Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 00 (2017) 1–16
PATTERN RECOGNITION

Efficient Algorithms for Graph Regularized PLSA

Xin Wang[a,b], Ming-Ching Chang[b], Siwei Lyu[b,*]

[a] *CuraCloud Corporation, Seattle, WA, 98104, USA*
[b] *Department of Computer Science, University at Albany, SUNY, NY 12222, USA*

Abstract

Probabilistic Latent Semantic Analysis (PLSA) is a popular data analysis method with the objective to discover the underlying semantic structure of input data. In this work, we describe a method for probabilistic topic analysis in image and text based on a new representation of graph-regularized PLSA (GPLSA). In GPLSA, data entities are mapped to an undirected graph, where similarities between topic compositions on the graph are measured by the divergence between discrete probabilities. Such divergence is essentially incorporated as a graph-regularizer that augments the original PLSA algorithm. Furthermore, we extend the GPLSA algorithms to multiple data modalities based on the connections between data entities of each modality. We propose efficient multiplicative iterative algorithms for GPLSA with three popular regularizers, namely ℓ_1, ℓ_2 and symmetric KL divergences. In each case, we derive simple efficient numerical solutions that require only matrix arithmetic operations during the optimization. The performance of proposed algorithms was evaluated on three applications: (1) image clustering as a single-modality topic analysis, (2) the text/image cross-modal retrieval, and (3) multi-lingual topic analysis as a multi-modality topic analysis. Experimental results demonstrate the efficacy of GPLSA over state-of-the-art methods in these applications.

Keywords: Probabilistic Latent Semantic Analysis, Graph Regularization, Topic Analysis, Clustering, Cross-modal Information Retrieval

1. Introduction

Probabilistic topic modeling has been studied extensively in recent years which aims to discover hidden thematic structures in large archival documents and to annotate a large document corpus with thematic information. Typically, a map is established between the high-dimensional word distribution vectors of the documents and the lower-dimensional topic vectors, where the semantic properties of these words and documents can be expressed in terms of probabilistic topic models. Basic probabilistic topic models include the well-known *probabilistic latent semantic analysis* (PLSA) [1] and *latent Dirichlet allocation* (LDA) [2]. Topic analysis has wide range of applications including activity detection [3], image analysis [4], pattern recognition [5, 6, 7, 8], natural scene categories [9], video processing [10, 11, 12], information retrieval [13], and co-authorship network analysis [14, 15, 16].

Graph is widely used in various domains for its representative capabilities of relations among entities, and thus the use of graph representation in topic modeling is becoming popular. For instance, social networks (such as the Facebook) operate based on a huge set of user profiles and friendship connections; publication archives such as

*Corresponding author
Email addresses: xinw@curacloudcorp.com (Xin Wang), mchang2@albany.edu (Ming-Ching Chang), lsw@cs.albany.edu (Siwei Lyu)

the Digital Bibliography and Library Project (DBLP) contain a vast authorship network. Several recent works have considered the integration of a graph structure with topic modeling, such that the topics of interest can be obtained using a *regularizer* defined along with the graph structure. The work in [17] combines topic modeling and social network to analyze the topics in an co-authorship network. On a related front, graph regularizer is exploited to mode the intrinsic structure of data distributions in *graph-regularized non-negative matrix factorization* (GNMF) [18], where encouraging results are obtained in document/image clustering. We will show later that GNMF can be regarded as a graph-regularized topic modeling problem, due to the close relationship between the PLSA and NMF algorithms [19, 20].

In this work, we proposed a set of efficient algorithms for the *graph-regularized probabilistic latent semantic analysis* (GPLSA), which is a general approach for topic analysis. GPLSA is capable of handling both single- or multiple- modalities of data by casting a graph structure into the PLSA topic modeling. We will show that in our general formulation of GPLSA, graph regularizers are essentially the divergences between the discrete probability distributions corresponding to the composition of topics from each data entry. The data entries are defined on a graph, where entities can be projected onto a lower dimensional semantic space. Our formulation enables entities sharing the same semantics to smooth out their effects with each other. We propose efficient algorithms for the learning of topics and their compositions using three widely used divergence definitions, namely, ℓ_1, ℓ_2 and symmetric KL divergences. GPLSA optimization is casted as the minimization of the underlying divergences, which encourages similarities between topic compositions of each data entry and its nearest neighbors on the graph. Our algorithm is efficient because the optimization steps consist of only simple matrix operations and the derivation of numerical solutions of scalar nonlinear equations. Our GPLSA algorithms also afford theoretical guarantee of convergence, unlike other state-of-the-art works [21, 22]. We apply the GPLSA algorithms in image clustering, cross-modal retrieval and multi-lingual topic analysis applications on public benchmarks for evaluation and comparison. Experimental results show noticeable improvements of GPLSA against other state-of-the-art methods.

Main contributions of this work are summarized in the following:

(1) We propose efficient algorithms for graph-regularized PLSA (GPLSA) as a general framework for single- or multi- modality topic analysis, where the graph regularizer is based on the divergence between discrete probability distributions. Similarities between topics are enforced in a joint latent space constraint by the graph, and topic distributions are enhanced by their nearest neighbors on the graph.

(2) We show improved results for the ℓ_1 regularizer over the baseline, completing the study of GNMF [18]. For ℓ_2 divergence as the regularizer, our GPLSA algorithm is more efficient and with a convergence guarantee. We further describe a new algorithm using symmetric KL divergences as the regularizer, and demonstrate that it is more effective compared to the ℓ_2 divergence.

(3) The proposed solution extends naturally from topic analysis of a single modality to multiple modalities. Our method enables capturing of similarities between documents across modalities, by learning a joint latent space for documents of different modalities. Our topic learning representation leverages the compatible yet complementary conceptual themes among each modality. Thus it is more effective than other methods relying on features derived from direct concatenation of modalities.

An early version of this work focusing on multi-modal Learning was published in [23, 24]. This paper improves our previous work in the following aspects. (1) We derive the efficient algorithms for the GPLSA problem in a general framework based on the extended works. We also provide in depth motivations and full technical details. In this regard, our previous work of [23] can be considered as a special case of our general formulation in this paper. (2) We provide simple efficient numerical solutions that require only matrix arithmetic operations for the optimization. We also provide a proof of convergence for the general form of the algorithm. (3) We provide additional diagnostic experiments regarding the clustering performance and multi-lingual topic analysis to demonstrate the effectiveness of our solutions.

The remaining of the paper is organized as follows. After introducing background works in Section 2, we review the PLSA algorithm with mathematical representations using matrix formation in Section 3. This formulation should facilitate the description of the GPLSA algorithm for the clustering task for single modality in section 4.1 and the extension for multi-modality retrieval in section 4.2. Section 5 describes experimental validation by applying the GPLSA algorithms to the image clustering, image/text cross-modal retrieval and multi-lingual topic analysis applications. Section 6 concludes the paper with discussions and future works.

2. Related Works

Existing works regarding the integration of a network structure with topic modeling can be organized into two categories. The first category focuses on graph constrained topic modeling for data with single modality, where topic selection preferences are smoothed among nearest neighbors on the graph [18, 25, 26]. For instance, GNMF [18] aims to find a compact representation which uncovers the hidden semantics from the documents and in the meantime represents the intrinsic geometric structure. A semantic representation space is found based on two assumptions that (1) that if two data points are connected along a graph edge, they should be sufficiently close to each other, and (2) the representations of these two data points with respect to the new basis are also close to each other. However, such jointly learned latent representations may do not have explicit probabilistic interpretations. Other approaches rely on Bayesian inference performed on the topic network [27, 28]. The *Relational Topic Model* (RTM) [27] uses LDA to model the documents and the relationships between them, but the Bayesian model is not efficient enough to learning and inference, which suffers from increased complexity problems for both steps.

Another category of structural or graph-contrained topic analysis methods are formulated for multi-modal data [29, 25, 26]. In an early work [29], the interdependencies between published documents take the form of citations which allow instant access to the referenced documents, the citations in the documents are considered as a separate modality in the document corpus. The topics learned form the individual modalities are weighted combined as the shared topics of the two modalities. More recently, there has been an effort to jointly model the documents topics and other auxiliary information provided within the dateset [14, 15, 16]. For example, the Wikipedia data documents typically include both texts and images, and the work in [30] uses Markov random field of topic models to associate images and texts based on their similarity. However, it assumes that each data entity and its associated auxiliary data share the same topic compositions across modalities in these methods. Such assumption might be too restrictive to be applied on the real-world multi-modal datasets. Recently, Bayesian based method tries to study multi-modal topics, a Markov random field (MRF) augmented probabilistic topic model is proposed more recently in [28], which incorporates the similarities between associated topic compositions of different data modalities using MRF. Although good performance in [28] are achieved, the Bayesian MRF method has the problem of increased complexity in both of learning and inference algorithms that usually be resolved with Monte-Carlo methods.

Graph-contrained topic analysis methods are also widely applied on multi-lingual text which can be regarded as multi-modal data[28]. For instance, [31] incorporates a bilingual dictionary based on translation bipartite graph into cross-lingual PLSA to extract common topics in cross-languages. Similariy, [32] presents a novel multilingual topic model, they first build a bipartite graph matching over terms in both languages assuming that words have similarity on document level contexts, then the matching topics are learned as the distributions of these matching pairs instead of being distributions over terms. However, both of the method relay on term pairs in the dictionary and their assumption of matching terms may result loosing of correlated information between the languages. In contrast, in this work, we aim to extract topics from different information sources (images and texts, or texts in different languages) that reflect their intrinsic conceptual similarities. It inherits the advantage of topic models that the learned topics are often intuitive and interpretable.

Therefore, it is useful to extend simpler topic analysis methods such as PLSA to learning from multi-modal data, whose efficient implementation can be used for rapid analysis of large multi-modal dataset and initializations of more sophisticated Bayesian methods. Several works also suggested that connecting multiple modalities using a graph structure is crucial for strong performance of the learning algorithms on multi-modality datasets [33, 34, 35, 30], with applications shown in [36, 37, 28]. Two specific methods of extending PLSA to multi-modal learning with co-regularization has been studied in two recent works [21, 22]. Because of the close relation between PLSA and NMF algorithms [19, 20], coPLSA can also be regarded as a co-regularized NMF problem. However, most existing co-regularized NMF methods use ℓ_2 divergence for both main objective and co-regularizer, and do not consider the normalization constraint. The co-regularizer used in [21] is based on the mutual similarities of data in the topic space, and that of [22] is the ℓ_2 divergence between the topic assignments in the latent space. The common drawback of both methods, however, is that the optimization procedure cannot guarantee monotonic improvement of the objective function before a stationary point is reached. As such, the algorithms in these previous works do not afford guarantees to converge and usually lead to inferior performance.

3. Background: PLSA Algorithm

We first introduce notations and definitions to be used throughout the paper. A d-dimensional vector $\mathbf{v}$ is stochastic if $v_i \geqslant 0$ and $\sum_{i=1}^{d} v_i = 1$, and corresponds to a categorical probability distribution over d outcomes. A $d{\times}n$ nonnegative matrix V is stochastic if its column vectors are stochastic.

For two d-dimensional stochastic vectors $\mathbf{v}$ and $\mathbf{w}$, we define their ℓ_1, ℓ_2 and Kulback-Leibler (KL) divergences, as: $\mathcal{D}_{\ell_1}(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{d} |v_i - w_i|$, $\mathcal{D}_{\ell_2}(\mathbf{v},\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{d}(v_i - w_i)^2$, $\mathcal{D}_{\mathrm{KL}}(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{d} v_i \log \frac{v_i}{w_i}$, and their symmetric KL divergence is defined as $\mathcal{D}_{\mathrm{sKL}}(\mathbf{v},\mathbf{w}) = \mathcal{D}_{\mathrm{KL}}(\mathbf{v},\mathbf{w}) + \mathcal{D}_{\mathrm{KL}}(\mathbf{w},\mathbf{v})$. Accordingly, we define the divergence between two stochastic matrices V and W as the sum of the divergences between their corresponding columns, as $\mathcal{D}_*(W,V) = \sum_j \mathcal{D}_*(W_{\cdot,j}, V_{\cdot,j})$, where $\mathcal{D}_*$ can be replaced with $\mathcal{D}_{\ell_1}$, $\mathcal{D}_{\ell_2}$, $\mathcal{D}_{\mathrm{KL}}$ or $\mathcal{D}_{\mathrm{sKL}}$. For stochastic vectors/matrices, these divergences are non-negative and equal to zero if and only if the two vectors/matrices are identical.

Making analogy to a collection of text documents, we use a "bag-of-word" representation [1] of a dataset, where each data entity (a "document") is represented as the normalized frequencies over some basic features ("words" in a "vocabulary"). PLSA is performed based on a simple probabilistic generative model of the dataset [2]: each word in a document is a sample from a mixture model; each component of the mixture model is a categorical distributions over the vocabulary (a "topic"); the mixing weights of the mixture model correspond to a probability distribution over the topics, and provides the topic composition of the data entity.

Specifically, given n documents $(\mathbf{d}_1, \cdots, \mathbf{d}_n)$ over a vocabulary of size d, $(\mathbf{w}_1, \cdots, \mathbf{w}_d)$, we use stochastic matrix V of dimension $d \times n$ to represent conditional probabilities, as $V_{ij} \equiv \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{doc} = \mathbf{d}_j)$. Assuming the documents are associated with m topics, $(\mathbf{t}_1, \cdots, \mathbf{t}_m)$, we use stochastic matrices W of dimension $d \times m$ and H of dimension $m \times n$ to represent conditional probabilities, as $W_{ik} \equiv \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{topic} = \mathbf{t}_k)$ and $H_{kj} \equiv \texttt{Prob}(\text{topic} = \mathbf{t}_k | \text{doc} = \mathbf{d}_j)$, respectively. According to the document generation model, documents and words are conditionally independent from each other. As such, these probabilities satisfy.

$$\texttt{Prob}(\text{word} = \mathbf{w}_i | \text{doc} = \mathbf{d}_j) = \textstyle\sum_k \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{topic} = \mathbf{t}_k)\texttt{Prob}(\text{topic} = \mathbf{t}_k | \text{doc} = \mathbf{d}_j)$$

With the matrix notations, this is equivalent to $V = WH$. Given a dataset represented in matrix V, PLSA attempts to find its decomposition into W and H, formulated as an optimization problem: $\min_{W,H} \mathcal{D}_{KL}(V, WH)$, with the constraint that both W and H are stochastic matrices. After dropping irrelevant constant terms, minimizing the KL divergence is equivalent to maximizing

$$\mathcal{J}(W,H) = \textstyle\sum_{ij} V_{ij} \log (WH)_{ij}. \tag{1}$$

This optimization problem can be solved with block coordinate ascent by iteratively optimizing W or H while fixing the other until converging to a local optimum. The individual optimization step for W and H is solved with the EM algorithm[38, 39]. To facilitate subsequent discussions, we briefly review the EM algorithm using the matrix notations introduced early in this section.

Optimizing W: Introducing a different stochastic matrix $\hat{W}$, we first define an auxiliary function

$$\mathcal{F}(W,\hat{W}) = \textstyle\sum_{ijk} \frac{V_{ij}\hat{W}_{ik}H_{kj}}{(\hat{W}H)_{ij}} \log\left(\frac{W_{ik}}{\hat{W}_{ik}}\left(\hat{W}H\right)_{ij}\right) = \sum_{ik} M_{ik} \log W_{ik} + \text{const.} \tag{2}$$

In the last step, terms irrelevant to W are collected into a constant. Nonnegative matrix $M = \hat{W} \otimes \left[(V \oslash (\hat{W}H))H^T\right]$ is formed with element-wise matrix multiplication $\otimes$ and division $\oslash$. An application of the Jensen's inequality shows that $\mathcal{F}(W,\hat{W}) \leqslant \mathcal{J}(W,H)$ with equality holds when $W = \hat{W}$, *i.e.*, $\mathcal{F}(W,\hat{W})$ is a tight lower-bound of $\mathcal{J}(W,H)$. Derivation of Eq.(2) and proof of $\mathcal{F}(W,\hat{W})$ being a tight lower-bound of $\mathcal{J}(W,H)$ are provided in the Appendix A.

The EM algorithm optimizing W uses the above lower-bound to improve the objective function in an iterative manner: Starting with an initial values $W = W^{(0)}$, we iteratively solve for $W^{(t+1)} \leftarrow \operatorname{argmax}_W \mathcal{F}(W, W^{(t)})$ with the constraint W being stochastic. As we have $\mathcal{J}(W^{(t)},H) = \mathcal{F}(W^{(t)},W^{(t)}) \leqslant \mathcal{F}(W^{(t+1)},W^{(t)}) \leqslant \mathcal{J}(W^{(t+1)},H)$, the sequence $(W^{(0)}, W^{(1)}, \cdots)$ monotonically increases $\mathcal{J}(W,H)$ until reaching a local maximum.

During each iteration step of the EM algorithm, we solve for $\operatorname{argmax}_W \mathcal{F}(W, W^{(t)})$, which using Eq.(2) reduces to

$$\max_W \quad \textstyle\sum_{ik} M_{ik} \log W_{ik}, \text{ s.t. } W_{ij} \geqslant 0 \ \ \& \ \ \sum_i W_{ij} = 1. \tag{3}$$

The solution to this problem is given by $W_{ik} = \frac{M_{ik}}{\sum_{i'} M_{i'k}}$ (proof given in the Appendix A), in which the normalization step and the non-negativity of M assures W to be a stochastic matrix.

Optimizing H: The EM algorithm optimizing H with fixed W proceeds similarly. First using an auxiliary stochastic matrix $\hat{H}$ we define function

$$\mathcal{G}(H,\hat{H}) = \sum_{ijk} \frac{V_{ij}W_{ik}\hat{H}_{kj}}{(W\hat{H})_{ij}} \log\left(\frac{H_{kj}}{\hat{H}_{kj}}\left(W\hat{H}\right)_{ij}\right) = \sum_{kj} Q_{kj} \log H_{kj} + \text{const}, \tag{4}$$

with matrix $Q = \hat{H} \otimes \left[W^T(V \oslash (W\hat{H}))\right]$. With a similar argument, we can show that $\mathcal{G}(H,\hat{H})$ is also a tight lower-bound of $\mathcal{J}(W,H)$ (proof given in the Appendix A), on the basis of which the EM algorithm is obtained. Specifically, each step of the EM algorithm solves

$$\max_H \quad \sum_{kj} Q_{ik} \log H_{kj}, \text{ s.t. } H_{kj} \geqslant 0 \ \ \& \ \ \sum_k H_{kj} = 1, \tag{5}$$

of which the solution is given by $H_{kj} = \frac{Q_{kj}}{\sum_{k'} Q_{k'j}}$ (proof given in the Appendix A).

4. Algorithm for graph-regularized PLSA

We start with a general setting of GPLSA problem, then extend the formulation to multi-modality problem.

4.1. General Formulation

In this section, we introduce our GPLSA algorithms. Formally, given a dataset represented with a stochastic matrix V of size $d \times n$, we seek factorization $V \approx WH$, with a stochastic matrix W of size $d \times k$ and a stochastic matrix H of size $k \times n$. Consider a graph with n vertices, where each vertex corresponds to a data point represented by each column of the stochastic matrix V. For each data point v_i, we add edges between v_i and the other data points to formulate the relation matrix R among the vertices[1]. The graph regularizers are the divergences between discrete probability distributions corresponding to the composition of topics from a data entry, which enable effects from entities of the same semantics to be smoothed by each other. Specifically, we formulate GPLSA algorithms as a graph regularizer constrained optimization problem as,

$$\min_{W,H} \sum_{kj} \mathcal{D}_{KL}(V, WH) + \lambda \mathcal{D}_*(H_{kj}, H_{kl}) R_{lj}, \tag{6}$$

with the constraint that W and H are stochastic matrices. Parameter $\lambda > 0$ balances the contribution of the GPLSA objectives of the standard PLSA term and the co-regularization term. In the following, $\mathcal{D}_*$ will be replaced with the ℓ_1, ℓ_2 and symmetric KL divergences. Dropping irrelevant constant terms, the objective function of Eq.(6) can be further simplified to

$$\max_{W,H} \sum_{kj} \mathcal{J}(W,H) - \lambda \mathcal{D}_*(H_{kj}, H_{kl}) R_{lj}, \tag{7}$$

with the same constraints on the factors.

As in the case of PLSA, in the learning step of GPLSA, the objective function in Eq.(7) is optimized with a block-coordinate descent scheme, by alternating steps between the optimization of each of W and H while fixing the other factors. In the following, we describe the EM steps of these sub-problems.

Optimizing W: The step optimizing each W is the same as the optimization of W in standard PLSA. As such, the solution can be obtained via solving a sequence of optimization problem given in Eq.(3).

Optimizing H: The optimization of H is different because of the graph-regularizer. The optimization of H_{kj} with fixed W and the other columns in H, i.e. H_{kl}, where $l \in n$ and $l \neq j$, after removing irrelevant constant terms, becomes

$$\begin{aligned} \max_H \sum_{kj} & \mathcal{J}(W,H) - \lambda \mathcal{D}_*(H_{kj}, H_{kl}) R_{lj}, \\ \text{s.t. } & H_{kj} \geqslant 0 \ \ \& \ \ \sum_k H_{kj} = 1. \end{aligned} \tag{8}$$

Using the auxiliary function $\mathcal{G}$ defined in Eq.(4), we can also obtain a tight lower-bound of the above objective function, as: $\mathcal{G}(H,\hat{H}) - \lambda \mathcal{D}_*(H_{kj}, H_{kl}) R_{lj} \leqslant \mathcal{J}(W,H) - \lambda \mathcal{D}_*(H_{kj}, H_{kl}) R_{lj}$ with equality when $\hat{H}_{kj} = H_{kj}$, which follows

[1] In this work, we follow the default settings of GNMF using the 0-1 weighting scheme for the relation matrix R, where $R_{jl} = R_{lj}$, $l, j \in n$.

from the property of $\mathcal{G}$. Note that in this lower-bound, the second term $\lambda\mathcal{D}_*(H_{kj}, H_{kl})R_{lj}$ does not depend on the auxiliary variable $\hat{H}_{kj}$.

Then, a similar EM algorithm can be developed to optimize H_{kj} iteratively, which improves the lower-bound in each iteration. Starting with the initial $H_{kj} = H^0_{kj}$, we iteratively solve for

$$\begin{aligned} H^{(t+1)}_{kj} \leftarrow & \operatorname*{argmax}_{H_{kj}} \mathcal{G}(H_{kj}, H^{(t)}_{kj}) - \lambda\mathcal{D}_*(H_{kj}, H_{kl})R_{lj}, \\ & \text{s.t. } H_{kj} \geqslant 0 \ \ \& \ \ \sum_k H_{kj} = 1. \end{aligned} \tag{9}$$

We provide efficient algorithms for symmetric KL, ℓ_2 and ℓ_1 divergences with convergence guarantees. The essential optimization problem we need to solve is

$$\begin{aligned} \max_{H_{kj}} & \sum_{kj} Q_{kj} \log H_{kj} - \lambda\mathcal{D}_*(H_{kj}, H_{kl})R_{lj}, \\ & \text{s.t. } H_{kj} \geqslant 0 \ \ \& \ \ \sum_k H_{kj} = 1. \end{aligned} \tag{10}$$

With respect to three types of co-regularizer, namely, symmetric KL, ℓ_2 and ℓ_1 divergences, the optimal solution to Eq.(10) are given as non-linear functions of a scalar variable η_j that corresponds to the Lagrangian multiplier of the normalizing constraint, $\sum_k H_{kj} = 1$.

because, these solutions are given in the following equations (proof given in the Appendix B):

- For $\mathcal{D}_* = \mathcal{D}_{\text{sKL}}$,
$$H_{kj}(\eta_j) = \frac{Q_{kj} + \lambda H_{kl} R_{lj}}{\lambda C \mathcal{W}\Big(\frac{Q_{kj}+\lambda H_{kl}R_{lj}}{\lambda C} \exp(\frac{\eta_j}{\lambda C} - \frac{S_{kj}}{C} + 1)\Big)}, \tag{11}$$
where $\mathcal{W}_0(\cdot)$ is the principal branch of the Lambert $\mathcal{W}$ function [40] that is defined implicitly as $z = W(z)e^{W(z)}$ for $z > 0$ [2]. R_{lj} represents the relation weight between document l and j, $\sum_l R_{lj}$ denotes the weight summation between document j and all other documents l, $l \in \{1, n\}$ and $l \neq j$. $(\log H)_{kl}R_{lj}$ and $H_{kl}R_{lj}$ are the constant terms with respect to the current variable, i.e., the j th column of matrix H, we denote $C = \sum_l R_{lj}$, $S_{kj} = (\log H)_{kl}R_{lj}$ for simplicity.

- For $\mathcal{D}_* = \mathcal{D}_{\ell_2}$,
$$H_{kj}(\eta_j) = \frac{\lambda H_{kl}R_{lj} - \eta_j + \sqrt{(\eta_j - \lambda H_{kl}R_{lj})^2 + 4\lambda(\sum_l R_{lj})Q_{kj}}}{2\lambda(\sum_l R_{lj})}, \tag{12}$$

- For $\mathcal{D}_* = \mathcal{D}_{\ell_1}$,
$$H_{kj}(\eta_j) = \begin{cases} \frac{Q_{kj}}{\eta_j + \lambda\sum_l R_{lj}} & -\lambda\sum_l R_{lj} < \eta_j < \frac{Q_{kj}}{H_{kl}} - \lambda\sum_l R_{lj}, \\ H_{kl} & \frac{Q_{kj}}{H_{kl}} - \lambda\sum_l R_{lj} \leq \eta_j \leq \frac{Q_{kj}}{H_{kl}} + \lambda\sum_l R_{lj}, \\ \frac{Q_{kj}}{\eta_j - \lambda\sum_l R_{lj}} & \frac{Q_{kj}}{H_{kl}} + \lambda\sum_l R_{lj} < \eta_j. \end{cases} \tag{13}$$
Compared with the two other divergence types, the update steps for ℓ_1 divergence in Eq.(13) corresponds to a piecewise function. The computation only involves arithmetic operations and thresholding, which is substantially more simple and efficient. The use of ℓ_1 co-regularizer has another important property, that the resulting H_{kj} can have identical components as H_{kl}. This is usually not the case for the ℓ_2 and symmetric KL regularizers.

We use $H_{kj}(\eta_j)$ in Eqs.(11,12,13) to emphasize the fact that they are functions of the scalar parameter η_j. To determine the value of η_j, which in turn leads to the optimal solution to H, we can solve the following 1D nonlinear equation corresponding to the normalization constraint in Eq.(10),

$$\sum_k H_{kj}(\eta_j) = 1. \tag{14}$$

[2] The Lambert $\mathcal{W}$ function can be numerically evaluated and is provided in popular numerical tools such as MATLAB (function `lambertw`) or SciPy (function `scipy.special.lambertw`). It has been used in algorithms that enforce entropic priors [41]. It also appears in a variant of PLSA to encourage sparsity over the obtained W or H factors [42].

For each type of graph regularizers, we use the corresponding $H_{kj}(\eta_j)$ in Eqs.(11,12,13). For each column index j, Eq.(14) is solved numerically, *e.g.*, with Newton-Raphson when $H_{kj}(\eta_j)$ is differentiable (*e.g.*, $\mathcal{D}_* = \mathcal{D}_{\ell_2}$ or $\mathcal{D}_{\text{sKL}}$) or bi-section when otherwise (*e.g.*, $\mathcal{D}_* = \mathcal{D}_{\ell_1}$).

In summary, we solve the GPLSA problems with an iterative algorithm that alternates between the optimization of individual W and H factors while fixing the others. The optimization of W factor is performed with another iterative EM algorithm based on individual optimization steps given in Eq.(3). The optimization of H factor is achieved by iterating steps that first solve Eq.(14) and then determine the factors with Eq.(11), Eq.(12) or Eq.(13). In practice, all iterative algorithms converges within 5-10 steps.

4.2. Extend GPLSA to multi-modality

Our GPLSA algorithms can be easily extended to the case with more than one modalities. Formally, given L modalities of the dataset, represented with stochastic matrices $V^{(l)}$ ($l \in 1, ..., L$) of size $d_l \times n$, we seek factorization $V^{(l)} \approx W^{(l)}H^{(l)}$, with stochastic matrices $W^{(l)}$ of size $d_{(l)} \times m$ and a matrix $H^{(l)}$ of size $m \times n$ representing the m modality-specific topic matrices and the topic compositions of the dataset, respectively. In GPLSA, association of different modalities to their common data entry is achieved by coupling the factorizations $V^{(l)} \approx W^{(l)}H^{(l)}$, *i.e.*, besides individual GPLSA objectives to each modality. The relation matrix R of the graph regularizers is constructed among $H^{(l)}$ based on the assumption that different data modalities admit similar underlying semantic structure, thus the algorithms aim to minimize the difference of H matrices from different modalities corresponding to each respective topic compositions.

Specifically, GPLSA algorithms are formulated as a constrained optimization problem as

$$\min_{W^{(l)},H^{(l)}} \sum_{l=1,...,L} \mathcal{D}_{KL}(V^{(l)}, W^{(l)}H^{(l)}) + \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}), \tag{15}$$

with the constraint that $W^{(l)}$ and $H^{(l)}$ are stochastic matrices, for simplicity, we use $H^{(\backslash l)}$ to denote the other H factor other than $H^{(l)}$. Parameter $\lambda > 0$ balances the contribution of the GPLSA objectives of each modality and the co-regularization term. In the following, $\mathcal{D}_*$ will be replaced with the symmetric KL, ℓ_2 or ℓ_1 divergences[3]. Dropping irrelevant constant terms, the objective function of Eq.(15) can be further simplified to

$$\max_{W^{(l)},H^{(l)}} \sum_{\ell=1,...,L} \mathcal{J}(W^{(l)}, H^{(l)}) - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}), \tag{16}$$

with the same constraints on the factors.

As in the case of PLSA, in the learning step of GPLSA, the objective function in Eq.(16) is optimized with a block-coordinate descent scheme, by alternating between steps optimizing each of $W^{(l)}$, $H^{(l)}$, while fixing the other factors. In the following, we describe the steps of these sub-problems.

Optimizing $W^{(l)}$: The step optimizing each $W^{(l)}$ is the same as the optimization of W in PLSA. As such, the solution can be obtained via solving a sequence of optimization problem given in Eq.(3).

Optimizing $H^{(l)}$: The optimization of $H^{(l)}$ with fixed $W^{(l)}$ and $H^{(\backslash l)}$, after removing irrelevant constant terms, becomes

$$\begin{aligned} \max_{H^{(l)}} \quad & \sum_{kj} Q_{kj} \log H^{(l)}_{kj} - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}), \\ \text{s.t.} \quad & H^{(l)}_{kj} \geqslant 0 \;\; \& \;\; \sum_k H^{(l)}_{kj} = 1. \end{aligned} \tag{17}$$

Similar to GPLSA on single modality, with respect to three types of co-regularizer for difference modality (namely, symmetric KL, ℓ_2 and ℓ_1 divergences), the optimal solution to Eq.(17) are given as non-linear functions of a scalar variable η_j that corresponds to the Lagrangian multiplier of the normalizing constraint, $\sum_k H^{(l)}_{kj} = 1$. Specifically, these solutions are given in the following equations, with proofs given in the Appendix C:

- For $\mathcal{D}_* = \mathcal{D}_{\text{sKL}}$,

$$H^{(l)}_{kj}(\eta_j) = \frac{Q_{kj} + \lambda H^{(\backslash l)}_{kj}}{\lambda \mathcal{W}_0\left(\frac{Q_{kj}+\lambda H^{(\backslash l)}_{kj}}{\lambda H^{(\backslash l)}_{kj}} \exp\left(1 + \frac{\eta_j}{\lambda}\right)\right)}, \tag{18}$$

[3] It is also possible to incorporate other regularization terms on the factors $W^{(l)}$ and $H^{(l)}$ to express other preference on the factors such as sparsity. Furthermore, we can us similar methods to enforce consistencies in parts of factor $W^{(l)}$. However, for simplicity, in the current work we do not consider these types of regularizers.

	Accuracy (%)				Normalized Mutual Information (%)			
K	Baseline	Regularizer	GNMF [18]	GPLSA	Baseline	Regularizer	GNMF[18]	GPLSA
10	43.89	L1	—	**60.00**	54.14	L1	—	**65.95**
		L2	79.58	**85.00**		KL	87.92	**88.62**
		KL	84.31	**87.36**		L2	88.44	**90.18**
15	65.37	L1	—	**67.87**	71.08	L1	—	**76.54**
		L2	84.35	**85.83**		L2	85.31	**86.22**
		KL	84.17	**86.39**		KL	87.60	**89.18**
20	60.49	L1	—	**73.68**	73.86	L1	—	**81.69**
		L2	72.22	**80.97**		L2	87.60	**88.36**
		KL	73.68	**81.81**		KL	85.18	**90.14**

Table 1: *Comparison of results regarding clustering performance of GPLSA* vs. *GNMF [18] on the COIL20 dataset in terms of accuracy and normalized mutual information.*

- For $\mathcal{D}_* = \mathcal{D}_{\ell_2}$,

$$H_{kj}^{(l)}(\eta_j) = \frac{1}{2}\sqrt{\left(H_{kj}^{(\backslash l)} - \frac{\eta_j}{\lambda}\right)^2 + \frac{4Q_{kj}}{\lambda}} + \frac{1}{2}\left(H_{kj}^{(\backslash l)} - \frac{\eta_j}{\lambda}\right), \tag{19}$$

- For $\mathcal{D}_* = \mathcal{D}_{\ell_1}$,

$$H_{kj}^{(l)}(\eta_j) = \begin{cases} \frac{Q_{kj}}{\eta_j+\lambda}, & -\lambda < \eta_j < \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} - \lambda, \\ H_{kj}^{(\backslash l)}, & \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} - \lambda \leqslant \eta_j \leqslant \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} + \lambda, \\ \frac{Q_{kj}}{\eta_j-\lambda}, & \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} + \lambda < \eta_j. \end{cases} \tag{20}$$

5. Experiments

In this section, we first evaluate the GPLSA algorithms on image clustering to justify how well they performs on topic analysis from data of a single modality. We then apply them to cross-modality retrieval from documents containing both images and texts. Specifically, our evaluation focused on the task of text retrieval from an image query (*i-2-t*), and image retrieval from a query with a text document (*t-2-i*). The evaluations are performed on three benchmark data sets described in the following. Finally, we also apply them on multi-lingual topic analysis problems.

5.1. Image Clustering

We use the Columbia Image Library (COIL-20) data set [18], which contains 32×32 gray scale images of 20 classes (objects). There are 72 images taken for each object with different view angles; example images are shown in Fig.1.

Figure 1: Image samples from COIL-20 data set.

The clustering results are evaluated by two metrics, the accuracy (AC) and the normalized mutual information metric (NMI). The AC is defined as:

$$AC = \frac{\sum_{i=1}^{n} \delta(r_i, s_i)}{n}, \tag{21}$$

where r_i is the estimated cluster label, s_i is the ground truth label, and n is the number of examples. $\delta(r_i, s_i)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise. The NMI are defined as follow:

$$NMI(C, \bar{C}) = \frac{MI(C, \bar{C})}{max(H(C), H(\bar{C}))}, \tag{22}$$

Methods	Topic space similarity				Semantic space similarity			
	TVGraz		Wikipedia		TVGraz		Wikipedia	
	i-2-t	t-2-i	i-2-t	t-2-i	i-2-t	t-2-i	i-2-t	t-2-i
SCM [33]	0.460	0.450	0.267	0.219	0.664	0.649	0.362	0.273
Link PLSA [29]	0.349	0.349	0.247	0.247	0.803	0.803	0.605	0.605
ℓ_1 GPLSA	0.359	0.365	0.317	0.307	0.723	0.726	0.667	0.658
ℓ_2 GPLSA	0.450	0.445	0.360	0.358	0.846	0.845	0.706	0.701
sKL GPLSA	**0.481**	**0.481**	**0.413**	**0.413**	**0.850**	**0.850**	**0.726**	**0.724**

Table 2: *Performance comparison of GPLSA with 3 regularizers and two other multi-modal learning algorithms in terms of mean average precision (mAP) on two standard public image/text benchmark datasets. See texts for details.*

where $H(C)$ and $H(\bar{C})$ are the entropies of $H(C)$ and $H(\bar{C})$, the mutual information (MI) is defined as:

$$MI(C, \bar{C}) = \sum_{c_i \in C, \bar{c}_j \in \bar{C}} p(c_i, \bar{c}_j) \log_2 \frac{p(c_i, \bar{c}_j)}{p(c_i)p(\bar{c}_j)}, \tag{23}$$

where C is the set of ground truth clusters, $\bar{C}$ is the clustering result from the testing algorithm. $p(c_i)$ and $p(\bar{c}_j)$ are the probabilities of a randomly selected image, which belongs to clusters c_i and $\bar{c}_j$, respectively. $p(c_i, \bar{c}_j)$ is the joint probability that this randomly selected image belongs to the clusters c_i and $\bar{c}_j$ at the same time. In case two sets of clusters are identical, NMI is 1. In case two sets are independent, NMI is 0.

We compare GPLSA with (1) a baseline algorithm with K-means clustering in the original feature space and (2) the GNMF method [18], which has been shown to achieve superior performance against classic clustering algorithms (i.e. K-means, SVD, NCut, standard NMF)[4]. After applying GPLSA and GNMF, a low dimensional representation for each image is obtained. The clustering of images is then performed upon this low dimensional representation using standard K-means for a final evaluation. We follow the default settings of GNMF using the 0-1 weighting scheme and use 5 nearest neighbors for the relation matrix R; and we set the number of topics equals to the number of clusters, the optimized balance parameter λ in GPLSA algorithms is chosen by cross-validation on a subset of the training data. We evaluate the performance with different number of clusters (K in Table 1), where the evaluated clusters (classes) are randomly chosen from 20 classes (objects).

Table 1 shows the results of GPLSA algorithms with ℓ_1, ℓ_2 and symmetric KL regularizers on the COIL-20 dataset. The results using the ℓ_1 regularizer consistently outperform the baseline. We note that the use of ℓ_1 regularizer in GNMF is not available, thus the comparison is omitted. GPLSA outperforms GNFM in both cases of symmetric KL and ℓ_2 regularizers. Finally, the symmetric KL graph-regularizer achieves the best overall performance. This is expectable as we solve the objective function with the log term using the Lambert $\mathcal{W}$ function, and in comparison GNMF relies on a rough approximate to solve a nonlinear equation.

5.2. Cross-Modal Image/Text Retrieval

For multi-modality data, the difference in original data representations of images and text are encoded into the corresponding topics with the GPLSA model. And with their topic compositions, images and texts are projected into a compatible semantic space, this new representation can then be used to do establish connections between images and text documents and facilitates cross-modality retrieval. Two standard benchmark image/text datasets were used in our experiments: *Wikipedia* [34] and *TVGraz* [43]. The *Wikipedia* dataset consists of 2866 image/text pairs of 30 semantic categories, includes a standard training and testing sets split with 2173 and 693 image/text pairs. The *TVGraz* dataset consists of 2058 image/text pairs of 10 semantic categories with an average document length of 289 words, and is split into training and testing sets with 1558 and 500 image/text pairs. Images and texts in both datasets are converted to bag-of-word representation, where for images we used 1024 visual keywords as a result of clustering the SIFT features from all training images, and 6203 unique text words were selected after stemming and removal of the common stop words.

[4]The GNMF code and dataset are released and available at http://www.cad.zju.edu.cn/home/dengcai/Data/GNMF.html

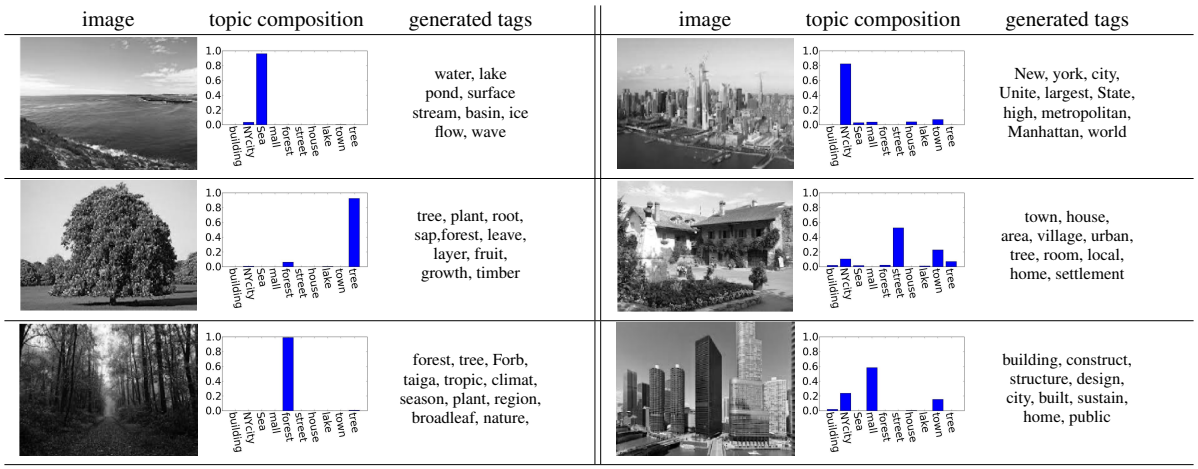

Figure 2: *Example images with generated topic compositions and tags obtained with sKL-GPLSA from the Wikipedia dataset.*

In the learning phase, we determine modality-specific topics using the GPLSA learning algorithm (Section 4.2) on the training sets. We extract 100 topics from the *Wikipedia* dataset and 50 topics from the *TVGraz* dataset, and the optimized balance parameter λ in GPLSA algorithms is chosen by cross-validation on a subset of the training data. When performing retrieval tasks on the testing set, we first recover the topic composition of the queried image or text using the PLSA algorithm (only the optimization of H matrix) using the learned modality-specific topics. The similarities between topic compositions of the queried image and texts in testing set (in task *i-2-t*) or queried text and images in testing set (in task *t-2-i*) are then evaluated and ranked. Two similarity measures are used in our experiments, the centered normalized correlation between the topic compositions and the centered normalized correlations between topic compositions transformed by a multi-class logistic regression function learned during training, which maps topic compositions to the semantic categories pre-defined for each dataset. As such, the former evaluates correlations of topic compositions directly, while the latter can be regarded as the correlation in a more semantically meaningful space induced from the topic compositions [33]. We use the mean average precision (MAP) scores over all testing data as performance metric. The average precision score for each query is computed as the mean precision value for the top 10 relevant retrievals. Here, we determine a relevant retrieval occurs if the retrieved text/image is from the same semantic category as the image/text used in query.

Table 2 shows the results of GPLSA algorithms with ℓ_1, ℓ_2 and symmetric KL regularizers on the two datasets. For comparison, we also include retrieval performance based on a link-PLSA model that requires the topic compositions of associated text and image to be identical. The link-PLSA algorithm can be implemented as described in [29]. Furthermore, all results were compared to a baseline established by the method of *semantic correlation matching* (SCM) [33], which represents the state-of-the-art performance in text/image cross-modal retrieval tasks. Results in Table 2 suggests that for the two cross-modal retrieval tasks, GPLSA algorithms in general achieve better performance than the link PLSA algorithm, and also outperform the SCM method that is based on kernel canonical correlation analysis. This may be attributed to, on the one hand, the more semantic relevance of the representation (as probability mixture of thematic topics of the text/images) obtained with GPLSA, and on the other hand, its less restrict assumption that allows for mis-match of topic compositions of associated text and images. This is further corroborated by observing that the MAP scores for *i2t* and *t2i* tasks are similar with GPLSA algorithms, suggesting the diminished representational difference between the two modalities in the topic space found by GPLSA. Furthermore, all three variants of the GPLSA algorithms achieves better performance and efficient computation, but symmetric KL graph-regularizer leads to the best overall performance. Last, combining with more semantic abstraction, as concluded in [33], can also significantly improve the retrieval performance.

In Fig.2 we further show several test images from the *Wikipedia* dataset with their corresponding topic compositions over a subset of topics obtained with the symmetric KL GPLSA algorithm (the names of each topic is manually assigned based on the top words from each topic to facilitate understanding), together with text tags that are generated by sampling from the topic mixtures associated with each image. The visualized topic compositions and the generated text tags of these images obtained with GPLSA span wide semantic ranges, and can shed some light on their effects

Topic 1 - (Turkey nuclear)		Topic 2 - (Europe peace)	
German	English	German	English
turkei (turkey)	Turkey	israel (israel)	peace
frag (question)	nuclear	europa (Europe)	Israel
prasident (president)	question	palastinens (palestinians)	Palestinian
moglich (possible)	President	union (union)	European
kommission (commission)	people	staat (State)	negoti
antwort (answer)	Commission	region (region)	territory
nuklear (nuclear)	concern	ost (east)	Union
fall (event)	time	friedensprozess (peace process)	East
zeit (time)	Council	abkomm (Agreem)	agreement
bedenk (Bedenk)	possibl	gebiet (area)	State
Topic 3 - (fishing environment)		**Topic 4 - (vehicle)**	
German	English	German	English
schiff (ship)	fish	herstell (producible)	car
fischerei (fishing)	disaster	fahrzeug (vehicle)	manufacture
katastroph (disaster)	fisher	kost (costly)	cost
erika (heather)	affect	autos (cars)	recycle
fisch (fish)	pollute	verantwort (responsible)	vehicle
europa (Europe)	ship	standpunkt (viewpoint)	environment
schad (defective)	control	gemeinsam (common)	industry
kontroll (control)	damage	automobilindustri (automotive-industrial)	common
umwelt (environment)	sea	recycling (recycling)	consume
fischereisektor (fishing sector)	environment	verbrauch (consumption)	response

Table 3: *Top words of leaned topics on the German-English corpus. See text for details.*

English document: The UN Economic Commission for Europe has also examined the objectives for reductions in the same sources of emissions as in the proposal for a directive now under discussion, and, as a result of these talks, the so-called Gothenburg Protocol was signed. There is a clear difference between this proposal and the Commission's. Mr President, the industry committee, after much discussion and serious consideration, is overwhelmingly opposed to the Commission's proposed ceilings, and this is across all groups and nationalities. I understand that the groups, the whole Parliament, are split on this issue, between supporters of the Committee on the Environment, Public Health and Consumer Policy and supporters of the industry committee's line.
Summarization in German: europa, kommission, union, wirtschaft, parlament, bericht, land, frag, wichtig, prasident, moglich, polit, ziel, mitgliedstaat, sozial
German document: In Kreisen der UNO-Wirtschaftskommission fr Europa sind ebenfalls Ziele fr die Verringerung derselben Emissionsquellen untersucht worden wie in dem jetzt debattierten Vorschlag fr eine Richtlinie, und im Ergebnis dieser Verhandlungen wurde das sogenannte Gteborger Protokoll unterzeichnet. Zwischen diesem Vorschlag und dem Vorschlag der Kommission gibt es einen deutlichen Unterschied. Herr Prsident, der Industrieausschu lehnt nach ausfhrlicher Diskussion und ernsthafter berlegung die von der Kommission vorgeschlagenen Hchstgrenzen mit der berwltigenden Mehrheit aller Fraktionen und Nationalitten ab. Soweit mir bekannt ist, sind die Fraktionen, ist das gesamte Parlament in zwei Lager gespalten, von denen das eine den Ausschu fr Umweltfragen, Volksgesundheit und Verbraucherpolitik untersttzt und das andere den Industrieausschu.
Summarization in English: European, commission, policy, President, Union, service, region, import, report, social, Member, develop, State, country, economic

Table 4: *German-English document summarization results. See text for details.*

in improving the precisions of semantic matching with the queried text document.

5.3. Multi-Lingual Topic Analysis

In multi-lingual topic analysis, our purpose it to produce a topic-level summarization of documents using GPLSA in different languages. From such a summarization, one can quickly grasp the basic subjects concerning a document written in unknown language by another familiar language. As such analysis bypasses the need of machine translation, it can be used in occasions where fast analysis of a vast number of documents in foreign languages is required.

We used a data set that is a subset of the multilingual corpora in the European Parliament Proceedings Parallel Corpus (EPPPC) [44], which contains newswire articles written in different European languages with aligned sentences. In our experiments, we selected 1100 documents of German ⇔ English pairs. Considering different languages as different information sources, topics are learned from a training set containing 1000 German-English documents. Performance of the topic learning algorithms are evaluated on a test set of the remaining 100 documents. After stemming and removing a standard list of the stop words using NLTK[5], the resulting English dictionary contains 10115 words, while the German dictionary contains 19532 words.

We then applied the GPLSA algorithm on this corpus to extract 100 common topics in the two languages. Table 3 shows the top words of two languages from 4 examples of the learned topics (the German words are shown with their

[5]http://www.nltk.org/

English translations in parentheses). As these result shows, common topics capture the correspondence of German and English words without precisely aligning them. More importantly, these words are grouped together under topics of the same concept.

As a simple application, we generate English summarization for German documents that were not covered by the training set. Similar to the previous experiments on image-text documents, we take the simple method of first recovering the topic assignments of the German documents using the learned German topic matrix, then combine it with the English topic matrix to generate corresponding keywords in English. Two examples from the results are shown in Table 4. For readers who do not know German, the extracted keywords in English can provide informative guidelines about the topics of the document (in this case, both are about 'European politics').

6. Conclusion

We have presented efficient algorithms for graph regularized PLSA (GPLSA) to probabilistic topic analysis of both single- and multiple- modality data representation. In GPLSA, topic compositions of a data entity are mapped to a graph and the similarities between topic compositions on the graph are measured with divergences between discrete probabilities. We propose efficient multiplicative iterative algorithms for GPLSA with ℓ_1, ℓ_2 and symmetric KL divergences as regularizers. The optimization problem for each case affords simple numerical solutions that require only matrix arithmetic operations and 1D nonlinear equations. Experimental results in various real-data sets show that the proposed algorithms enhances the performance of the sate-of-the art frameworks.

There are several directions that the current work can be further improved. First, the correlation matrix controls the smoothness of topics in GPLSA model. Thus, learning a suitable correlation matrix is critical to GPLSA algorithms. Secondly, we are working on adapting the GPLSA algorithms to datasets with more sophisticated structures over topics, such as allowing a hierarchical structure of the topics with higher layers capturing more abstract semantic notions. At last, we are also interested in incorporating other type of constraints such as sparseness in multi-modal topic analysis. We will also seek more applications of GPLSA algorithms, for instance to video analysis or multi-modal social data analysis.

Appendix A.

Proof. **Proof of $\mathcal{J}(W,H)$ is tight lower-bounded by $\mathcal{F}(W,\hat{W})$.** First, note that $\left(\hat{W}H\right)_{ij} = \sum_k \hat{W}_{ik}H_{kj}$, or $\sum_k \frac{\hat{W}_{ik}H_{kj}}{\left(\hat{W}H\right)_{ij}} = 1$. Next, consider the concavity of the logarithm function, we first rearrange terms in the definition of $\mathcal{F}(W,\hat{W})$ and then apply Jensen's inequality to the inner term to get

$$\sum_{ij} V_{ij}\left\{\sum_k \frac{\hat{W}_{ik}H_{kj}}{\left(\hat{W}H\right)_{ij}} \log\left(\frac{W_{ik}H_{kj}}{\hat{W}_{ik}H_{kj}}\left(\hat{W}H\right)_{ij}\right)\right\} \leqslant \sum_{ij} V_{ij}\log\left(\sum_k W_{ik}H_{kj}\right) = \sum_{ij} V_{ij}\log(WH)_{ij}.$$

As the last term is $\mathcal{J}(W,H)$, this proves the inequalities $\mathcal{F}(W,\hat{W}) \leqslant \mathcal{J}(W,H)$. Furthermore, equality trivially holds if we have $W = \hat{W}$.

Next, we show the other part of Eq.(2), $\mathcal{F}(W,\hat{W}) = \sum_{ik} M_{ik}\log W_{ik}$+const. We start with the definition of $\mathcal{F}(W,\hat{W})$, as

$$\sum_{ijk} \frac{V_{ij}\hat{W}_{ik}H_{kj}}{\left(\hat{W}H\right)_{ij}} \log\left(\frac{W_{ik}}{\hat{W}_{ik}}\left(\hat{W}H\right)_{ij}\right) \sum_{ik}\hat{W}_{ik}\sum_j\left\{\frac{V_{ij}}{\left(\hat{W}H\right)_{ij}}(H^T)_{jk}\right\}\log W_{ik} + \text{const.}$$

where $\sum_j\left\{\frac{V_{ij}}{(\hat{W}H)_{ij}}(H^T)_{jk}\right\} = \left[(V \oslash (\hat{W}H))H^T\right]_{ik}$, it then follows that the term in front of $\log W_{ik}$ is the element of matrix $M = \hat{W} \otimes \left[(V \oslash (\hat{W}H))H^T\right]$.

Optimal solution to Eq.(3). We first introduce Lagrangian multiplier for each equality constraint η_k in the optimization problem, and form the Lagrangian as:

$$\sum_{ik} M_{ik}\log W_{ik} - \sum_k \eta_k\left(\sum_i W_{ik} - 1\right).$$

Taking derivative of the Lagrangian with regards to each W_{ik} and solving the equation when setting the result to zero yield $W_{ik} = \frac{M_{ik}}{\eta_k}$. Further considering the constraint $\sum_{i'} W_{i'k} = 1$, we have $\eta_k = \sum_{i'} M_{i'k}$, thus proves the result.

Proof of $\mathcal{J}(W,H)$ is tight lower-bounded by $\mathcal{G}(H,\hat{H})$. As in the case of showing $\mathcal{F}(W,\hat{W}) \leqslant \mathcal{J}(W,H)$, we first use the fact that $\left(W\hat{H}\right)_{ij} = \sum_k W_{ik}\hat{H}_{kj}$, or $\sum_k \frac{W_{ik}\hat{H}_{kj}}{\left(W\hat{H}\right)_{ij}} = 1$. We then rearrange terms of $\mathcal{G}(H,\hat{H})$ and then apply Jensen's inequality to obtain

$$\sum_{ij} V_{ij}\left\{\sum_k \frac{W_{ik}\hat{H}_{kj}}{\left(W\hat{H}\right)_{ij}} \log\left(\frac{W_{ik}H_{kj}}{W_{ik}\hat{H}_{kj}}\left(W\hat{H}\right)_{ij}\right)\right\} \leqslant \sum_{ij} V_{ij}\log\left(\sum_k W_{ik}H_{kj}\right) = \sum_{ij} V_{ij}\log(WH)_{ij}.$$

As the last term is $\mathcal{J}(W,H)$, this proves that $\mathcal{G}(H,\hat{H}) \leqslant \mathcal{J}(W,H)$. Furthermore, equality is trivially held when we have $H = \hat{H}$.

Next, we show the other part of Eq.(4), $\mathcal{G}(H,\hat{H}) = \sum_{kj} Q_{kj}\log H_{kj} + \text{const}$. We start with the definition of $\mathcal{G}(H,\hat{H})$, as

$$\sum_{ijk} \frac{V_{ij}W_{ik}\hat{H}_{kj}}{\left(W\hat{H}\right)_{ij}} \log\left(\frac{H_{kj}}{\hat{H}_{kj}}\left(W\hat{H}\right)_{ij}\right) = \sum_{kj}\hat{H}_{kj}\sum_i \left\{(W^T)_{ki}\frac{V_{ij}}{\left(W\hat{H}\right)_{ij}}\right\}\log H_{kj} + \text{const}.$$

Where $\sum_i \left\{(W^T)_{ki}\frac{V_{ij}}{\left(W\hat{H}\right)_{ij}}\right\} = \left[W^T(V \oslash (W\hat{H}))\right]_{kj}$, it then follows that the term in front of $\log H_{kj}$ is the element of matrix that can be written as $Q = \hat{H} \otimes \left[W^T(V \oslash (W\hat{H}))\right]$.

Optimal solution to Eq.(5). We first introduce Lagrangian multiplier for each equality constraint η_j in the optimization problem, and form the Lagrangian as:

$$\sum_{kj} Q_{kj}\log H_{kj} - \sum_j \eta_j\left(\sum_k H_{kj} - 1\right).$$

Taking derivative of the Lagrangian with regards to each H_{kj} and solving the equation when setting the result to zero yield $H_{kj} = \frac{Q_{kj}}{\eta_j}$. Further considering the constraint $\sum_{k'j} H_{k'j} = 1$, we have $\eta_j = \sum_{k'} Q_{k'j}$, thus proves the result.

Appendix B.

Proof. **Proof of Eq.(11)**. We first simplify the objective function of Eq.(10) by dropping irrelevant constant terms to obtain the new objective function as

$$\sum_{kj} Q_{kj}\log H_{kj} - \lambda\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{r}\left(H_{kj}\log\frac{H_{kj}}{H_{kl}} + H_{kl}\log\frac{H_{kl}}{H_{kj}}\right)R_{lj}.$$

Next, we introduce Lagrangian multiplier η_j to each equality constraint on the column of H_{kj} in Eq.(10) and form the Lagrangian. Ignoring the nonnegative constraint (it will be shown to be satisfied in our setup later in this section), the first KKT condition of the problem with regards to H_{kj} is that the derivative of the Lagrangian with regards to H_{kj} vanishes, as

$$\frac{Q_{kj}}{H_{kj}} - \eta_j - \lambda((\sum_l R_{lj})\log H_{kj} + (\sum_l R_{lj}) - (\log H)_{kl}R_{lj} - \frac{H_{kl}R_{lj}}{H_{kj}}) = 0.$$

We define $C = \sum_l R_{lj}$, $S_{kj} = (\log H)_{kl}R_{lj}$, $A = \frac{Q_{kj}+\lambda S_{kj}}{\lambda C}$ and $B = \frac{\eta_j}{\lambda C} - \frac{S_{kj}}{C} + 1$, and rearrange term followed by exponentiation of both sides of the above equation to obtain

$$\frac{A}{H_{kj}} = \log H_{kj} + B \Rightarrow Ae^B = \exp\left(\log\left(H_{kj}e^B\right)\right)\log\left(H_{kj}e^B\right).$$

Using the definition of the principal branch of the Lambert $\mathcal{W}$-function, this further reduces to

$$\log H_{kj}e^B = \mathcal{W}_0\left(Ae^B\right) \Rightarrow H_{kj}e^B = e^{\mathcal{W}_0(Ae^B)} \Rightarrow H_{kj}e^B = \frac{Ae^B}{\mathcal{W}_0\left(Ae^B\right)} \Rightarrow H_{kj} = \frac{A}{\mathcal{W}_0\left(Ae^B\right)},$$

where in the last step we use the property of $\mathcal{W}$ function that $e^{\mathcal{W}_0(z)} = \frac{z}{\mathcal{W}_0(z)}$. Now replacing A and B with their definitions yields Eq.(11).

Proof of Eq.(12). With the ℓ_2 regularizer, the objective function of Eq.(10) becomes

$$\sum_{kj} Q_{kj} \log H_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \frac{1}{2} \left(H_{kj} - H_{kl}\right)^2 R_{lj}. \tag{B.1}$$

Similar to the symmetric KL divergence case, we form Lagrangian with Lagrangian multiplier η_j to each equality constraint of Eq.(10), and take its derivative with regards to H_{kj} and set it to zero,

$$\frac{Q_{kj}}{H_{kj}} - \eta_j - \lambda \left[(\sum_l R_{lj}) H_{kj} - H_{kl} R_{lj} \right] = 0, \tag{B.2}$$

rearranging terms leads to a quadratic equation, as

$$\lambda (\sum_l R_{lj}) H_{kj}^2 + \left(\eta_j - \lambda (H_{kl} R_{lj})\right) H_{kj} - Q_{kj} = 0, \tag{B.3}$$

where the positive root of this equation is given by Eq.(12), and the other root is negative.

Proof of Eq.(13). For the ℓ_1 divergence regularizer, the objective function of Eq.(10) becomes

$$\sum_{kj} Q_{kj} \log H_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \left(|H_{kj} - H_{kl}|\right) R_{lj}. \tag{B.4}$$

First we form the Lagrangian by introducing a multiplier η_j for each column of H_{kj}. Since the resulting Lagrangian is separable for each column of matrix H_{kj}, we focus on terms that are relevant to one element H_{kj}, which is

$$\sum_{k} Q_{kj} \log H_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \left(|H_{kj} - H_{kl}|\right) R_{lj} - \eta_j \left(\sum_k H_{kj} - 1 \right).$$

After rearranging terms to remove the absolute value and dropping constants, we have

$$\sum_{k} Q_{kj} \log H_{kj} - \lambda \sum_{H_{kj} < H_{kl}} \left(H_{kj} - H_{kl}\right) - \lambda \sum_{H_{kl} < H_{kj}} \left(H_{kl} - H_{kj}\right) - \eta_j \sum_k H_{kj}.$$

Next, we take derivative with regards to this function, which will be in two cases. For $H_{kj} > H_{kl}$, setting the derivative with regards to H_{kj} becomes $\frac{Q_{kj}}{H_{kj}} - \lambda \sum_l R_{lj} - \eta_j = 0 \Rightarrow H_{kj} = \frac{Q_{kj}}{\eta_j + \lambda \sum_l R_{lj}}$. This holds when $H_{kj} = \frac{Q_{kj}}{\eta_j + \lambda \sum_l R_{lj}} > H_{kl} \Rightarrow \frac{Q_{kj}}{H_{kl}} - \lambda > \eta_j \sum_l R_{lj}$. To make H_{kj} nonnegative, it is easy to see that $\eta_j > -\lambda \sum_l R_{lj}$. For $H_{kj} < H_{kl}$, setting the derivative with regards to H_{kj} becomes $\frac{Q_{kj}}{H_{kj}} + \lambda \sum_l R_{lj} - \eta_j = 0 \Rightarrow H_{kj} = \frac{Q_{kj}}{\eta_j - \lambda \sum_l R_{lj}}$. This holds when $H_{kj} = \frac{Q_{kj}}{\eta_j - \lambda \sum_l R_{lj}} < H_{kl} \Rightarrow \frac{Q_{kj}}{H_{kl}} + \lambda \sum_l R_{lj} < \eta_j$. When these two conditions are not satisfied, the optimal solution is given by setting $H_{kj} = H_{kl}$. Combining all these results yields Eq.(13). □

Appendix C.

Proof of Eq.(18). We first simplify the objective function of Eq.(17) by dropping irrelevant constant terms to obtain the new objective function as

$$\sum_{kj} Q_{kj} \log H_{kj}^{(l)} - \lambda \sum_{kj} \left(H_{kj}^{(l)} \log \frac{H_{kj}^{(l)}}{H_{jk}^{(\backslash l)}} - H_{jk}^{(\backslash l)} \log H_{kj}^{(l)} \right)$$

Next, we introduce Lagrangian multiplier η_j to each equality constraint on the column of $H^{(l)}$ in Eq.(17) and form the Lagrangian. Ignoring the nonnegative constraint (it will be shown to be satisfied automatically later), the first KKT condition of the problem with regards to $H_{kj}^{(l)}$ is that the derivative of the Lagrangian with regards to $H_{kj}^{(l)}$ vanishes, as

$$\frac{Q_{kj} + \lambda H_{kj}^{(\backslash l)}}{H_{kj}^{(l)}} + \lambda \log \frac{H_{kj}^{(\backslash l)}}{H_{kj}^{(l)}} - \lambda - \eta_j = 0.$$

We define $A = \frac{Q_{kj}}{\lambda} + H_{kj}^{(\backslash l)}$ and $B = 1 - \log H_{kj}^{(\backslash l)} + \frac{\eta_j}{\lambda}$, and rearrange term followed by exponentiation of both sides of the above equation to obtain

$$\frac{A}{H_{kj}^{(l)}} = \log H_{kj}^{(l)} + B \Rightarrow Ae^B = \exp\left(\log\left(H_{kj}^{(l)} e^B\right)\right) \log\left(H_{kj}^{(l)} e^B\right)$$

Using the definition of the principal branch of the Lambert $\mathcal{W}$-function, this further reduces to

$$\log H_{kj}^{(l)} e^B = \mathcal{W}_0\left(Ae^B\right) \Rightarrow H_{kj}^{(l)} e^B = e^{\mathcal{W}_0(Ae^B)} \Rightarrow H_{kj}^{(l)} e^B = \frac{Ae^B}{\mathcal{W}_0\left(Ae^B\right)} \Rightarrow H_{kj}^{(l)} = \frac{A}{\mathcal{W}_0\left(Ae^B\right)},$$

where in the last step we use the property of $\mathcal{W}$ function that $e^{\mathcal{W}_0(z)} = \frac{z}{\mathcal{W}_0(z)}$. Now replacing A and B with their definitions yields Eq.(18).

Proof of Eq.(19). With the ℓ_2 co-regularizer, the objective function of Eq.(17) becomes

$$\sum_{kj} Q_{kj} \log H_{kj}^{(l)} - \lambda \sum_{kj} \tfrac{1}{2} \left(H_{kj}^{(l)} - H_{kj}^{(\backslash l)}\right)^2$$

Similar to the symmetric KL divergence case, we form Lagrangian with Lagrangian multiplier η_j to each equality constraint of Eq.(17), and take its derivative with regards to $H_{kj}^{(l)}$ and set it to zero,

$$\frac{Q_{kj}}{H_{kj}^{(l)}} - \lambda(H_{kj}^{(l)} - H_{kj}^{(\backslash l)}) - \eta_j = 0.$$

Rearranging terms, this leads to a quadratic equation

$$\lambda \left(H_{kj}^{(l)}\right)^2 - \left(\lambda H_{kj}^{(\backslash l)} - \eta_j\right) H_{kj}^{(l)} - Q_{kj} = 0,$$

the positive root of which is given by Eq.(19) (the other root is negative).

Proof of Eq.(20). For the ℓ_1 divergence co-regularizer, the objective function of Eq.(17) becomes

$$\sum_{kj} Q_{kj} \log H_{kj}^{(l)} - \lambda \sum_{kj} \left|H_{kj}^{(l)} - H_{kj}^{(\backslash l)}\right|.$$

First we form the Lagrangian by introducing a multiplier η_j for each column of $H^{(l)}$. Since the resulting Lagrangian is separable for each column of matrix $H^{(l)}$, we focus on terms that are relevant to one element $H_{kj}^{(l)}$, which is

$$\sum_{k} Q_{kj} \log H_{kj}^{(l)} - \lambda \sum_{k} \left|H_{kj}^{(l)} - H_{kj}^{(\backslash l)}\right| - \eta_j \left(\sum_k H_{kj}^{(l)} - 1\right).$$

After rearranging terms to remove the absolute value and dropping constants, we have

$$\sum_{k} Q_{kj} \log H_{kj}^{(l)} - \lambda \sum_{H_{kj}^{(l)} < H_{kj}^{(\backslash l)}} \left(H_{kj}^{(l)} - H_{kj}^{(\backslash l)}\right) - \lambda \sum_{H_{kj}^{(l)} < H_{kj}^{(\backslash l)}} \left(H_{kj}^{(\backslash l)} - H_{kj}^{(l)}\right) - \eta_j \sum_k H_{kj}^{(l)}.$$

Next, we can take derivative with regards to this function, which will be in two cases. For $H_{kj}^{(l)} > H_{kj}^{(\backslash l)}$, setting the derivative with regards to $H_{kj}^{(l)}$ becomes $\frac{Q_{kj}}{H_{kj}^{(l)}} - \lambda - \eta_j = 0 \Rightarrow H_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j + \lambda}$. This holds when $H_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j + \lambda} > H_{kj}^{(\backslash l)} \Rightarrow \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} - \lambda > \eta_j$. To make $H_{kj}^{(l)}$ nonnegative, it is easy to see that $\eta_j > -\lambda$. For $H_{kj}^{(l)} < H_{kj}^{(\backslash l)}$, setting the derivative with regards to $H_{kj}^{(l)}$ becomes $\frac{Q_{kj}}{H_{kj}^{(l)}} + \lambda - \eta_j = 0 \Rightarrow H_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j - \lambda}$. This holds when $H_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j - \lambda} < H_{kj}^{(\backslash l)} \Rightarrow \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} + \lambda < \eta_j$. When these two conditions are not satisfied, the optimal solution is given by setting $H_{kj}^{(l)} = H_{kj}^{(\backslash l)}$. Combining all these results yields Eq.(20). □

References

[1] T. Hofmann, Probabilistic latent semantic analysis, in: UAI, 1999, pp. 289–296.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, in: J. Mach. Learn. Res., Vol. 3, JMLR, 2003, pp. 993–1022.

[3] X. Wang, X. Ma, E. Grimson, Unsupervised activity perception by hierarchical bayesian models, in: CVPR, 2007.

[4] S. Nikolopoulos, S. Zafeiriou, I. Patras, I. Kompatsiaris, High order PLSA for indexing tagged images, Signal Processing 93 (8) (2013) 2212–2228.
[5] T. Hospedales, S. Gong, T. Xiang, A Markov clustering topic model for mining behaviour in video.
[6] J. Sivic, B. C. Russell, A. Zisserman, I. Ecole, N. Supérieure, Unsupervised discovery of visual object class hierarchies, in: In CVPR, 2008.
[7] D. Küttel, M. D. Breitenstein, L. J. V. Gool, V. Ferrari, What s going on? discovering spatio-temporal dependencies in dynamic scenes, in: CVPR, 2010.
[8] J. Li, S. Gong, T. Xiang, Global behaviour inference using probabilistic latent semantic analysis, in: BMVC, 2008, pp. 1–10.
[9] L. Fei-fei, A bayesian hierarchical model for learning natural scene categories, in: In CVPR, 2005, pp. 524–531.
[10] J. C. Niebles, H. Wang, L. Fei-fei, Unsupervised learning of human action categories using spatial-temporal words, in: In Proc. BMVC, 2006.
[11] J. Varadarajan, J.-M. Odobez, Topic models for scene analysis and abnormality detection, in: 9th International Workshop in Visual Surveillance, 2009.
[12] X. Wang, K. T. Ma, G.-W. Ng, W. E. L. Grimson, Trajectory analysis and semantic region modeling using a nonparametric bayesian model, in: CVPR, 2008.
[13] X. Wei, W. B. Croft, LDA-based document models for ad-hoc retrieval, in: SIGIR, 2006.
[14] R. Nallapati, A. Ahmed, E. P. Xing, W. W. Cohen, Joint latent topic models for text and citations., in: KDD, 2008, pp. 542–550.
[15] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004.
[16] Y. Liu, A. Niculescu-Mizil, W. Gryc, Topic-link LDA: Joint models of topic and author community, in: ICML, 2009.
[17] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, in: Proceeding of the 17th international conference on World Wide Web, 2008, pp. 101–110.
[18] D. Cai, X. He, J. Han, T. S. Huang, Graph regularized non-negative matrix factorization for data representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2011) 1548–1560.
[19] E. Gaussier, C. Goutte, Relation between PLSA and NMF and implications, in: SIGIR, 2005, pp. 601–602.
[20] C. Ding, T. Li, W. Peng, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing factorization, Computational Statistics and Data Analysis 52 (2008) 3913–3927.
[21] Y. Jiang, J. Liu, Z. Li, P. Li, H. Lu, Co-regularized PLSA for multi-view clustering, in: ACCV, 2012.
[22] Y. Jiang, J. Liu, Z. Li, H. Lu, Collaborative PLSA for multi-view clustering, in: International Conference on Pattern Recognition, 2012.
[23] X. Wang, M.-C. Chang, Y. Ying, S. Lyu, Co-regularized PLSA for multi-modal learning, in: AAAI, 2016.
[24] www.aaai.org.
[25] J. G. Jialu Liu, Chi Wang, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: SIAM Data Mining Symposium, 2013.
[26] X. He, M.-Y. Kan, P. Xie, X. Chen, Comment-based multi-view clustering of web 2.0 items, in: International Conference on World Wide Web, 2014.
[27] J. Chang, D. Blei, Relational topic models for document networks, in: Artificial Intelligence and Statistics, 2009.
[28] S. Virtanen, Y. Jia, A. Klami, T. Darrell, Factorized multi-modal topic model, in: UAI, 2012.
[29] D. Cohn, T. Hofmann, The missing link - a probabilistic model of document content and hypertext connectivity, in: Advances in Neural Information Processing Systems, 2001.
[30] Y. Jia, M. Salzmann, T. Darrell, Learning cross-modality similarity for multinomial data, in: Proceedings of the 2011 International Conference on Computer Vision, ICCV, 2011.
[31] D. Zhang, Q. Mei, C. Zhai, Cross-lingual latent topic extraction., in: ACL, 2010, pp. 1128–1137.
[32] J. Boyd-Graber, D. M. Blei, Multilingual topic models for unaligned text, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, 2009, pp. 75–82.
[33] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 521–535.
[34] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM International Conference on Multimedia, 2010, pp. 251–260.
[35] X. Mao, B. Lin, Cai, X. He, J. Pei, Parallel field alignment for cross media retrieval, in: Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 897–906.
[36] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Advances in Neural Information Processing Systems 25, 2012, pp. 2231–2239.
[37] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: International Conference on Machine Learning, 2011.
[38] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B 39 (1) (1977) 1–38.
[39] R. Neal, G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, 1998, pp. 355–368.
[40] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, D. E. Knuth, On the Lambert W function, in: Advances in Computational Mathematics, 1996, pp. 329–359.
[41] M. Brand, Pattern discovery via entropy minimization., in: AISTATS, 1999.
[42] M. Shashanka, B. Raj, P. Smaragdis, Sparse overcomplete latent variable decomposition of counts data, in: J. Platt, D. Koller, Y. Singer, S. Roweis (Eds.), Advances in Neural Information Processing Systems 20, 2007, pp. 1313–1320.
[43] I. Khan, A. Saffari, H. Bischof, TVGraz: Multi-modal learning of object categories by combining textual and visual features, in: AAPR Workshop, 2009, pp. 213–224.
[44] European parliament proceedings parallel corpus (EPPPC), http://www.statmt.org/europarl/.