Chapter 1
Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization

Konstantin Vorontsov

Abstract Probabilistic Topic Modeling with hundreds of its models and applications has been an efficient text analysis technique for almost twenty years. This research area has evolved mostly within the frame of the Bayesian learning theory. For a long time, the possibility of learning topic models with a simpler conventional (non-Bayesian) regularization remained underestimated and rarely used. The framework of Additive Regularization for Topic Modeling (ARTM) fills this gap. It dramatically simplifies the model inference and opens up new possibilities for combining topic models by just adding their regularizers. This makes the ARTM a tool for synthesizing models with desired properties and gives rise to developing the fast online algorithms in the BigARTM open-source environment equipped with a modular extensible library of regularizers. In this paper, a general iterative process is proposed that maximizes a smooth function on unit simplices. This process can be used as inference mechanism for a wide variety of topic models. This approach is believed to be useful not only for rethinking probabilistic topic modeling, but also for building the neural topic models increasingly popular in recent years.

1.1 Introduction

Topic modeling is a popular natural language processing technique, which has been actively developed since the late 1990s and still finds many applications [6, 10, 30, 16]. A probabilistic topic model reveals the latent thematic structure of a text document collection representing each topic by a probability distribution over words, and describing each document with a probabilistic mixture of topics.

Topic modeling can be considered as a soft clustering of documents. Unlike conventional hard clustering, a document is allocated among several topical clusters

Konstantin Vorontsov
Federal Research Center “Computer Science and Control” of RAS and
M.V.Lomonosov Moscow State University, e-mail: voron@mlsa-iai.ru

instead of belonging entirely to one cluster. Topic models are also called soft biclustering, since the words are also distributed over topics.

The problem of topic modeling of a text document collection is posed as a low-rank matrix factorization. This is an ill-posed problem, which may have infinitely many solutions. Regularizers are introduced to impose additional restrictions on the model and make the solution more stable [50]. In complex problems, there can be several regularizers.

Starting with the LDA, Latent Dirichlet Allocation model [7], Bayesian learning remains the dominant approach in topic modeling. Its main disadvantage is that the inference process is unique for each model, and the more complex the model, the more difficult its calculations. There are currently no easy ways to automate the inference as well as to construct complex models from the simpler ones. Bayesian regularization is introduced via prior distributions, however, the use of optimization criteria is more convenient and commonly accepted. Many models assume Dirichlet prior distributions, which simplifies Bayesian inference due to the conjugacy property. It was mathematical convenience that predetermined the special role of the Dirichlet distribution in topic modeling, despite the lack of convincing linguistic justifications. Finally, the Bayesian inference is inconvenient to combine with neural network learning procedures [63]. The above barriers prevent the topic modeling from the widespread adoption. Topic models more complicated than LDA are rarely used in the text analysis industry. Hundreds of models remain “the studies for one paper”.

The disadvantages mentioned above are overcome in the Additive Regularization of Topic Models (ARTM), which is an approach based on classical non-Bayesian regularization [54, 56]. As shown in [33], a wide class of Bayesian topic models can be restated in terms of ARTM. After that, it is possible to transfer regularizers from one model to another or to combine the regularizers from various models into a composite model with the required properties. For learning any ARTM models, a general algorithm is used, in which regularizers can be added as plug-ins. The modular technology for ARTM is implemented in the open source library `BigARTM`, `http://bigartm.org` [57, 21]. Let us emphasize that ARTM is a general framework for inferring and combining topic models rather than another model or method.

In this paper, an even more general approach is proposed. A theorem on the maximization of a smooth function on unit simplices is proven. From this theorem, a family of iterative EM-like algorithms can be inferred for learning topic models of various structures with arbitrary smooth regularizers. In fact, topic modeling becomes a theory of a single theorem.

An iteration of the general algorithm is not much different from the gradient step of a neural network learning process. This observation opens up new perspectives for learning neural topic models, as well as learning neural networks with non-negativity and normalization constraints imposed on some of the parameter vectors.

1.2 Maximization on unit simplices

Define the norm operator, which transforms an arbitrary numeric vector $(x_i)_{i\in I}$ into a non-negative normalized vector:

$$p_i = \operatorname*{norm}_{i\in I}(x_i) = \frac{(x_i)_+}{\sum\limits_{k\in I}(x_k)_+}, \text{ for all } i \in I,$$

where $(x)_+ = \max\{0, x\}$ is a positive part operation. If $x_i \leqslant 0$ for all $i \in I$, then the result of the norm operator is the null vector. Otherwise, the vector $(p_i)_{i\in I}$ lies on the unit simplex and defines a discrete probability distribution on a finite set I.

Theorem 1 *Let the function $f(\Omega)$ be continuously differentiable with respect to the set of vectors $\Omega = (\omega_j)_{j\in J},\ \omega_j = (\omega_{ij})_{i\in I_j}$. If ω_j is the vector of the local extremum of the mathematical programming problem*

$$f(\Omega) \to \max_{\Omega}, \quad \sum_{i\in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j,\ j \in J$$

and if $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$ for some i, then ω_j satisfies the equations

$$\omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right). \tag{1.1}$$

Proof. The Lagrangian of the optimization problem with non-negativity and normalization constraints is

$$\mathscr{L}(\Omega) = f(\Omega) - \sum_{j\in J}\lambda_j\left(\sum_{i\in I_j}\omega_{ij} - 1\right) + \sum_{j\in J}\sum_{i\in I_j}\mu_{ij}\omega_{ij},$$

with λ_j and μ_{ij} factors corresponding to normalization and nonnegativity constraints respectively. Equate the partial derivatives of the Lagrangian to zero, as required by the Karush–Kuhn–Tucker conditions:

$$\frac{\partial \mathscr{L}}{\partial \omega_{ij}} = \frac{\partial f}{\partial \omega_{ij}} - \lambda_i + \mu_{ij} = 0; \quad \mu_{ij}\omega_{ij} = 0. \tag{1.2}$$

Multiplying both sides of the equation (1.2) by ω_{ij}, one gets

$$\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = \omega_{ij}\lambda_j.$$

Denote the left side of the equality by A_{ij}. Then $A_{ij} = \omega_{ij}\lambda_j$. According to the condition of the theorem, there exists i such that $A_{ij} > 0$. Consequently, $\lambda_j > 0$. If $\frac{\partial f}{\partial \omega_{ij}} < 0$ for some i, then $\mu_{ij} = \lambda_i - \frac{\partial f}{\partial \omega_{ij}} > 0$, consequently, $\omega_{ij} = 0$.

Combining the equation $\omega_{ij}\lambda_t = A_{ij}$ for $A_{ij} > 0$ with a zero solution $\omega_{ij} = 0$ for $A_{ij} \leqslant 0$, we get $\omega_{ij}\lambda_j = (A_{ij})_+$. Summing these equations over i, express the dual variable: $\lambda_j = \sum_{i\in I_j}(A_{ij})_+$. Substituting λ_j into the formula $\omega_{ij} = \frac{1}{\lambda_j}(A_{ij})_+$, we get the required equation (1.1).

The theorem is proven.

The simple iteration method can be used to solve the system numerically. The update formula (1.1) is similar to the gradient maximization step $\omega_{ij} = \omega_{ij} + \eta\frac{\partial f}{\partial \omega_{ij}}$. In both cases, the gradient of $f(\Omega)$ is calculated. Three differences are worth noting: instead of an additive gradient step, a multiplicative update is used, the vector is projected onto the unit simplex by the norm operator, and the step size η is irrelevant.

Assuming that (1.1) is always applicable consider the iterative process

$$\omega_{ij}^{t+1} = \operatorname*{norm}_{i\in I_j}\left(\omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^t}\right), \quad t = 0, 1, 2, \ldots$$

Theorem 2 *Let $f(\Omega)$ be an upper bounded, continuously differentiable function, and all Ω^t, starting from some iteration t^0, satisfy the following conditions:*

- $\forall j \in J\ \ \forall i \in I_j\ \ \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ *(keeping zeros)*
- $\exists \epsilon > 0\ \ \forall j \in J\ \ \forall i \in I_j\ \ \omega_{ij}^t \notin (0, \epsilon)$ *(separation from zero)*
- $\exists \delta > 0\ \ \forall j \in J\ \ \exists i \in I_j\ \ \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ *(nondegeneracy)*

Then $f(\Omega^{t+1}) > f(\Omega^t)$ and $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ under $t \to \infty$.

This theorem was proved by I. A. Irkhin as a generalization of his convergence results for the EM-algorithm in topic modeling [27].

1.3 Probabilistic topic modeling

Consider the collection D of text documents composed of terms from a vocabulary W. The *terms* can be words, lemmatized words, n-grams or phrases, depending on the methods used for text preprocessing. Each document $d \in D$ is a sequence of terms $w_1, w_2, \ldots, w_{n_d}$, where n_d means the document length. Under the "bag of words" hypothesis, the order of terms does not matter, then the document d can be represented compactly by a conditional distribution $\hat{p}(w \mid d) = \frac{n_{dw}}{n_d}$, where n_{dw} counts how many times the term w occurs in the document d.

Conditional independence is the assumption that each topic generates terms regardless of the document: $p(w \mid t) = p(w \mid d, t)$. According to this assumption and the law of total probability,

$$p(w \mid d) = \sum_{t\in T} p(w \mid t)\, p(t \mid d) = \sum_{t\in T} \varphi_{wt}\theta_{td}. \tag{1.3}$$

Probabilistic Topic Model, PTM (1.3) describes how documents are generated from the known distributions $p(w\,|\,t)$ and $p(t\,|\,d)$. Learning PTM from data is an inverse problem: given a collection estimate model parameters $\varphi_{wt} = p(w\,|\,t)$ and $\theta_{td} = p(t\,|\,d)$. In the matrix form, $\Phi = (\varphi_{wt})_{W\times T}$ and $\Theta = (\theta_{td})_{T\times D}$.

Log-likelihood maximization is usual learning criterion for PTMs:

$$\ln \prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta} \tag{1.4}$$

with linear constraints that make columns nonnegative and normalized:

$$\sum_{w\in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{1.5}$$

For a better understanding of topic modeling consider the learning problem (1.4)–(1.5) from four points of view.

Firstly, it is a problem of approximate low-rank matrix factorization. The rank $|T|$ is usually much smaller than both $|D|$ and $|W|$ dimensions. The problem is ill-posed because its solution is not unique: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$ for infinitely many nonsingular S matrices. Regularization can be added to the main criterion in order to make the solution better defined and more stable using an extra knowledge or data.

Secondly, it is a document auto-encoder. The encoder $f_\Phi\colon \frac{n_{dw}}{n_d} \to \theta_d$ transforms $|W|$-dimensional sparse vector representation of the document $\hat{p}(w\,|\,d)$ into $|T|$-dimensional topical embedding $\theta_d = p(t\,|\,d)$. Linear decoder $g_\Phi\colon \theta_d \to \Phi\theta_d$ attempts to reconstruct the original representation as accurately as possible. Matrix Φ is a parameter of both encoder and decoder. The matrix $\Theta = (\theta_1, \ldots, \theta_D)$ is the result of all documents encoding. This important difference between the matrices Φ and Θ becomes obscure if considered only from the matrix factorization point of view.

Thirdly, it is a soft bi-clustering of both documents and terms by topical clusters T. Each document d and each term w are softly allocated to all clusters according to the distributions $p(t\,|\,d)$ and $p(t\,|\,w)$ respectively, instead of being hardly assigned to only one cluster. The model is also capable of estimating topic distribution for a term in a document $p(t\,|\,d,w)$, for a sentence $p(t\,|\,s)$, and for arbitrary text fragment. In general, we call a distribution $p(t\,|\,x)$ for an object x the *topical embedding* of x.

Fourth, it is a language model that predicts the occurrence of words in documents. Admittedly, conventional topic models are bad competitors in this role. Good word predictions are possible only from local contexts, however, they are violated by the bag-of-words hypothesis. In topic modeling, many ways have been proposed to go beyond this hypothesis and process text as a sequence of terms. Another flaw is more fatal: one can hardly expect that the appearance of a word is determined only by its topics, even if they were estimated from the local context. Deep neural networks based on attention models [51] and transformer architecture, such as BERT [17] and GPT-3 [11] capture the entire set of linguistic phenomena and predict words in a text much better than PTMs and even better than humans do. However, these models are non-interpretable: it is impossible to understand which phenomena are captured, and what each coordinate of the text embedding means.

In contrast to neural models, topical embeddings are interpretable. The topic can tell about itself addressing frequent words from the $p(w\,|\,t)$ distribution, or extracting topical phrases with automatic topic labeling [37] or summarization methods. Moreover, topical embedding $p(t\,|\,x)$ can tell about non-textual object x in words or phrases of natural language.

Thus, topic modeling is aimed not so much at predicting words in documents as revealing the thematic structure of a text collection, determining the semantics of documents and related objects, explaining topics in natural language.

1.4 Additive regularization

To solve the ill-posed problem of stochastic matrix factorization, we add regularization criterion $R(\Phi,\Theta)$ to the log-likelihood (1.4), under non-negativity and normalization constraints (1.5):

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{1.6}$$

Generally, several requirements may be imposed, each formalized by a regularizer $R_i(\Phi,\Theta)$, $i = 1,\dots,k$. The scalarization approach for multicriteria optimization leads to the *Additive Regularization for Topic Modeling* (ARTM), proposed in [54]:

$$R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta),$$

where non-negative regularization coefficients τ_i, $i = 1,\dots,k$, are hyperparameters of the learning algorithm.

Theorem 3 *Let the function $R(\Phi,\Theta)$ be continuously differentiable. Then the point (Φ,Θ) of the local extremum of the problem* (1.6), (1.5) *satisfies the system of equations with auxiliary variables $p_{tdw} = p(t\,|\,d,w)$, if zero columns of the matrices Φ, Θ are excluded from the solution:*

$$p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\varphi_{wt}\theta_{td}\bigr); \tag{1.7}$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{1.8}$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{1.9}$$

Proof can be found in [56], but it can be easier derived from the theorem 1. Let's rewrite (1.7) as follows:

$$p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\varphi_{wt}\theta_{td}\bigr) = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}} = \frac{\varphi_{wt}\theta_{td}}{p(w\,|\,d)}.$$

Let's apply the formula (1.1) to the function (1.6) and substitute the auxiliary variables p_{tdw} in the resulting expressions:

$$\begin{aligned}
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\frac{\partial L}{\partial\varphi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\sum_{d\in D}\frac{n_{dw}\theta_{td}}{p(w\,|\,d)} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right)\\
&= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right);\\
\theta_{td} &= \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial L}{\partial\theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in d}\frac{n_{dw}\varphi_{wt}}{p(w\,|\,d)} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)\\
&= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right).
\end{aligned}$$

Zero columns in the Φ and Θ matrices appear in those cases when the positive coordinate condition in the theorem 1 is not satisfied. Zero columns can be removed from the matrices, which is allowed by the condition of the theorem.

The theorem is proven.

A topic t is *degenerate* if $n_{wt} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}} \leqslant 0$ for all $w \in W$.

The degeneracy of the topic is a consequence of the excessively strong sparsing effect of the regularizer R. Zeroing the column of the matrix Φ means that the model prefers to abandon this topic. Reducing the number of topics can be a desirable side effect of regularization.

A document d is *degenerate* if $n_{td} + \theta_{td}\frac{\partial R}{\partial\theta_{td}} \leqslant 0$ for all $t \in T$.

The degeneracy of the document means that the model is not capable to describe it. May be, the document is too short or doesn't match the topical structure of the collection.

Learning a topic model is a numerical solution of the (1.7)–(1.9) system. The simple iteration method leads to the Expectation–Maximization (EM) algorithm, in which two steps are performed at each iteration: *E-step* (1.7) and *M-step* (1.8)–(1.9). With a rational implementation of this algorithm each iteration is performed in one linear pass through the collection. For each term w in each document d the topical embedding $p(t|d,w)$ is calculated by the E-step formula and is immediately used to update the counters n_{wt} and n_{td}.

Fast online algorithm, implemented in the `BigARTM` library [57], uses parallelization, splitting the collection into batches, controlling the update rate of Φ matrix, and a few more tricks to increase the computational speed [21, 3]. As a result, `BigARTM` outperforms other freely available topic modeling tools such as Gensim and Vowpal Wabbit by up to 20 times on some tasks [33].

Probabilistic Latent Semantic Analysis (PLSA) is historically the first probabilistic topic model [23]. In ARTM it corresponds to zero regularizer

$$R(\Phi,\Theta) = 0.$$

Latent Dirichlet Allocation (LDA) [7] is the first and most cited Bayesian model. It imposes restrictions on the columns of the Φ and Θ matrices in the form of Dirichlet prior distributions. In ARTM it corresponds to the cross-entropy regularizer [33]

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W}\beta_{wt}\ln\varphi_{wt} + \sum_{d\in D}\sum_{t\in T}\alpha_{td}\ln\theta_{td}. \tag{1.10}$$

If the hyperparameters β_{wt}, α_{td} are positive, then the regularization smoothes the conditional distributions φ_{wt}, θ_{td} bringing them closer to the given vectors $\text{norm}_w(\beta_{wt})$, $\text{norm}_t(\alpha_{td})$. If β_{wt}, α_{td} are negative, then the effect of the regularizer is sparsing instead of smoothing, as can be seen from the M-step formulas:

$$\varphi_{wt} = \underset{w\in W}{\text{norm}}\bigl(n_{wt} + \beta_{wt}\bigr); \qquad \theta_{td} = \underset{t\in T}{\text{norm}}\bigl(n_{td} + \alpha_{td}\bigr).$$

In the Bayesian interpretation, hyperparameters are bounded from below: $\beta_{wt} > -1$, $\alpha_{td} > -1$, due to the properties of the Dirichlet distribution. Therefore, sparsing effect is restricted and weak. There are no such restrictions in the ARTM interpretation, since a priori Dirichlet distributions are not introduced into the model.

1.5 Comparison with Bayesian learning

Let for generality X be the observed data set (e.g. the text documents collection), $p(X\,|\,\Omega)$ be a probabilistic data model with Ω parameters (e.g. the Φ and Θ matrices), $p(\Omega\,|\,\gamma)$ be *a priori distribution* of model parameters with hyperparameters γ (in the LDA model, the Dirichlet distributions with hyperparameters β_{wt}, α_{td}). Then the *posterior distribution* of Ω parameters is given by the Bayes' formula:

$$p(\Omega\,|\,X,\gamma) = \frac{p(\Omega, X\,|\,\gamma)}{p(X\,|\,\gamma)} \propto p(X\,|\,\Omega)\,p(\Omega\,|\,\gamma),$$

where the symbol $\propto$ means "equals up to normalization". Bayesian inference is useful in many data analysis problems where we do something with model parameters: testing statistical hypotheses, interval estimating, sampling, etc. However, in the practice of topic modeling Bayesian inference is performed only to get a point estimate of the Ω parameters:

$$\Omega := \arg\max_{\Omega} p(\Omega\,|\,X,\gamma).$$

Maximizing a posteriori (MAP) gives a point estimate for Ω, bypassing the intermediate step of the approximate and tedious posterior inference:

$$\Omega := \arg\max_{\Omega}\bigl(\ln p(X\,|\,\Omega) + \ln p(\Omega\,|\,\gamma)\bigr).$$

The logarithm of the prior distribution can be considered as a classical non-Bayesian regularization criterion $R(\Omega) = \ln p(\Omega \,|\, \gamma)$. In this form, it can be separated from a particular model and brought to another model.

Additive regularization generalizes log-priors to any regularizers, including those that do not have a probabilistic nature, as well as their linear combinations, without violating the convergence properties:

$$\Omega := \arg\max_{\Omega} \Bigl(\ln p(X \,|\, \Omega) + \sum_i \tau_i R_i(\Omega) \Bigr).$$

The main disadvantage of Bayesian inference is that it requires sophisticated calculations unique to each model, which makes it difficult to regularly combine multiple requirements and constraints. In Bayesian learning, there are no conventional regularization mechanisms based on criteria, since there is actually no optimization problem for Ω. Additional information can be introduced either through the prior distribution or through the very structure of the model. If the prior distributions are not Dirichlet distributions, then the inference becomes noticeably more complicated. Non-unified inference incur implementing and testing costs for each model.

The Dirichlet distribution plays a special role in Bayesian topic modeling. Although it has no convincing linguistic justification, most models are built on it in the literature. The reason is solely in the mathematical convenience of the Dirichlet prior conjugated with a multinomial distribution. In ARTM there is no reason to prefer the Dirichlet distribution to other regularizers.

The additivity of regularizers leads to a modular topic modeling technology, which is implemented in the `BigARTM` open source project [57]. In applications, composite models with desired properties can be built by adding ready-to-use regularizers from the library, without new mathematical calculations and coding. The development of such a technology within the Bayesian framework is hardly possible.

1.6 Overview of models and regularizers

Many topic models, originally formulated in the Bayesian paradigm can be reformulated in terms of classical non-Bayesian regularization [33].

Combination of smoothing, sparsing and decorrelation regularizers has proven itself well in practice in many studies [55, 56, 61]. Topic decorrelation regularizer

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws}$$

not only makes the topics more diverse, but also groups common words into separate background topics and purges all other topics from them [49].

Semi-supervised topic models use the smoothing regularizer of Φ matrix to set the seed words for some of the topics so that subject topics of interest can crystallize in their place during the iterative process. This technique has been used for searching rare topics in social media, such as symptoms, diseases, and their treatments [41, 42]; crime and extremism [36, 47]; ethnicities and interethnic relations [8, 34, 40]. For example, to search for a given number of ethno-relevant topics within the ARTM framework, smoothing regularization was applied using the vocabulary of ethnonyms. After that, the topic model was able to determine how topics are specialized by ethnicity [1, 2]. In particular, multi-ethnic topics were found, helping sociologists to identify the aspects of interethnic relations.

Multimodal topic model describes documents containing not only words, but also terms of other modalities: categories, authors, time, tags, entities, users, etc. Each modality $m \in M$ has its own dictionary of terms W^m, own matrix Φ^m with normalized columns, and own log-likelihood criterion. The problem is to maximize the weighted sum of these criteria over modalities:

$$\sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi^m_{wt} \theta_{td} + R\big(\{\Phi^m\}, \Theta\big) \;\to\; \max_{\Phi,\Theta}. \qquad (1.11)$$

Multimodal data helps to determine the document topics more accurately. Conversely, the topic model can be used to reveal the semantics of modalities or predict missing modality metadata.

Classification topic model is a special case of the multimodal PTM with the modality C of categories or classes. The model predicts class probabilities for a document $p(c \,|\, d)$ with a linear classifier using topic probabilities $p(t \,|\, d)$ as features:

$$p(c \,|\, d) = \sum_{t\in T} p(c \,|\, t) p(t \,|\, d) = \sum_{t\in T} \varphi_{ct} \theta_{td}.$$

Experiments showed that this topic model outperforms conventional multiclass classification methods on large text collections with a large number of unbalanced, overlapping, interdependent classes [46]. Similar results on the same collections were reproduced for the multimodal ARTM in [53]. *Unbalanced* classes can contain both a small and a very large number of documents. *Overlapping* classes means that a document may belong to many classes. *Interdependent* classes share terms and topics, therefore, they can compete and interfere when classifying a document.

Multilingual topic model is another case of multimodal PTM, when languages act as modalities. Linking parallel texts into a common document is enough for synchronizing topics across languages in cross-language document search tasks [58]. Regularizers based on bilingual dictionaries have been proposed in [18], however, the parallel texts linking remains the main contribution to the search quality.

Triple matrix topic model arises from the assumption that topics are generated not by a document, but by one of the modalities, for example, categories, authors, or tags. The author-topic model ATM [45], the tag weighted topic model TWTM [35], and the model for detecting behaviour dynamics in video [24] can be viewed as triple matrix factorization:

$$p(w \,|\, d) = \sum_{t\in T} p(w \,|\, t) \sum_{a\in A} p(t \,|\, a) p(a \,|\, d) = \sum_{t\in T} \varphi_{wt} \sum_{a\in A} \psi_{ta} \pi_{ad},$$

where A is a dictionary of authors, tags, or behaviours respectively. The EM-like algorithm given in [33] for this model can be easily obtained as a corollary of the maximization theorem on unit simplices.

Hierarchical topic models divide topics into smaller subtopics recursively. There is a wide variety of approaches and methods for learning and evaluating topical hierarchies [62]. The top-down level-wise strategy based on ARTM has been proposed in [15] and improved in [5]. The hierarchy is built from top to bottom, each child level having greater number of topics than the parent level has. Each level is a conventional flat topic model, which is linked with the parent level by conditional probabilities $\psi_{st} = p(s \,|\, t)$ of subtopics $s \in S$ in parent topics $t \in T$. The regularizer tries to approximate parent topics φ_{wt} by a probabilistic mixture of child topics φ_{ws} with coefficients ψ_{st}:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \varphi_{ws} \psi_{st}. \tag{1.12}$$

The maximization of $R(\Phi, \Psi)$ coincides up to notation with the main topic modeling task (1.4), with parent topics t considered as *pseudo-documents* with term frequencies $n_{wt} = n_t \varphi_{wt}$. This regularizer can be implemented by simply adding $|T|$ pseudo-documents to the collection before building each child level. The linking matrix Ψ is produced by the model in the columns of the Θ matrix corresponding to pseudo-documents.

Multimodal hierarchical topic models perform well in document-by-document topic-based search [25, 26]. Combining decorrelation, sparsing, and smoothing regularizers along with modalities of n-grams, authors and categories significantly improves search quality. In experiments with exploratory search in technology blogs, both precision and recall reach 90%. Optimal (in terms of search quality) dimension of topical embeddings at the third level of the hierarchy turned out to be several times higher than that of the flat model. This means that the gradual fragmentation of topics into smaller subtopics allows topical embeddings to keep more useful information about documents.

Topic model for mining polarized opinions is actually a two-level hierarchy, in which the upper level determines topics in news [20]. The second level is based on unusual modalities, dividing the topic into subtopics with polarized opinions about the topic. The modalities are: named entities with positive and negative sentiments, named entities with their semantic roles, triplets “subject, predicate, object”. Experiments have shown that each of the three modalities is important for improving the polarized opinions detection. A similar two-level hierarchy has been proposed in [43], where syntactic modalities were used at the child level to divide parent level topics into more detailed client intents in the collection of contact center dialogs.

Hyperparameter optimization strategies. Additive regularization loses to Bayesian modeling in only one aspect. The more regularizers are used, and the more regularization coefficients have to be selected, the more careful balancing they require. Early studies have shown that regularizers can interfere with each other, and that understanding their interactions leads to sequential strategies of adding regularizers to the model [55].

Adding regularizers during the iteration process in the order $\{\Phi$ decorrelation, Θ sparsing, Φ smoothing$\}$ has been proven to be a successful strategy for topic-based exploratory search [61, 25]. In further experiments, the hyperparameter space was extended with modality weights, pseudo-document weights and the number of topics at each level in the hierarchical model [26]. When regularizer starts from a given iteration, learning algorithm must be restarted from this point many times with hyperparameter values iterated over a coarse grid. The model quality is controlled visually by multiple criteria during the iterative process.

Later, this technique was extended and implemented in TopicNet open source library, which operates on top of `BigARTM` hiding technical details from the user [12]. The user specifies only the high-level regularization strategy. TopicNet automates computational experiments on hyperparameter optimization, providing logging and visualization.

A more general framework for hyperparameter optimization in ARTM is based on evolutionary algorithms and representation of a learning process as a multi-stage strategy for changing hyperparameters [31]. Later this approach was extended by a surrogate model for PTM evaluation, which reduced the time for automatic selection of hyperparameters [32].

1.7 Hypergraph topic models of transactional data

Topic models of text collections describe occurrences of words in documents. Multimodal topic models describe documents that may contain the terms of several modalities: words, tags, categories, authors, etc. In all these cases, the model describes pairwise interactions between documents and terms. In more complex applications, the initial data may describe transactions between three or more objects. For example, "user u clicked ad b on page s" in an advertising network; "user u wrote word w on blog page d" in a social network; "buyer b bought item g from seller s" in a sales network; "client u departed from airport x to airport y by airline a" in passenger air transportation; "user u rated the film f in a contextual situation s" in a recommender system. Another modality could be transaction time. In all of the examples above, a multi-object transaction can not be reduced to the pair interactions.

Transactional data can be represented by a hypergraph $\Gamma = \langle V, E \rangle$ defined by the set of term vertices V and the set of transaction edges E. Each edge e of E is a subset of two or more vertices, $e \subset V$. The task is to restore unknown topic distributions of vertices $p(t \mid v)$ from the observed dataset of transactions.

Each vertex has modality m from the set M. Denote by V_m the set of vertices having modality m. In conventional topic models, there are two modalities: terms $V_1 = W$ and documents $V_2 = D$; each edge transaction $e = (d, w)$ means that the term w occurs in the document d; thus, the hypergraph is a bipartite graph.

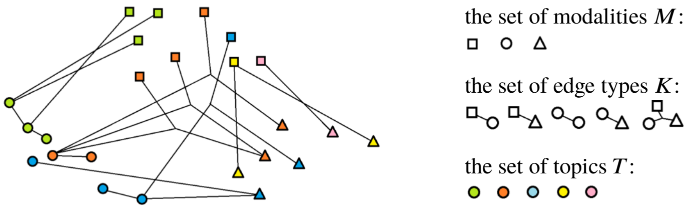

Fig. 1.1 An example of a hypergraph with vertices of three modalities, edges of five types, and five topics.

In more complicated applications, transactions can be of various types. For example, in the advertising network, along with triplet data "user u clicked ad b on page s", there may be pair data "user u visited page s", "page s contains term w", "ad b contains term w", "user's u query contains term w".

Let K be the set of transaction types. *Transactional data* of type k is a dataset of edges $E_k \subset E$. Each edge $e \in E_k$ occurs in the dataset n_{ke} times, having a latent topic $t \in T$. Figure 1.1 shows an example of a hypergraph.

Assume that each transaction $e \in E$ has one dedicated vertex d called *container*, and denote the edge by $e = (d, x)$, where x is the set of all other vertices of the edge. Similar to a document, a container has a distribution of topics $p(t \,|\, d)$. Denote the set of all containers by D.

We accept several hypotheses of conditional independence. Assume that neither the distribution of topics $p(t \,|\, d)$ in a container d, nor distributions of vertices in topics $p(v \,|\, t)$ depend on the type of the edge k. Next, suppose that the process of generating the edge $(d, x) \in E_k$ consists of two steps. First, a topic t is generated from the distribution $p(t \,|\, d)$. Then the set of vertices $x \subset V$ is generated so that each vertex $v \in x$ of the modality m is generated from the distribution $p(v \,|\, t)$ over the set V_m independently of the other edge vertices.

The topic model expresses the probabilities of hypergraph edge through conditional distributions associated with their vertices:

$$p(x \,|\, d) = \sum_{t \in T} p(t \,|\, d) \prod_{v \in x} p(v \,|\, t) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt}.$$

In matrix notation, the model parameters are matrices Θ and Φ_m, $m \in M$, as in the multimodal topic model (1.11).

Learning the hypergraph model is log-likelihoods maximization for all edge types k with weights τ_k, under the usual non-negativity and normalization constraints, improved by the regularizer $R(\Phi, \Theta)$:

$$\sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{kdx} \ln\biggl(\sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt}\biggr) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}; \tag{1.13}$$

$$\sum_{v \in V_m} \varphi_{vt} = 1,\ \varphi_{vt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1,\ \theta_{td} \geqslant 0.$$

Theorem 4 *Let the function $R(\Phi,\Theta)$ be continuously differentiable. Local maximum point (Φ,Θ) of the problem (1.13) satisfies the system of equations with respect to model parameters φ_{vt}, θ_{td} and auxiliary variables $p_{tdx} = p(t\,|\,d,x)$, if zero columns of the matrices Φ_m, Θ are excluded from the solution:*

$$p_{tdx} = \operatorname*{norm}_{t\in T}\Bigl(\theta_{td}\prod_{v\in x}\varphi_{vt}\Bigr); \tag{1.14}$$

$$\varphi_{vt} = \operatorname*{norm}_{v\in V_m}\Bigl(\sum_{k\in K}\sum_{dx\in E_k}[v\in x]\,\tau_k n_{kdx}p_{tdx} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Bigr); \tag{1.15}$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{k\in K}\sum_{dx\in E_k}\tau_k n_{kdx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Bigr); \tag{1.16}$$

Proof. Let us apply the theorem 1 on maximization on unit simplices, extracting the expression for the auxiliary variables p_{tdx} defined in (1.14):

$$\begin{aligned}
\varphi_{vt} &= \operatorname*{norm}_{v\in V_m}\Bigl(\varphi_{vt}\sum_{k\in K}\tau_k\sum_{dx\in E_k}n_{kdx}\frac{\theta_{td}}{p(x\,|\,d)}\frac{\partial}{\partial\varphi_{vt}}\prod_{u\in x}\varphi_{ut} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Bigr)\\
&= \operatorname*{norm}_{v\in V_m}\Bigl(\sum_{k\in K}\sum_{dx\in E_k}\tau_k n_{kdx}[v\in x]p_{tdx} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\Bigr);\\
\theta_{td} &= \operatorname*{norm}_{t\in T}\Bigl(\theta_{td}\sum_{k\in K}\tau_k\sum_{x\in d}n_{kdx}\frac{1}{p(x\,|\,d)}\prod_{v\in x}\varphi_{vt} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Bigr)\\
&= \operatorname*{norm}_{t\in T}\Bigl(\sum_{k\in K}\sum_{x\in d}\tau_k n_{kdx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Bigr).
\end{aligned}$$

The theorem is proven.

The hypergraph model is a broad generalization of conventional PTMs. Despite this, the derivation of the EM-algorithm out of the theorem 1 is no more difficult than in the conventional case. This algorithm is implemented in `BigARTM` project.

1.8 Hypergraph recommender topic models

Let U be a finite set of users, I be a finite set of items that users can take or prefer. The probabilistic topic model predicts user preferences:

$$p(i\,|\,u) = \sum_{t\in T}p(i\,|\,t)\,p(t\,|\,u).$$

This model is equivalent to the topic model of a text collection, up to terminology: documents → users, terms → items, topics → interests. Following the analogy, the bag-of-words transforms into the bag-of-transactions hypothesis. In this case,

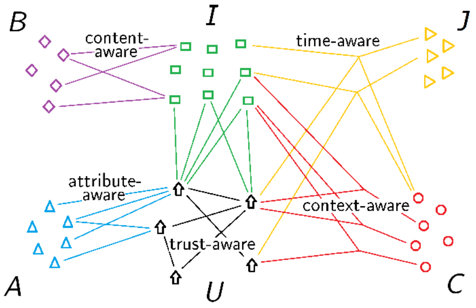

Fig. 1.2 Types of transactions between six modalities in a recommender system: users U, items I, user attributes A, item properties B, contextual situations C, time intervals J.

dataset can be considered as n_{ui} counters of the user u transactions with the item i. Depending on the application, transactions may be purchases, visits, likes, etc.

There is a well-known "cold start" issue in recommender systems. Nothing to recommend to a new user, since there is no history of his preferences. Nobody to recommend a new item, since no one has chosen it yet. To solve this problem, additional data about users and items can be involved. In particular, these may be data n_{ua} on the user attributes $a \in A$ or data n_{ib} on the item properties $b \in B$. If items have text descriptions, then B is a dictionary of terms used in these descriptions. Such recommender systems are called, respectively, attribute-aware and content-aware.

Users' advice to each other can also be used as additional data. These are pairwise interactions between users $n_{uu'}$ or trust-aware data.

User preferences may change over time or depend on the situation. To take into account such information, two more modalities are introduced: the set of situations C and the set of time intervals J. Interactions between them are described by transactions of three or more terms, for example, n_{uic} for "user u selected item i in situation c", or n_{uicj} for "user u selected item i in situation c in time interval j. Such systems are called, respectively, context-aware and time-aware.

Many types of $***$-aware models were introduced separately in the literature [14]. The hypergraph model can combine them all and learn topical embeddings for any interacting terms regardless their nature, fig. 1.2.

The recommender system data is different from the text collections as it has no natural analogue of a document or container. The set of transactions (u,i) may increase with time for both the user u and the item i, unlike unchanging documents.

Assume that the edges of the hypergraph $x \subset V$ do not contain container vertex. The edge generative process first generates a topic t from the distribution $\pi_t = p(t)$ which is common to the entire collection. Then the vertices $v \in x$ are generated independently of each other from distributions $\varphi_{vt} = p(v \,|\, t)$ over modalities V_m:

$$p(x) = \sum_{t\in T} p(t) \prod_{v\in x} p(v \,|\, t) = \sum_{t\in T} \pi_t \prod_{v\in x} \varphi_{vt}.$$

Topic models, in which documents act as one of the modalities, are called symmetric [52]. As before, maximization problem is for regularized log-likelihood under normalization and non-negativity constraints:

$$\sum_{k\in K}\tau_k\sum_{x\in E_k}n_{kx}\ln\biggl(\sum_{t\in T}\pi_t\prod_{v\in x}\varphi_{vt}\biggr)\ +\ R(\Phi,\pi)\to\max_{\Phi,\pi};\tag{1.17}$$

$$\sum_{v\in V_m}\varphi_{vt}=1,\ \varphi_{vt}\geqslant 0;\qquad \sum_{t\in T}\pi_t=1,\ \pi_t\geqslant 0.$$

Theorem 5 *Let the function $R(\Phi,\pi)$ be continuously differentiable. Local maximum point (Φ,π) of the problem (1.17) satisfies the system of equations with respect to model parameters φ_{vt}, π_t and auxiliary variables $p_{tx}=p(t\,|\,x)$, if zero columns of the Φ_m matrices are excluded from the solution:*

$$p_{tx}=\operatorname*{norm}_{t\in T}\biggl(\pi_t\prod_{v\in x}\varphi_{vt}\biggr).\tag{1.18}$$

$$\varphi_{vt}=\operatorname*{norm}_{v\in V_m}\biggl(\sum_{k\in K}\sum_{x\in E_k}[v\in x]\,\tau_k n_{kx}p_{tx}+\varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\biggr);\tag{1.19}$$

$$\pi_t=\operatorname*{norm}_{t\in T}\biggl(\sum_{k\in K}\sum_{x\in E_k}\tau_k n_{kx}p_{tx}+\pi_t\frac{\partial R}{\partial\pi_t}\biggr).\tag{1.20}$$

Proof follows straightforwardly from the maximization theorem on unit simplices, as in the case of the previous theorem.

In `BigARTM` the symmetrized model is not implemented, but it is not difficult to simulate it. To to this, the collection is split in some way into documents (for example, by transaction time), then a strong regularizer is introduced for smoothing the columns of the Θ matrix towards the (n_t) vector summed over all documents.

1.9 Sequential text topic models

The bag-of-words hypothesis is one of the most criticized assumptions in topic modeling. Many approaches was proposed in the literature in order to go beyond the bag-of-words restrictive assumption, either completely or partially.

Topic models with n-grams exploit the fact that stable combinations of n consecutive words often, though not always, represent subject domain terms or names. The n-grams may tell much more about topics than the same words treated independently. Topics built on the n-gram dictionary are better interpretable, than those built on unigrams [60, 29]. There are two approaches to using n-grams in topic modeling. In the first one, the dictionary of n-grams is built at the stage of text preprocessing using automatic extraction of terms, keywords, or collocations [19]. Then, the n-gram dictionary is used as a modality. The second approach is more complicated, in which topic modeling is combined with n-gram extraction [59, 60]. Concentration of distribution $p(t \,|\, w)$ in one or more topics is usually a strong indication that the n-gram w is a subject domain term.

Word network topic model predicts the appearance of a word nearby to another word, instead of predicting it in the document. "Nearby" means, say, no more than 10 words away or in one sentence. Define for each word $u \in W$ a pseudo-document d_u consisting of all words that occur nearby to the word u throughout the collection. Denote by n_{uw} the number of occurrences of the word w in a pseudo-document d_u.

The word network topic model WNTM [65] and the earlier word topic model WTM [13] predict a word in the neighborhood of other word:

$$p(w \,|\, u) = \sum_{t \in T} p(w \,|\, t)\, p(t \,|\, d_u) = \sum_{t \in T} \varphi_{wt} \theta_{tu}.$$

The log-likelihood can be used either as a regularizer for other topic model, or as the main learning criterion. In the first case, topic model is learned by the document collection augmented by pseudo-documents. In the second case, only pseudo-documents are used:

$$\sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} \varphi_{wt} \theta_{tu} \;\to\; \max_{\Phi,\Theta}.$$

According to *the distributional hypothesis* the meaning of a word is determined by the distribution of all words, in whose environment it occurs [22]. Words found in similar contexts have similar semantics, and in the model they should receive similar embeddings. Word embeddings implemented in the `word2vec` program [38, 39] are also learned from word co-occurrence data. They encapsulate the meanings of words so well that paired associations turn into vector equalities:

$$\text{king} - \text{queen} = \text{man} - \text{woman};$$
$$\text{Moscow} - \text{Beijing} = \text{Russia} - \text{China}.$$

The additively regularized WNTM also has this property [44], unlike conventional topic models. Moreover, topical embeddings are coordinate-wise interpretable, unlike word2vec and neural embeddings.

Sentence topic model can be considered as a special case of hypergraph topic model. Vertices of the hypergraph are words, edges are sentences. This approach is equivalent to the sentence topic model senLDA [4] and Twitter-LDA short message model [64] first proposed in terms of Bayesian learning. The hypergraph representation gives a lot of freedom in defining edges. These can be not only sentences, but also noun phrases, syntagmas, lexical chains, and in general any group of words, with reasonable assumption that they are generated by a common topic.

E-step regularization. The idea behind using intradocument word order data is to impose regularization constraints on topical embeddings $p_{tdw} = p(t \,|\, d, w)$. They specialize topical embeddings $p(t \,|\, w)$ from the global context of the collection to the narrower document context. Further narrowing of the context to the local neighborhoods of words requires processing the document as a sequence of word embeddings.

Define the regularizer $R(\Pi, \Phi, \Theta)$ as a function of the matrices Φ, Θ and a three-dimensional matrix of auxiliary variables $\Pi = (p_{tdw})_{T\times D\times W}$. According to (1.7), the elements of Π matrix are functions of Φ and Θ matrices. Therefore, the regularizer has a form $\tilde{R}(\Phi, \Theta) = R(\Pi(\Phi, \Theta), \Phi, \Theta)$. Then, theorem 3 can be applied to it. However, it is more convenient to write the system of equations in terms of the partial derivatives of the regularizer R rather than $\tilde{R}$.

Consider the problem of the regularized log-likelihood maximization under non-negativity and normalization constraints (1.5):

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Pi(\Phi, \Theta), \Phi, \Theta) \;\to\; \max_{\Phi,\Theta}. \qquad (1.21)$$

Theorem 6 *Let the function* $R(\Pi, \Phi, \Theta)$ *be continuously differentiable and does not depend on* p_{tdw} *for all* $w \notin d$. *Then the point* (Φ, Θ) *of the local extremum of the problem* (1.21), (1.5) *satisfies the system of equations with auxiliary variables* p_{tdw} *and* $\tilde{p}_{tdw}$, *if zero columns of the matrices* Φ, Θ *are excluded from the solution:*

$$p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\varphi_{wt}\theta_{td}\bigr);$$

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right); \qquad (1.22)$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad (1.23)$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \qquad (1.24)$$

Proof. First, we define the function $p_{zdw}(\Phi, \Theta) = \frac{\varphi_{wz}\theta_{zd}}{\sum_t \varphi_{wt}\theta_{td}}$ and find its partial derivatives. For any $t, z \in T$

$$\begin{aligned}
\varphi_{wt}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} &= \varphi_{wt}\frac{[z{=}t]\theta_{td}\sum_u \varphi_{wu}\theta_{ud} - \theta_{td}\varphi_{wz}\theta_{zd}}{(\sum_u \varphi_{wu}\theta_{ud})^2} \\
&= p_{tdw}[z{=}t] - p_{tdw}p_{zdw}; \qquad (1.25)\\
\theta_{td}\frac{\partial p_{zdw}}{\partial \varphi_{td}} &= \theta_{td}\frac{[z{=}t]\varphi_{wt}\sum_u \varphi_{wu}\theta_{ud} - \varphi_{wt}\varphi_{wz}\theta_{zd}}{(\sum_u \varphi_{wu}\theta_{ud})^2} \\
&= p_{tdw}[z{=}t] - p_{tdw}p_{zdw}; \qquad (1.26)
\end{aligned}$$

Note that the resulting expressions (1.25) and (1.26) are the same.

Let us introduce an auxiliary function Q of the variables Π, Φ, Θ:

$$Q_{tdw}(\Pi, \Phi, \Theta) = \frac{\partial R(\Pi, \Phi, \Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi, \Phi, \Theta)}{\partial p_{zdw}}.$$

Let us differentiate the superposition $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$, given that $\partial p_{zdw'}/\partial\varphi_{wt} = 0$ if $w \neq w'$; $\partial p_{zd'w}/\partial\theta_{td} = 0$ if $d \neq d'$; $\partial R/\partial p_{tdw} = 0$ if $w \notin d$:

$$\varphi_{wt}\frac{\partial\tilde{R}}{\partial\varphi_{wt}} = \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}} + \sum_{d\in D}\varphi_{wt}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\varphi_{wt}}; \tag{1.27}$$

$$\theta_{td}\frac{\partial\tilde{R}}{\partial\theta_{td}} = \theta_{td}\frac{\partial R}{\partial\theta_{td}} + \sum_{w\in d}\theta_{td}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\theta_{td}}. \tag{1.28}$$

Using (1.25) and (1.26), we get the identity

$$\varphi_{wt}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\varphi_{wt}} = \theta_{td}\sum_{z\in T}\frac{\partial R}{\partial p_{zdw}}\frac{\partial p_{zdw}}{\partial\theta_{td}} = p_{tdw}Q_{tdw}.$$

Let us substitute the resulting expressions into (1.27) and (1.28), which we then substitute into the system of equations from the theorem 3:

$$p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\varphi_{wt}\theta_{td}\bigr);$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(\sum_{d\in D}n_{dw}p_{tdw} + \sum_{d\in D}Q_{tdw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \tag{1.29}$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\sum_{w\in d}n_{dw}p_{tdw} + \sum_{w\in d}Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right). \tag{1.30}$$

Substituting of the auxiliary variable $\tilde{p}_{tdw}$ according (1.22) allows us to rewrite the equations (1.29)–(1.30) in the required form (1.23)–(1.24).

The theorem is proven.

In the EM-algorithm, topical embeddings $p_{tdw} = p(t|d,w)$ are calculated for each word w in the document d. Then they are transformed into new vectors $\tilde{p}_{tdw}$ and used at the M-step instead of p_{tdw}. We call this technique *E-step regularization* or *E-step post-processing*. This is an optional procedure, its the presence or absence does not affect the implementation of other computations in any way.

Moreover, the post-processing formula does not necessarily need to be derived from the regularization criterion. You can do the opposite: transform sequence of topical embeddings using a heuristic post-processing, for example, smoothing, sparsing, or segmentation. In fact, this will correspond to a regularization under some criterion $R(\Pi)$, which is not obligatory to be written out explicitly.

This approach was used in [48] to improve the quality of topical segmentation of documents.

One-pass topic modeling. In the EM-algorithm, the computation of document topical embedding $\theta_d = (\theta_{td})_{t\in T}$ requires many iterations over the document. Nevertheless, θ_d can be calculated in a one linear pass through the document [28]. The explicit formula $\theta_{td}(\Phi)$ follows from the M-step equation or from the total probability formula, where the distribution $p(t)$ is assumed to be fixed:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(t|w)\,p(w|d) = \sum_{w\in d} \frac{n_{dw}}{n_d} \operatorname*{norm}_{t\in T}\bigl(\varphi_{wt}p(t)\bigr).$$

Although formally this equality constraint is not an optimization criterion, in fact it plays the role of a regularizer and can be used in combination with other regularizers within the ARTM framework.

Theorem 7 *Let the functions $\theta_{td}(\Phi)$ and $R(\Phi,\Theta)$ be continuously differentiable. Then the point Φ of the local extremum of the problem* (1.6), (1.5) *with equality constraints $\theta_{td} = \theta_{td}(\Phi)$ satisfies the system of equations with auxiliary variables $p_{tdw} = p(t\,|\,d,w)$, n_{td}, and p'_{tdw}, if zero columns of the matrices Φ, Θ are excluded from the solution:*

$$\begin{aligned}
p_{tdw} &= \operatorname*{norm}_{t\in T}\bigl(\varphi_{wt}\theta_{td}\bigr);\\
n_{td} &= \sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}};\\
p'_{tdw} &= p_{tdw} + \frac{\varphi_{wt}}{n_{dw}}\sum_{s\in T}\frac{n_{sd}}{\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\varphi_{wt}};\\
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\biggl(\sum_{d\in D} n_{dw}p'_{tdw} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\biggr).
\end{aligned}$$

Like the E-step post-processing, modification of the EM-algorithm leads to the transformation of the topical embeddings p_{tdw} into p'_{tdw}, which are substituted into the usual M-step equation for the Φ matrix, without affecting the implementation of the remaining computations.

Experiments on three text collections [28] have shown that the one-pass algorithm is not only much faster but also improves the model in terms of sparseness, difference, logLift and coherence topic quality measures. The `BigARTM` and TopicNet libraries were used for the experiments.

The one-pass topic modeling opens up possibilities for fast computation of local contextual topical embeddings and processing of text as a sequence of words beyond the bag-of-words restrictive assumption.

1.10 Discussion and conclusions

Hundreds of Bayesian topic models described in thousands of papers over the past two decades, can be reformulated in terms of classical non-Bayesian regularization. After this, they can be inferred easily, literally by one line of calculations out of the theorem on the maximization of a smooth function on unit simplices. One may wonder why this opportunity has not been noticed over so long time, especially given that Bayesian inference is laborious and unique to each model, which brings many technical inconveniences to researchers.

Many areas of data analysis and machine learning including image and signal processing are being developed according to the same general scenario. First, the formal model and the optimization problem are stated; then various specific structures, auxiliary criteria and regularizers are added; and finally, the transition to Bayesian regularization takes place. This transition usually occurs when there is a practical need for evaluation not only the model parameters themselves, but also their posterior distributions.

In Probabilistic Topic Modeling, the typical development scenario was violated and the community moved to Bayesian learning skipping the natural stage of development within the classical regularization. The very paradox is that in the practice of topic modeling, posterior distributions are used only for maximum likelihood point estimation.

Additive regularization (ARTM) is an attempt to fill the gap, though it might be late as the focus of community interest has already shifted to deep neural networks, attention models, and transformer architectures. Topic modeling is now focused more on the fusion with neural networks in search of opportunities to combine the best of two worlds [63].

Both worlds of models, neural-based and topic-based, generate vector representations of words and texts.

Both worlds tend to models homogenization [9], that is, to have a unified vector space that embeds any heterogeneous objects of any nature based on data about their interactions. It was demonstrated above how the hypergraph topic models implement this idea.

Both worlds of models can generate global and local embeddings. It has been shown above how the topic models can process a sequential text. The neural network models are much more complicated, their embeddings are able to absorb all the information about the connections between words, but it is out of our understanding which connections and how exactly are taken into account. Topic models are much simpler, their embeddings take into account only the lexical co-occurrence of words, while retaining interpretability. The coordinate-wise interpretability is a direct consequence of the fact that topic embeddings are non-negative normalized vectors on a unit simplex.

Avoiding the Bayesian inference makes topic models closer to neural models, thus making their deeper integration possible. As soon as non-negativity and normalization constraints are imposed, any vector parameter of a neural network can be learned with the use of the multiplicative gradient steps from the theorem of maximization on unit simplices. These are the promising opportunities for future research.

This work was supported by the Russian Foundation for Basic Research (project no. 20-07-00936).

References

[1] Apishev M, Koltcov S, Koltsova O, Nikolenko S, Vorontsov K (2016) Additive regularization for topic modeling in sociological studies of user-generated text content. In: MICAI 2016, 15th Mexican International Conference on Artificial Intelligence, Springer, Lecture Notes in Artificial Intelligence, vol 10061, pp 166–181

[2] Apishev M, Koltcov S, Koltsova O, Nikolenko S, Vorontsov K (2016) Mining ethnic content online with additively regularized topic models. Computacion y Sistemas 20(3):387–403

[3] Apishev MA, Vorontsov KV (2020) Learning topic models with arbitrary loss. In: Proceeding of the 26th Conference of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association, pp 30–37

[4] Balikas G, Amini M, Clausel M (2016) On a topic model for sentences. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '16, pp 921–924

[5] Belyy AV, Seleznova MS, Sholokhov AK, Vorontsov KV (2018) Quality evaluation and improvement for hierarchical topic modeling. In: Computational Linguistics and Intellectual Technologies. Dialogue 2018, pp 110–123

[6] Blei DM (2012) Probabilistic topic models. Communications of the ACM 55(4):77–84

[7] Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022

[8] Bodrunova S, Koltsov S, Koltsova O, Nikolenko SI, Shimorina A (2013) Interval semi-supervised LDA: Classifying needles in a haystack. In: Espinoza FC, Gelbukh AF, Gonzalez-Mendoza M (eds) MICAI (1), Springer, Lecture Notes in Computer Science, vol 8265, pp 265–274

[9] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji NS, Chen AS, Creel KA, Davis J, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie LE, Goel K, Goodman ND, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard TF, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Koh PW, Krass MS, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani SP, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko JF, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani YH, Ruiz C, Ryan J, R'e C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan KP, Tamkin A, Taori R, (2021) On the opportunities and risks of foundation models. CoRR abs/2108.07258, URL https://crfm.stanford.edu/assets/report.pdf

[10] Boyd-Graber J, Hu Y, Mimno D (2017) Applications of topic models. Foundations and Trends® in Information Retrieval 11(2-3):143–296

[11] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 1877–1901

[12] Bulatov V, Egorov E, Veselova E, Polyudova D, Alekseev V, Goncharov A, Vorontsov K (2020) TopicNet: Making additive regularisation for topic modelling accessible. In: Proceedings of The 12th Conference on Language Resources and Evaluation (LREC 2020), pp 6745–6752

[13] Chen B (2009) Word topic models for spoken document retrieval and transcription 8(1):2:1–2:27

[14] Chen R, Hua Q, Chang YS, Wang B, Zhang L, Kong X (2018) A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. IEEE Access 6:64301–64320, DOI 10.1109/ACCESS.2018.2877208

[15] Chirkova NA, Vorontsov KV (2016) Additive regularization for hierarchical multimodal topic modeling. Journal Machine Learning and Data Analysis 2(2):187–200

[16] Churchill R, Singh L (2022) The evolution of topic modeling. ACM Comput Surv 54(10s)

[17] Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186

[18] Dudarenko MA (2015) Regularization of multilingual topic models. Vychisl Metody Programm (Numerical methods and programming) 16:26–38

[19] El-Kishky A, Song Y, Wang C, Voss CR, Han J (2014) Scalable topical phrase mining from text corpora. Proc VLDB Endowment 8(3):305–316

[20] Feldman DG, Sadekova TR, Vorontsov KV (2020) Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. In: Computational Linguistics and Intellectual Technologies. Dialogue 2020, pp 268–283

[21] Frei O, Apishev M (2016) Parallel non-blocking deterministic algorithm for online topic modeling. In: AIST'2016, Analysis of Images, Social networks and Texts, Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), vol 661, pp 132–144

[22] Harris Z (1954) Distributional structure. Word 10(23):146–162

[23] Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, pp 50–57

[24] Hospedales T, Gong S, Xiang T (2012) Video behaviour mining using a dynamic topic model. International Journal of Computer Vision 98(3):303–323

[25] Ianina A, Vorontsov K (2019) Regularized multimodal hierarchical topic model for document-by-document exploratory search. In: Balandin S, Niemi V, Tutina T (eds) Proceeding Of The 25th Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 5–8, 2019., pp 131–138

[26] Ianina AO, Vorontsov KV (2020) Hierarchical interpretable topical embeddings for exploratory search and real-time document tracking. International Journal of Embedded and Real-Time Communication Systems (IJERTCS) 11(4)

[27] Irkhin IA, Vorontsov KV (2020) Convergence of the algorithm of additive regularization of topic models. Trudy Instituta Matematiki i Mekhaniki UrO RAN 26(3):56–68

[28] Irkhin IA, Bulatov VG, Vorontsov KV (2020) Additive regularization of topic models with fast text vectorization. Computer Research and Modeling 12(6):1515–1528

[29] Jameel S, Lam W (2013) An N-gram topic model for time-stamped documents. In: 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013, Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, pp 292–304

[30] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications 78(11):15169–15211

[31] Khodorchenko M, Teryoshkin S, Sokhin T, Butakov N (2020) Optimization of learning strategies for artm-based topic models. In: de la Cal EA, Villar Flecha JR, Quintián H, Corchado E (eds) Hybrid Artificial Intelligent Systems, Springer International Publishing, pp 284–296

[32] Khodorchenko M, Butakov N, Sokhin T, Teryoshkin S (2022) Surrogate-based optimization of learning strategies for additively regularized topic models. Logic Journal of the IGPL DOI 10.1093/jigpal/jzac019, URL https://doi.org/10.1093/jigpal/jzac019, jzac019, https://academic.oup.com/jigpal/advance-article-pdf/doi/10.1093/jigpal/jzac019/43022305/jzac019.pdf

[33] Kochedykov DA, Apishev MA, Golitsyn LV, Vorontsov KV (2017) Fast and modular regularized topic modelling. In: Proceeding of the 21st Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017, IEEE, pp 182–193

[34] Koltcov S, Koltsova O, Nikolenko S (2014) Latent Dirichlet allocation: Stability and applications to studies of user-generated content. In: Proceedings

of the 2014 ACM Conference on Web Science, ACM, New York, NY, USA, WebSci'14, pp 161–165

[35] Li S, Li J, Pan R (2013) Tag-weighted topic model for mining semi-structured documents. In: IJCAI'13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press, pp 2855–2861

[36] M A Basher AR, Fung BCM (2014) Analyzing topics and authors in chat logs for crime investigation. Knowledge and Information Systems 39(2):351–381

[37] Mei Q, Shen X, Zhai C (2007) Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, pp 490–499

[38] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. CoRR abs/1301.3781

[39] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. CoRR abs/1310.4546

[40] Nikolenko SI, Koltcov S, Koltsova O (2017) Topic modelling for qualitative studies. Journal of Information Science 43(1):88–102

[41] Paul MJ, Dredze M (2013) Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9–14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, pp 168–178

[42] Paul MJ, Dredze M (2014) Discovering health topics in social media using topic models. PLoS ONE 9(8)

[43] Popov A, Bulatov V, Polyudova D, Veselova E (2019) Unsupervised dialogue intent detection via hierarchical topic model. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, pp 932–938

[44] Potapenko A, Popov A, Vorontsov K (2017) Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. In: Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017, Springer, Cham, pp 167–180

[45] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, Arlington, Virginia, United States, UAI '04, pp 487–494

[46] Rubin TN, Chambers A, Smyth P, Steyvers M (2012) Statistical topic models for multi-label document classification. Machine Learning 88(1-2):157–208

[47] Sharma A, Pawar DM (2015) Survey paper on topic modeling techniques to gain usefull forcasting information on violant extremist activities over cyber space. International Journal of Advanced Research in Computer Science and Software Engineering 5(12):429–436

[48] Skachkov NA, Vorontsov KV (2018) Improving topic models with segmental structure of texts. In: Computational Linguistics and Intellectual Technologies. Dialogue 2018, pp 652–661

[49] Tan Y, Ou Z (2010) Topic-weak-correlated latent Dirichlet allocation. In: 7th International Symposium Chinese Spoken Language Processing (ISCSLP), pp 224–228

[50] Tikhonov AN, Arsenin VY (1977) Solution of ill-posed problems. W. H. Winston, Washington, DC

[51] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp 5998–6008

[52] Vinokourov A, Girolami M (2000) A probabilistic hierarchical clustering method for organising collections of text documents. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol 2, pp 182–185 vol.2, DOI 10.1109/ICPR.2000.906043

[53] Vorontsov K, Frei O, Apishev M, Romov P, Suvorova M, Yanina A (2015) Non-bayesian additive regularization for multimodal topic modeling of large collections. In: Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, ACM, New York, NY, USA, pp 29–37

[54] Vorontsov KV (2014) Additive regularization for topic models of text collections. Doklady Mathematics 89(3):301–304

[55] Vorontsov KV, Potapenko AA (2014) Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: AIST'2014, Analysis of Images, Social networks and Texts, Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), vol 436, pp 29–46

[56] Vorontsov KV, Potapenko AA (2015) Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications 101(1):303–323

[57] Vorontsov KV, Frei OI, Apishev MA, Romov PA, Suvorova MA (2015) BigARTM: Open source library for regularized multimodal topic modeling of large collections. In: AIST'2015, Analysis of Images, Social networks and Texts, Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), pp 370–384

[58] Vulic I, De Smet W, Tang J, Moens MF (2015) Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. Information Processing & Management 51(1):111–147

[59] Wallach HM (2006) Topic modeling: Beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, NY, USA, ICML '06, pp 977–984

[60] Wang X, McCallum A, Wei X (2007) Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, pp 697–702

[61] Yanina A, Golitsyn L, Vorontsov K (2018) Multi-objective topic modeling for exploratory search in tech news. In: Filchenkov A, Pivovarova L, Žižka J (eds) Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017, Springer International Publishing, Cham, pp 181–193

[62] Zavitsanos E, Paliouras G, Vouros GA (2011) Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes. Journal of Machine Learning Research 12:2749–2775

[63] Zhao H, Phung D, Huynh V, Jin Y, Du L, Buntine W (2021) Topic modelling meets deep neural networks: A survey. In: Zhou ZH (ed) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, pp 4713–4720

[64] Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing Twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'11, pp 338–349

[65] Zuo Y, Zhao J, Xu K (2016) Word network topic model: A simple but general solution for short and imbalanced texts. Knowledge and Information Systems 48(2):379–398