The Problems, Methods and Limitations of Machine Intelligence: Mining Texts, Graphs and Hypergraphs

Konstantin Vorontsov
Machine Intelligence Laboratory
(Moscow Institute of Physics and Technology, Russia)
and *Laboratory on Artificial Intelligence* (Sberbank, Russia)

Cognitive Technologies and Quantum Intelligence
ITMO University • Saint Petersburg • 17–19 May, 2018

Contents

Basic machine learning tasks

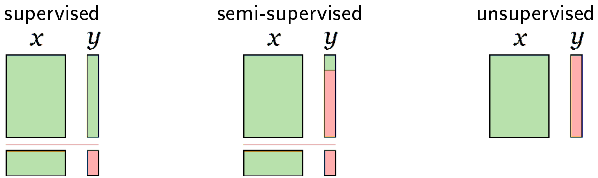
Given a training set of input–output pairs (x_i, y_i), $i = 1, \dots, m$
Find a model $y = f(x, \alpha)$, then predict outputs on a testing set

Supervised learning, e.g. regression, least squares method:

$$\sum_{i=1}^{m} \bigl(f(x_i, \alpha) - y_i\bigr)^2 \;\to\; \min_{\alpha}$$

Unsupervised learning, e.g. clustering, likelihood maximization:

$$\sum_{i=1}^{m} \log p(x_i, \alpha) \;\to\; \max_{\alpha}$$



Feature extraction: problems and approaches

In classical machine learning, objects x_i are represented by vectors.

In many applications, data come in raw non-vector form:

- natural language texts
- time series and signals: econometric, biomedical, etc.
- images and video
- networks: social, technical, transportation, etc.
- transaction data: logs, clickstream, e-commerce, banking, etc.

How to build a vector representation of a poorly structured object?

Approaches:

- feature engineering based on subject domain understanding
- architecture engineering for deep neural networks
- *learning vector representations (embeddings)*

Machine learning problems on graphs and hypergraphs

Graph is a most common structure to describe objects of any nature via their parts, links, interactions, or relationships.

Examples of graph data:

- text document collection is a bipartite graph:
  vertices: documents d and words w;
  edge (d, w) means that the word w occurs in the document d.
- social network data:
  vertices: users;
  edge (u, v) means that user u communicates with user v.
- financial transactions can be described by a *hypergraph*:
  vertices: clients c, firms f, and goods g;
  edge (c, f, g) means that a client c bought goods g from f.

PCA: Principal Component Analysis

Interactions between elements of two finite sets, e.g. W and D

Given

n_{wd}, how many times word $w \in W$ occur in document $d \in D$

Find

ϕ_w: vector representation (embedding) of word w
θ_d: vector representation (embedding) of document d

The problem is to build vectors capable to predict (d, w) pairs:

$$\sum_{d \in D} \sum_{w \in W} \big(n_{wd} - \langle \phi_w, \theta_d \rangle\big)^2 \to \min_{\Phi, \Theta}$$

Solution is a low-rank matrix factorization via gradient descent:

$$\underset{W \times D}{N} \approx \underset{W \times T}{\Phi} \cdot \underset{T \times D}{\Theta}, \qquad |T| \ll |W|, |D|$$

The shortcoming is that vector coordinates are not interpretable.

Recent embedding techniques for texts and graphs

word2vec: word embedding
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: paragraph and document embeddings
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: sentence embeddings
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: symbolic n-gram embeddings
`https://github.com/facebookresearch/fastText`

node2vec: graph nodes embeddings
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: more general graph embeddings
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: any things embeddings (from Facebook AI Research)
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

The shortcoming is that vector coordinates are not interpretable.

Interpretable topical embeddings

Intuitively,

- *Topic* corresponds to a subject area with its own terminology
- *Topic* is a set of terms that often co-occur in documents

More formally,

- *topic* is a probability distribution over terms (words, tokens): $p(w|t)$ is the frequency of term w in topic t
- *document profile* is a probability distribution over *topics*: $p(t|d)$ is the frequency of topic t in document d

When writing term w in document d author thought of topic t.

Topic model uncovers the set T of latent topics in a text collection and gives interpretable embeddings $p(t|w)$, $p(t|d)$.

Example. Multilingual topic model of Wikipedia

216 175 of Russian–English parallel not-aligned articles.
Top 10 words and their probabilities $p(w|t)$ in %:

topic #68				topic #79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Assessors evaluated 396 topics from 400 as paired and interpretable.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Example. Multilingual topic model of Wikipedia

216 175 of Russian–English parallel not-aligned articles.
Top 10 words and their probabilities $p(w|t)$ in %:

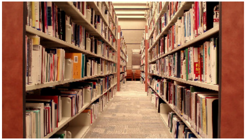
topic #88				topic #251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Assessors evaluated 396 topics from 400 as paired and interpretable.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

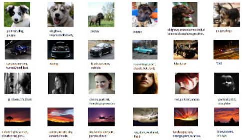
Topic modeling applications

exploratory search in digital libraries

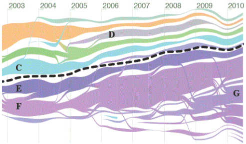


personalized search in topical communities

multimodal search for texts and images

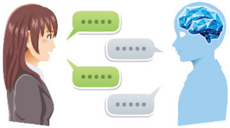


topic detection and tracking in news flows



navigation in big text collections

dialog management in chatbot intelligence



Topic modeling: the problem setup

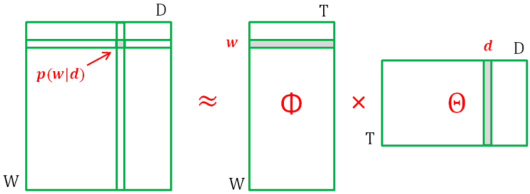
Given: a set of terms (words) W, a set of documents D,
n_{dw} = how many times term w appears in document d

Find: parameters $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ of the topic model

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td} = \sum_{t \in T} p(w|t)p(t|d).$$

subject to $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$.

This is a problem of *nonnegative matrix factorization*:



PLSA — Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]

Constrained maximization of the log-likelihood:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

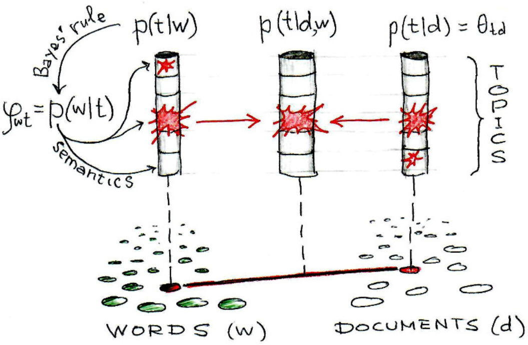
EM-algorithm is a simple iteration method for the nonlinear system

$$\begin{aligned}
&\text{E-step:} \quad && \begin{cases} p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\[1ex]
\phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}\,p_{tdw}\Big) \\[1ex]
\theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}\,p_{tdw}\Big) \end{cases}\\
&\text{M-step:} &&
\end{aligned}$$

where $\operatorname*{norm}_{t\in T}(x_t) = \frac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ is vector normalization.

Interpretable topical embeddings for words and documents

- Text collection is a bipartite graph with (d, w) edges
- Word w has a chance to occur in d when they share same topics
- Topic interpretation comes from $p(w|t)$ due to Bayes' rule



Well-posed and ill-posed problems in the sense of Hadamard (1923)

The problem is *well-posed* if

- a solution exists,
- the solution is unique,
- the solution is stable w.r.t. initial conditions.

Jacques Hadamard (1865–1963)

Matrix factorization is an *ill-posed* inverse problem.
If (Φ, Θ) is a solution, then (Φ', Θ') is also the solution:

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, where $\operatorname{rank} S = |T|$
- $\mathscr{L}(\Phi', \Theta') = \mathscr{L}(\Phi, \Theta)$
- $\mathscr{L}(\Phi', \Theta') \leqslant \mathscr{L}(\Phi, \Theta) + \varepsilon$ for approximate solutions

Additional *regularizing criteria* should narrow the set of solutions.

LDA — Latent Dirichlet Allocation [D.Blei, A.Ng, M.Jordan, 2003]

Maximize a posteriori probability (MAP) with Dirichlet prior.
The prior can be reinterpreted as cross-entropy minimization:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{cross-entropy regularizer}} \to \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw}p_{tdw} + \beta_w\Bigr) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw}p_{tdw} + \alpha_t\Bigr) \end{array} \right. \\ \text{M-step:} & \end{array}$$

ARTM — Additive Regularization for Topic Modeling

Maximize log-likelihood with regularization criterion $R(\Phi,\Theta)$:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{array}{ll}\text{E-step:} & \\ \\ \text{M-step:} & \\ \\ & \end{array}\begin{cases} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

K.Vorontsov. Additive regularization for topic models of text collections. 2014.

ARTM: combining topic models via additive regularization

Maximize log-likelihood with additive combination of regularizers:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

where τ_i are regularization coefficients.

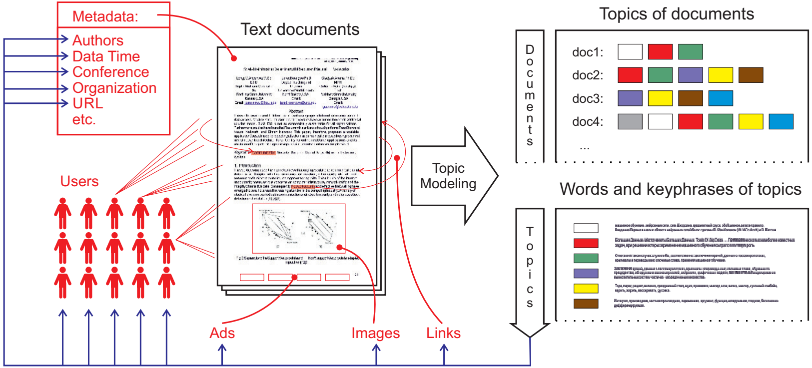
EM-algorithm is a simple iteration method for the system

$$\begin{array}{ll} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum\limits_{d\in D} n_{dw}p_{tdw} + \sum\limits_{i=1}^{n}\tau_i\phi_{wt}\frac{\partial R_i}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{w\in d} n_{dw}p_{tdw} + \sum\limits_{i=1}^{n}\tau_i\theta_{td}\frac{\partial R_i}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

K.Vorontsov, A.Potapenko. Additive regularization of topic models. Machine Learning, 2015.

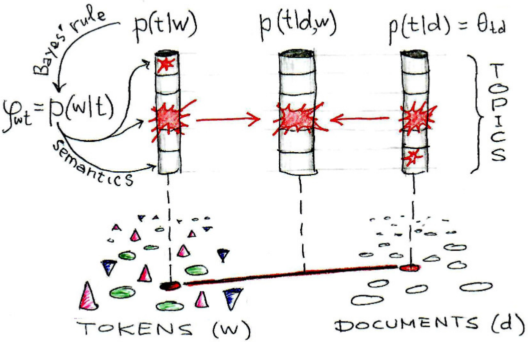
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topic distributions of terms $p(w|t)$ and *tokens* of other *modalities*: $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{tag}|t)$, $p(\text{category}|t)$, $p(\text{link}|t)$, $p(\text{object-on-image}|t)$, $p(\text{user}|t)$, etc.



Interpretable topical embeddings for multimodal documents

- Documents contain words and tokens of other *modalities*
- Examples of modalities: authors, date-time, tags, users, etc.
- Topics propagate semantics from words to other modalities



Multimodal extension of ARTM

W^m is a vocabulary of *tokens* of m-th *modality*, $m \in M$.

Maximize the sum of modality log-likelihoods with regularization:

$$\sum_{m\in M} \lambda_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}$$

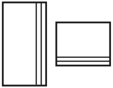
EM-algorithm is a simple iteration method for the system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W^m}\Big(\sum_{d\in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina. Non-Bayesian additive regularization for multimodal topic modeling of large collections. 2015.

Regularizers for the interpretability of topics

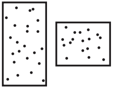
background



LDA: Smoothing background topics $B \subset T$:

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

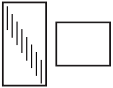
sparse



"Anti-LDA": Sparsing subject domain topics $S = T \setminus B$:

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

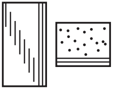
decorrelated



Making topics as different as possible:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \phi_{wt} \phi_{ws}$$

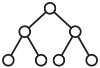
interpretable



Making topics more interpretable
by combining the above regularizers

Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

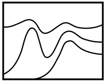
hierarchy



Hierarchical links between topics t and subtopics s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}.$$

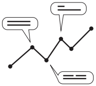
temporal



Topics dynamics over the modality of time intervals i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|.$$

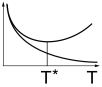
regression



Linear predictive model $\hat{y}_d = \langle v, \theta_d \rangle$ for documents:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big(y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2.$$

n of topics

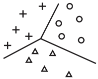


Sparsing $p(t)$ for topic selection:

$$R(\Theta) = -\tau \sum_{t \in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_d p(d) \theta_{td}.$$

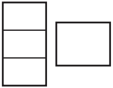
Special cases of the multimodal topic modeling

supervised



The modalities of classes or categories for text classification and categorization.

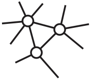
multilanguage



The modalities of languages with translation dictionary $\pi_{uwt} = p(u|w,t)$ for the $k \to \ell$ language pair:

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt} \phi_{wt}$$

graph



The modality of graph vertices v with doc sets D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} S_{uv} \sum_{t \in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|} \right)^2.$$

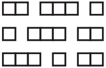
geospatial

The modality of geolocations g with proximity $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g' \in G} S_{gg'} \sum_{t \in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}} \right)^2$$

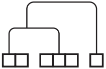
Beyond the “bag-of-words” restrictive hypothesis

n-gram



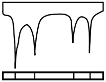
The modalities of n-grams, collocations, named entities

syntax



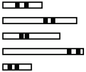
The modality of n-grams after SyntaxNet preprocessing

segmentation



E-step regularization affecting $p(t|d,w)$ distributions for segmentation and sentence topic models

coherence



Modeling co-occurrence data n_{uv} for biterms (u,v):

$$R(\Phi) = \tau \sum_{u,v} n_{uv} \ln \sum_t n_t \phi_{ut} \phi_{vt}$$

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

BigARTM: open source for fast and modular topic modeling

BigARTM features:

- Parallelism + modalities + regularizers + hypergraph NEW
- Out-of-core one-pass processing of large text collections
- Built-in library of regularizers and quality measures

BigARTM community:

- Open-source https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Documentation http://bigartm.org

BigARTM license and programming environment:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

Why does BigARTM simplify topic modeling for applications

<table>
<tr><th>Stages</th><th>Bayesian Inference for PTMs</th><th colspan="2">ARTM</th></tr>
<tr><td>Requirements analysis:</td><td>Requirements analysis</td><td colspan="2">Requirements analysis</td></tr>
<tr><td>Model formalization:</td><td>Generative model design</td><td>predefined criteria</td><td>user-defined criteria</td></tr>
<tr><td>Model inference:</td><td>Bayesian inference for the generative model (VI, GS, EP)</td><td colspan="2">One regularized EM-algorithm for any combination of criteria</td></tr>
<tr><td>Model implementation:</td><td>Researchers coding (Matlab, Python, R)</td><td colspan="2">Production code (C++)</td></tr>
<tr><td>Model evaluation:</td><td>Researchers coding (Matlab, Python, R)</td><td>predefined measures</td><td>user-defined measures</td></tr>
<tr><td>Deployment:</td><td>Deployment</td><td colspan="2">Deployment</td></tr>
</table>

conventions: ::: not unified stages ::: ::: unified stages :::

Bayesian modeling requires maths and coding at each stage.

ARTM introduces the modular "LEGO-style" technology, packing each requirement into a *regularization plugin*.

Benchmarking BigARTM vs. Gensim and Vowpal Wabbit

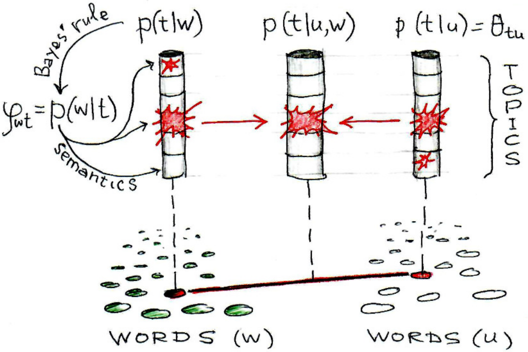
- 3.7M articles from Wikipedia, 100K unique words

		$T = 50$		$T = 200$	
	procs	time, m	perplexity	time, m	perplexity
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Interpretable topical embeddings for word co-occurrence

- The idea of *distributional semantics*: "Words that occur in the same contexts tend to have similar meanings" [Harris, 1954].
- Word induces a pseudo-document that joins all its contexts



Examples of vector operations in word similarity tasks

Take the best of the two approaches:

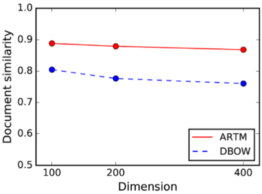
- **ARTM**: sparse interpretable vector components
- **word2vec**: interpretable vector addition and subtraction

vector operation	ARTM result	word2vec result
king − boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow − russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india − russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars − car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Quantitative estimation on document similarity tasks

ArXiv triplets dataset of 20K triplets of papers:
⟨ paper A, similar paper B, dissimilar paper C ⟩



- trained on 1M ArXiv plain texts
- tested on the ArXiv triplets
- DBOW is a well-known paragraph2vec architecture [Dai et. al, 2015]

ARTM-PWE (probabilistic word embeddings) outperforms DBOW (distributed bag-of-words) model.

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Transaction data

Data may contain not only pairs (d, w) but also *transactions* — triples, …, n-tuples of tokens of different nature (modality).

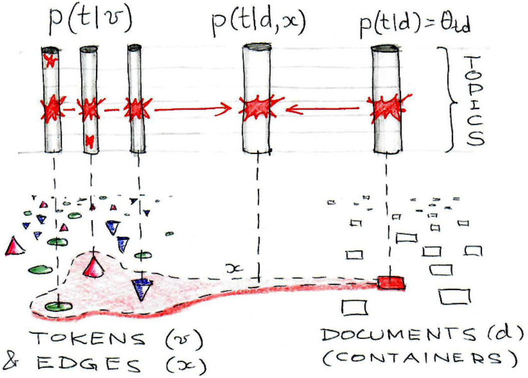
Examples of triple *transactions*:

- **Social network data:**
  (d, u, w) — the user u wrote the word w in the blog d
- **Advertising network data:**
  (u, d, b) — the user u clicked on the banner b on the page d
- **Recommender system data:**
  (u, m, s) — the user u rated the movie m in the situation s
- **Banking and retail data:**
  (b, s, g) — the buyer u bought the goods g from the seller s

The problem: *giving* an observable set of transactions *find* the latent distribution $p(t|v)$ of topics t for each token v.

Interpretable topical embeddings for transaction data

- A hypergraph is defined as a system of subsets of vertices
- Transaction = a subset of tokens = an edge of hypergraph
- Transaction occurs if its tokens share same topics

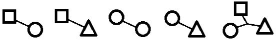


Hypergraph Topic Model: definitions and notations

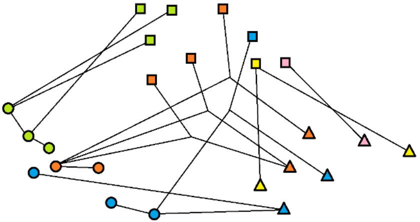
$\Gamma = \langle V, E \rangle$ is an oriented hypergraph,
vertices are tokens of different modalities, *edges* are transactions,
$V = V^1 \sqcup \dots \sqcup V^M$ is a disjoint union of vertex subsets by modality,

M is the set of modalities:

K is the set of edge types:



T is the set of topics:



X^k is the set of observable hyper-edges (transactions) of type k,
the edge (d, x): *container* vertex $d \in V$, common vertices $x \subset V$,
n_{dx} is the number of transactions (d, x) in the dataset X^k,
$p_k(d, x)$ is an unknown probability measure over edges of type k.

Hypergraph Topic Model: likelihood maximization

Hypergraph Topic Model of hyper-edges (transactions) of type k:

$$p_k(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{kvt},$$

$\theta_{td} = p(t|d)$ is the topic distribution of the container d
$\phi_{kvt} = p_k(v|t)$ is the vertex distribution of the topic t

Maximize the log-likelihood for transactions of type k:

$$\sum_{dx \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{kvt} \to \max_{\Phi,\Theta},$$

$$\phi_{kvt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{kvt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1.$$

Hypergraph extension of ARTM

Maximize the weighted sum of log-likelihoods with regularization:

$$\sum_{k\in K} \tau_k \sum_{dx\in X^k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \phi_{kvt} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

where parameter $\tau_k > 0$ is the weight of edges of type k.

EM-algorithm is a simple iteration method for the system

$$\begin{aligned}
&\text{E-step:} && p_{ktdx} = \operatorname*{norm}_{t\in T}\Bigl(\theta_{td} \prod_{v\in x} \phi_{kvt}\Bigr) \\
&\text{M-step:} && \phi_{kvt} = \operatorname*{norm}_{v\in V^m}\Bigl(\sum_{dx\in X^k} [v\in x]\, \tau_k\, n_{dx}\, p_{ktdx} + \phi_{kvt}\frac{\partial R}{\partial \phi_{kvt}}\Bigr) \\
& && \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{k\in K} \tau_k \sum_{dx\in X^k} n_{dx}\, p_{ktdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr)
\end{aligned}$$

Sentence topic models: TwitterLDA and senLDA

S_d is a set of sentences in document d

n_{sw} = how many times term w appears in sentence s

Topic model of sentence:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}$$

Maximization of the regularized log-likelihood

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

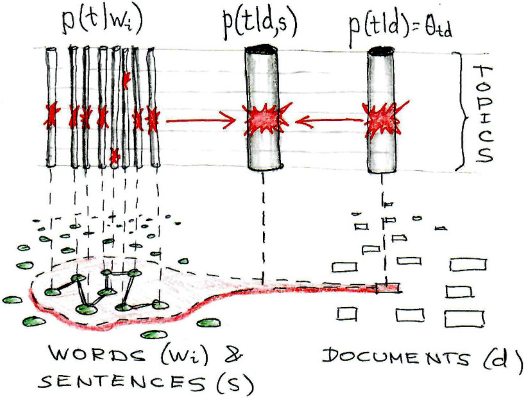
is a special case of hypergraph topic modeling with sentences considered as transactions.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Interpretable topical embeddings for sentences

- Sentence s occurs if its words share same topics
- Sentence is a most semantically definite unit of natural language
- Sentence can be represented by an edge of hypergraph



Conclusion

- *Vector representation* (embedding) is a common approach to make machine learning models applicable to graphs, hypergraphs and raw transaction data
- *Topic modeling* gives interpretable embeddings and propagates semantics from words through topics to other modalities
- Hundreds of known topic models can be expressed in *additive regularization* framework (ARTM) and combined
- ARTM originates the modular "LEGO-style" topic modeling technology implemented in the open source project BigARTM (now including hypergraphs)

http://bigartm.org

Discussion

- Is there anything in common between topical vector representations of and wave functions?
- What are the perspectives for implementing the EM-like algorithms on a quantum computer?
- Will quantum computing handle large amounts of text/transaction data with superlinear speed?

ARTM and BigARTM references

[1] *K.Vorontsov.* Additive regularization for topic models of text collections. Doklady Mathematics, 2014.

[2] *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.

[3] *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM, 2015.

[4] *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS, 2015.

[5] *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova.* BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST, 2015.

[6] *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST, 2016.

[7] *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov.* Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

[8] *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

[9] *A.Ianina, L.Golitsyn, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

[10] *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.

[11] *D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.* Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.