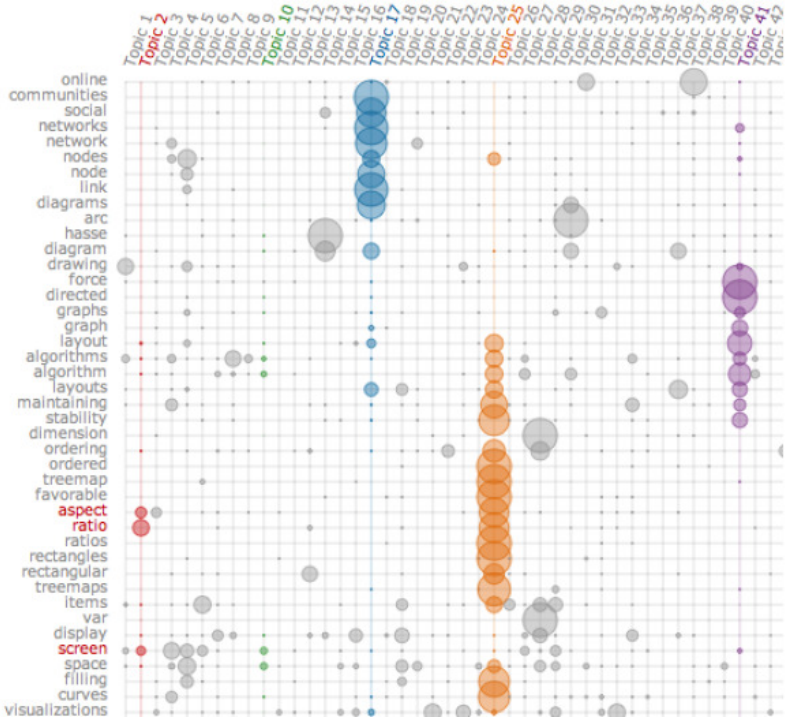
Topic Model Visualization



A. Chaney, D. Blei. Visualizing topic models // International AAAI Conference on Social Media and Weblogs, 2012.

Topic Model Visualization



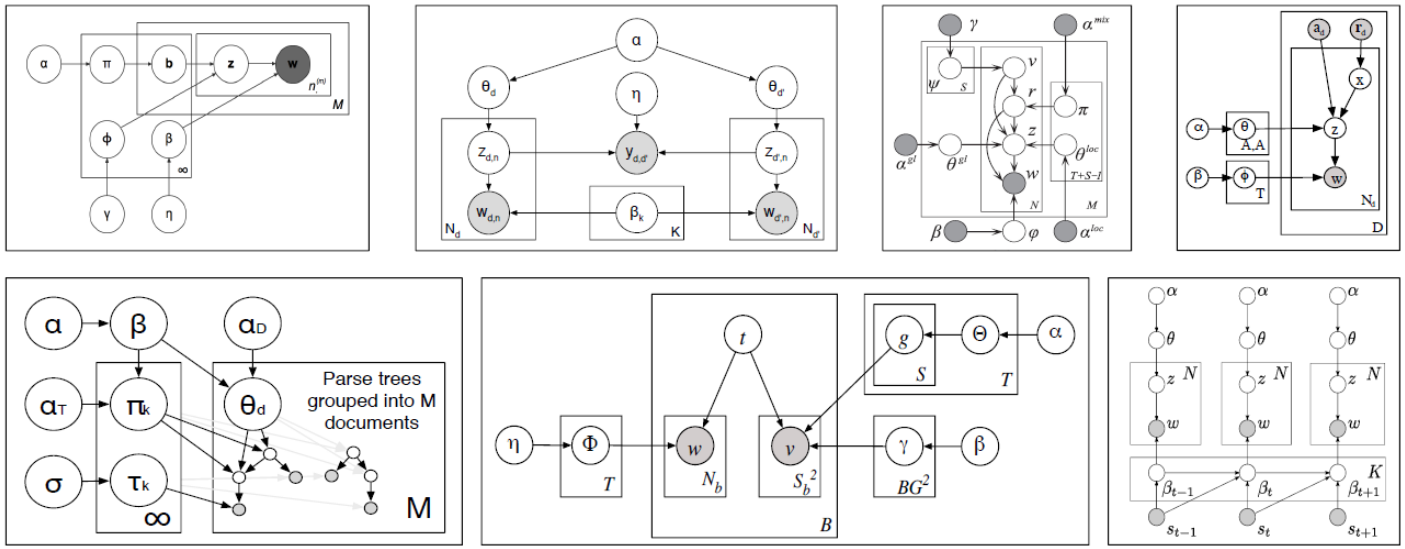


Jason Chuang, Christopher D. Manning, Jeffrey Heer.

Termite: Visualization Techniques for Assessing Textual Topic Models // Advanced Visual Interfaces, 2012

Topic Modeling mainstream

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



David Blei. Probabilistic topic models
Communications of the ACM. 2012. Vol. 55. No. 4. Pp. 77–84.

Topic Modeling Bibliography:
http://mimno.infosci.cornell.edu/topics.html

PLSA — Probabilistic Latent Semantic Analysis [Hofmann 1999]

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 ϕ, θ:
$\phi_{wt} = p(w|t)$ — 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:

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

under non-negativeness and normalization constrains

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

EM-algorithm

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

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s} \phi_{ws}\theta_{sd}};$$

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

M-step gives frequency estimates of conditional probabilities:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d} n_{dwt}}{\sum\limits_{d,w} n_{dwt}}, \qquad \theta_{td} = \frac{n_{td}}{n_d} \equiv \frac{\sum\limits_{w} n_{dwt}}{\sum\limits_{w,t} n_{dwt}},$$

Short notation via proportionality sign $\propto$:

$$\phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

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 $i_{\max}$;
Output: distributions ϕ, θ;

initialize ϕ_{wt}, θ_{td} for all $d \in D,\ w \in W,\ t \in T$;
for all iterations $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td}, n_t, n_d := 0$ for all $d \in D,\ w \in W,\ t \in T$;
  for all documents $d \in D$ and terms $w \in d$
    $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ for all $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw} p_{tdw}$ for all $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ for all $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ for all $d \in D,\ t \in T$;

Usually $i_{\max} = 20..50$ iterations are sufficient. Time is $O(n|T|i_{\max})$.

ARTM — Additive Regularization of Topic Model

The ill-posed problem:
likelihood maximization has infinitely many solutions.

Regularization of the ill-posed problem:

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

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\text{log-likelihood}} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\phi, \theta)}_{R(\phi,\theta)} \to \max_{\phi,\theta},$$

under non-negativeness and normalization restrictions

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

where $\tau_i > 0$ are *regularization coefficients*.

ARTM: EM-algorithm with regularized M-step

M-step for PLSA:

$$\phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

M-step for regularized PLSA:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \qquad \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+;$$

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

Classical Topic Models are particular cases of ARTM:

PLSA: $R(\phi, \theta) = 0$

LDA: $R(\phi, \theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

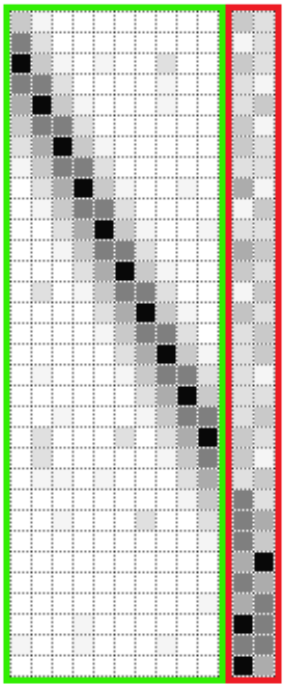
ARTM is a simple and powerful tool for capturing additional criteria, including linguistical requirements and resources.

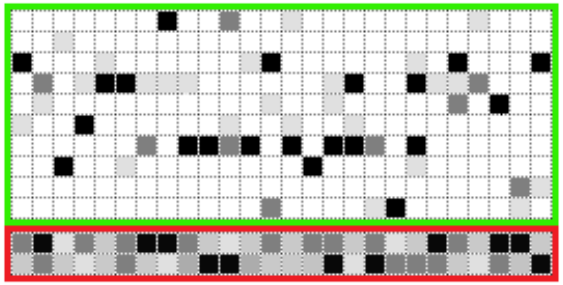
Assumptions: what topics would be well-interpretable?

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

ϕ_{wt} terms×topics θ_{td} topics×documents





Smoothing regularization (rethinking LDA)

The non-sparsity assumption for background topics $t \in B$:
ϕ_{wt} are similar to a given distribution β_w;
θ_{td} are similar to a given distribution α_t.

$$\sum_{t\in B} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

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:

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

which is new non-Bayesian interpretation of LDA [Blei 2003].

Sparsing regularizer (further rethinking LDA)

The sparsity assumption for domain-specific topics $t \in S$: distributions ϕ_{wt}, θ_{td} contain many zero probabilities.

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

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

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

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

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Regularization for topics decorrelation

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

We maximize covariances between column vectors ϕ_t:

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

The regularized M-step makes columns of Φ more distant:

$$\phi_{wt} \propto \Big(n_{wt} - \tau \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \Big)_+ .$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Regularization for topic selection

Assumption: infrequent topics are not well-interpretable.

We maximize KL-divergence $\mathrm{KL}\big(\frac{1}{|T|} \,\big\|\, p(t)\big)$ 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 Θ 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.

Regularization for topic coherence improving

Assumption: if topic is well-interpretable then its top words are *coherent* i. e. frequently appear nearby in the documents.

$C_{uw} = \hat{p}(w|u) = \frac{N_{uw}}{N_u}$ — *coherence* of a word pair $u, w \in W$, N_u, N_{uw} are document frequency of word w and word pair u, w.

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

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

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

The regularized M-step gives a kind of smoothing:

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

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Regularization for semi-supervised learning

Assumption: experts have provided us with topic labeling data:

- each document $d \in D_0 \subseteq D$ belongs to a subset of topics $T_d \subset T$;
- each topic $t \in T_0 \subseteq T$ contains a subset of words $W_t \subset W$.

ϕ^0_{wt} — uniform distribution over subset of terms W_t
θ^0_{td} — uniform distribution over subset of topics T_d

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

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

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

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

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Experiment 1: NIPS Collection. Combining regularizers

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 $\approx 1.3 \cdot 10^4$.

Topic model quality measures

Multi-criteria optimization requires multiple quality measures.

- Hold-out *perplexity*: $\mathcal{P} = \exp(-\frac{1}{n}\mathscr{L})$
- *Sparsity* — the number of zero elements ϕ_{wt} and θ_{td}
- Interpretability measures for each topic t:
  - topic *coherence* [Newman, 2010]
  - topic *kernel size*: $|W_t|$, kernel $W_t = \{w\colon p(t|w) > 0.25\}$
  - topic *purity*: $\sum\limits_{w \in W_t} p(w|t)$
  - topic *contrast*: $\frac{1}{|W_t|} \sum\limits_{w \in W_t} p(t|w)$
- Model degeneracy:
  - number of non-zero topics: $|T|$
  - the fraction of background words: $\frac{1}{n} \sum\limits_{d,w} \sum\limits_{t \in B} p(t|d,w)$

Sparsing + Smoothing + Decorrelation + Topic Selection

ARTM — Additive Regularization of Topic Model

M-step formula for combined regularization:

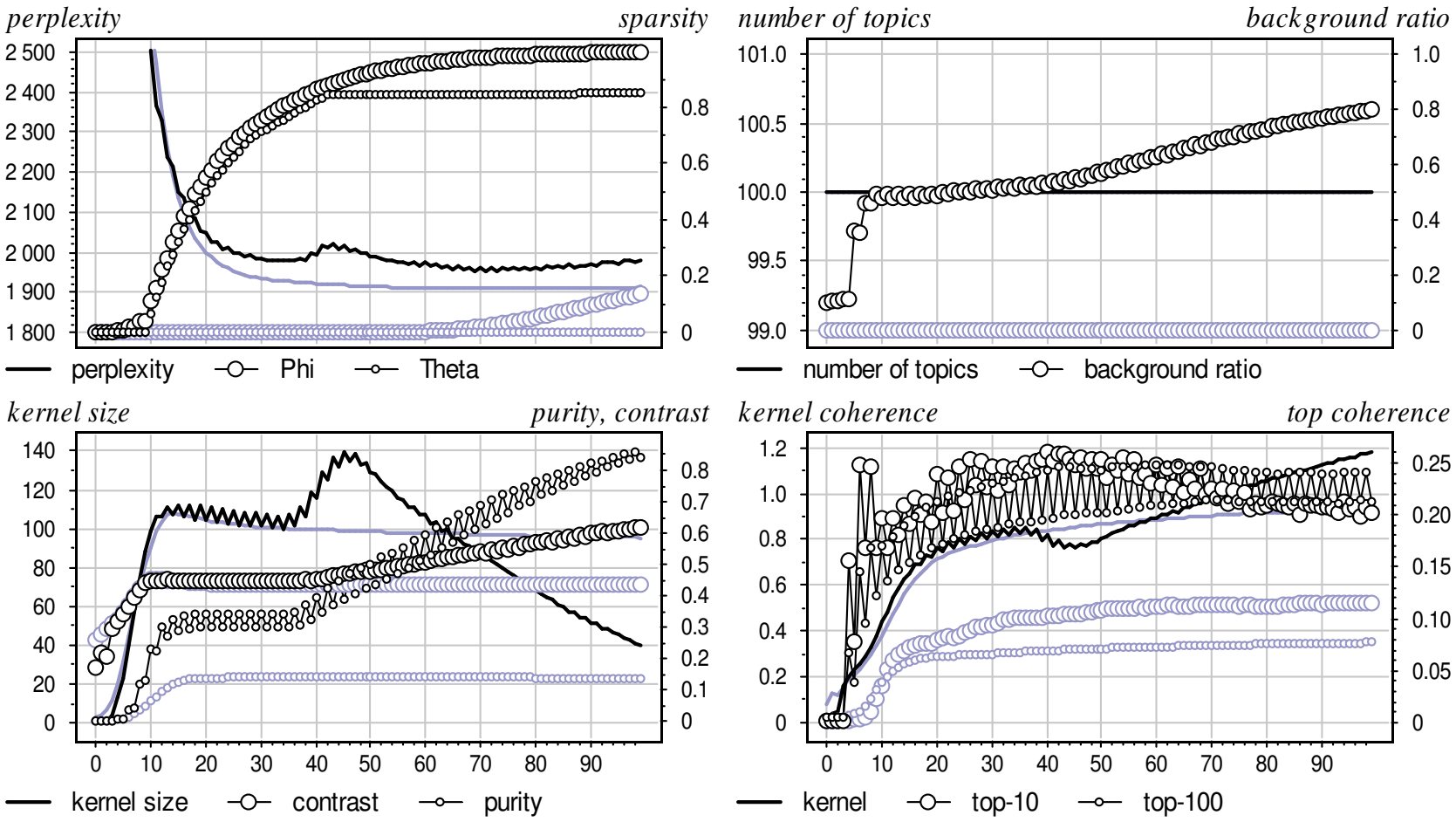
$$\phi_{wt} \propto \Bigg(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{smoothing}\\ \text{background}\\ \text{topics}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{sparsing}\\ \text{specific}\\ \text{topics}}} - \tau_3 \underbrace{\phi_{wt} \sum_{s \in S \setminus t} \phi_{ws}}_{\text{decorrelation}} \Bigg)_+$$

$$\theta_{td} \propto \Bigg(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{smoothing}\\ \text{background}\\ \text{topics}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{sparsing}\\ \text{specific}\\ \text{topics}}} - \tau_6 \underbrace{\frac{n_d}{n_t} \theta_{td}}_{\substack{\text{topic}\\ \text{selection}}} \Bigg)_+$$

A new problem arises: how to choose the *regularization path* $\tau = (\tau_1, \ldots, \tau_6)$ as a function of the iteration number?

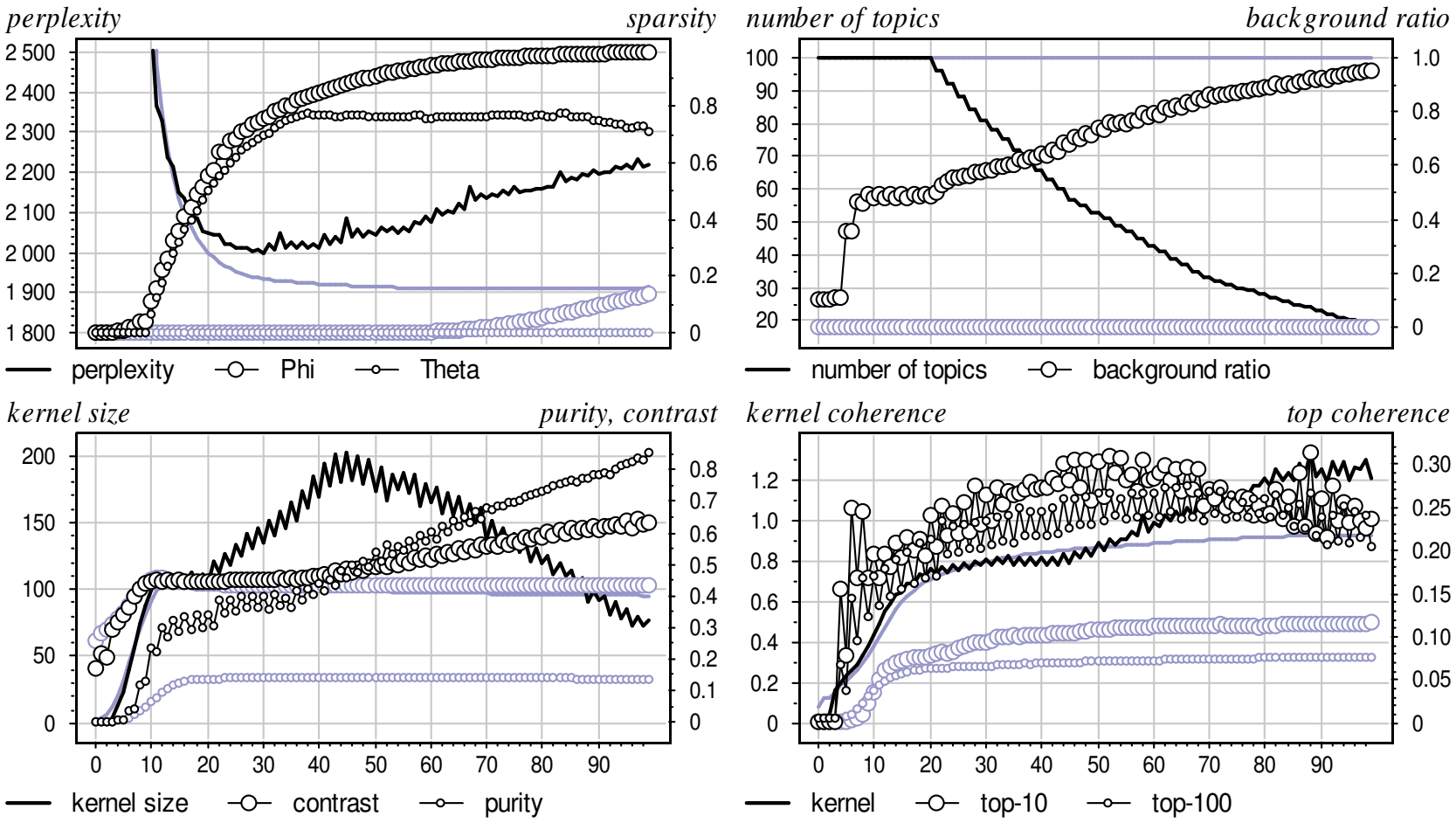
Sparsing + Smoothing + Decorrelation

Quality measures as functions of the iteration number
(grey lines — PLSA, black lines — ARTM)



Sparsing + Smoothing + Decorrelation + Topic Selection

Quality measures as functions of the iteration number
(grey lines — PLSA, black lines — ARTM)



Topics (kernel words are bold)

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

Topics (kernel words are bold)

PLSA:	ARTM:
music	**estimator**
rules	**music**
note	**musical**
representation	**notes**
neural	Mozer
events	**melody**
net	**composition**
set	**Bach**
time	**chorales**
musical	**melodic**
figure	**jackknife**
network	cooperative
notes	**subnet**
input	**GEM**
melody	**melodies**
structure	**ICL**
harmony	**tonal**
tau	**accent**
pitch	augmented
temporal	piece

PLSA:	ARTM:
linear	**compression**
matrix	**quantization**
algorithm	**singular**
vector	**pyramid**
network	**Sanger**
output	decomposition
vectors	generalized
figure	**SVD**
input	**SONN**
singular	**MDL**
data	**primitives**
mapping	**Tenorio**
image	**plasma**
dimensional	matrices
nonlinear	**compressed**
algorithms	reconstruction
inverse	**book**
error	**memorized**
optimal	**Laplace**
compression	**eigen**

Topics (kernel words are bold)

PLSA:	ARTM:
face	**face**
images	**faces**
faces	**facial**
recognition	**Cottrell**
set	**Pentland**
image	**gesture**
based	**lane**
HME	**emotion**
facial	**person**
representation	**steering**
view	**appearance**
figure	**Baluja**
model	**setpoint**
experts	camera
network	tracking
human	**pose**
expert	**Pomerleau**
space	**mouth**
examples	**Darrell**
system	**lighting**

PLSA:	ARTM:
query	**MLP**
set	**query**
queries	**queries**
data	**CART**
algorithm	**documents**
learning	retrieval
documents	relevant
number	**document**
performance	rank
words	sampling
MLP	instances
CART	**splits**
values	collection
cluster	Gibbs
experiments	**sex**
results	**ranking**
relevant	ordering
retrieval	recursive
classification	text
algorithms	axis

Conclusions from experiments

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?

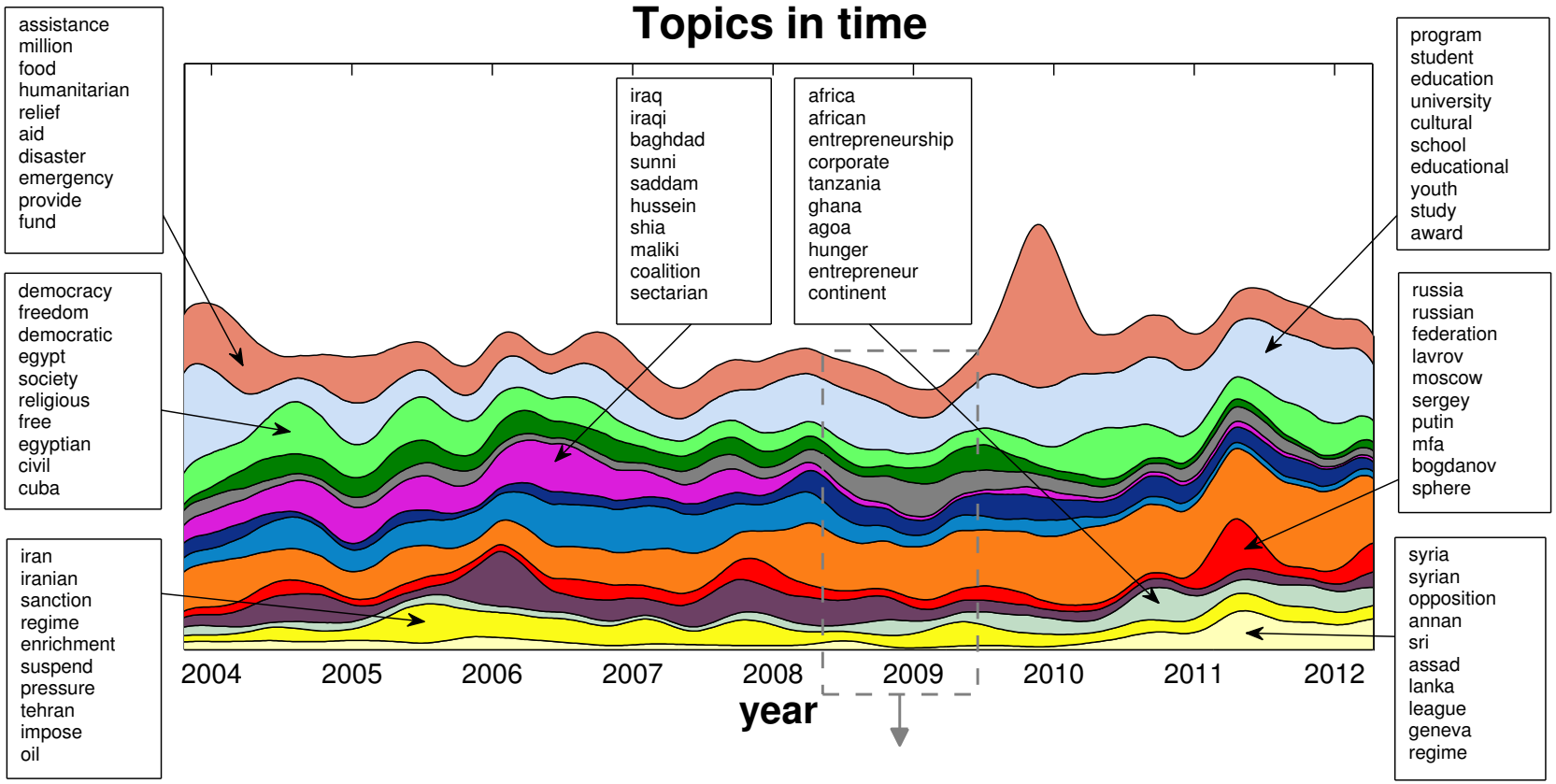
The set of regularizers:

- smoothing background topics
- sparsing domain-specific topics
- sparsing $p(t|\text{time})$ distributions
- penalizing noisy variations of $p(t|\text{time})$

Dataset:

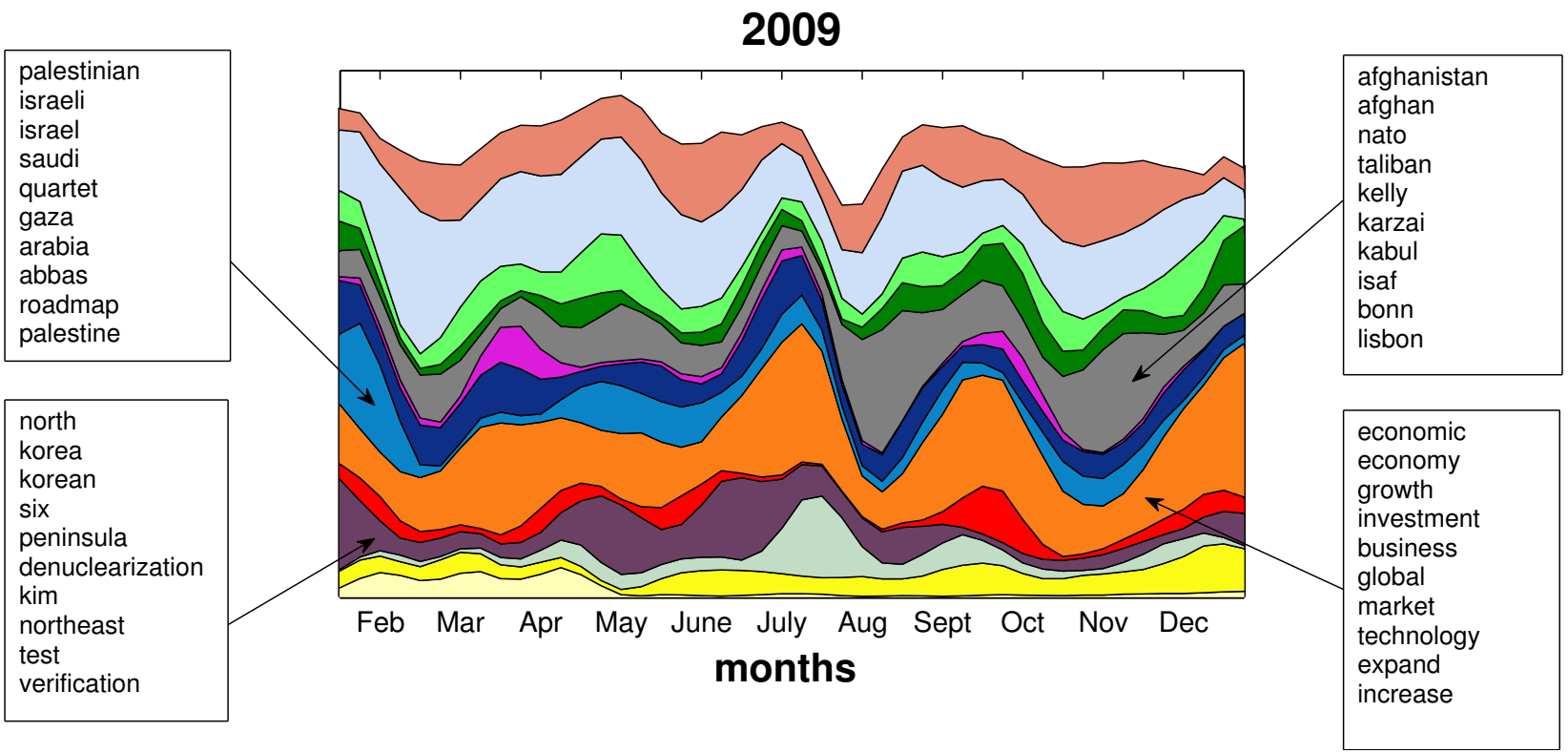
- 20 000 press releases of 4 countries, 2001–2013;

ARTM for temporal Topic Modeling



Experimental work *Nikita Doykov*, MSU, 2014

ARTM for temporal Topic Modeling



Experimental work of *Nikita Doykov*, MSU, 2014

Variety of regularizers for ARTM

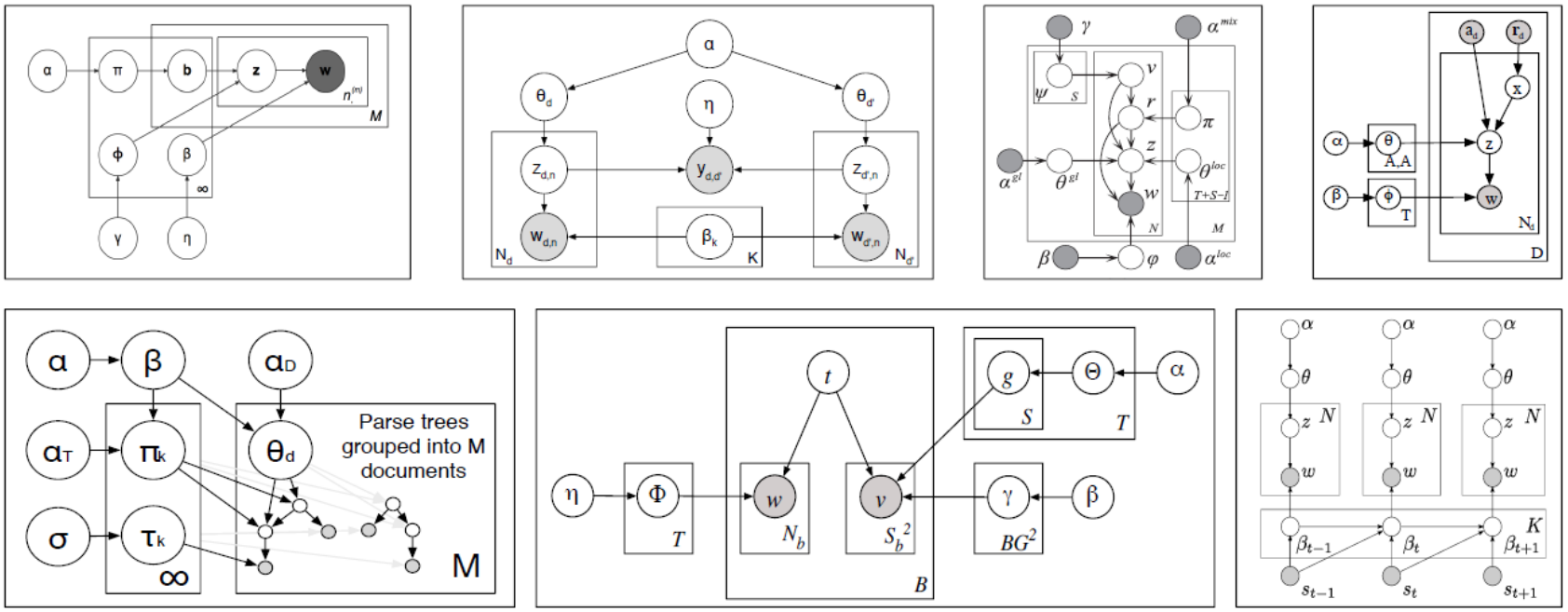
Understood and implemented:

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
Communications of the ACM. 2012. Vol. 55. No. 4. Pp. 77–84.

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

Bibliography

- *Hofmann T.* Probabilistic Latent Semantic Indexing. SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.
- *Teh Y. W., Newman D., Welling M.* A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. NIPS, 2006, Pp. 1353–1360.
- *Porteous I., Newman D., Ihler A., Asuncion A., Smyth P., Welling M.* Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation. KDD 2008.
- *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.
- *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic Evaluation of Topic Coherence // Human Language Technologies, HLT-2010, Pp. 100–108.
- *Yi Wang.* Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.
- *Sato I., Nakagawa H.* Rethinking Collapsed Variational Bayes Inference for LDA. Int'l Conf. on Machine Learning ICML, 2012.
- *Vorontsov K. V.* Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. Pleiades Publisher, 2014. Vol. 88, No. 3.
- *Vorontsov K. V., Potapenko A. A.*, Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'14. Springer. 2014. (to appear)

Vorontsov Konstantin
voron@yandex-team.ru
Wiki www.MachineLearning.ru (in Russian):

- User:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование