Document Clustering and Topic Modeling: A Unified Bayesian Probabilistic Perspective

Gianni Costa
ICAR-CNR
87036 Rende (CS), ITALY
costa@icar.cnr.it

Riccardo Ortale
ICAR-CNR
87036 Rende (CS), ITALY
ortale@icar.cnr.it

***Abstract*—Document clustering and topic modeling are fundamental tasks in text mining, that can be unified to reciprocally enhance each other. In this paper, we present a machine learning approach to the joint modeling and interdependent fulfilment of both tasks. In particular, document clustering and topic modeling are seamlessly interrelated under an innovative Bayesian generative model of clusters, topics and contents in text corpora. Such a model assumes that text corpora result from a generative process, in which clusters and topics act as connected latent factors. Essentially, clusters are initially associated with descriptive and actionable topic distributions, that enforce cluster coherence. The individual documents are then assigned to one respective cluster and worded accordingly. Under the devised model, document clustering and topic modeling can be simultaneously performed in an interdependent manner simply by Bayesian reasoning. For this purpose, the mathematical details regarding collapsed Gibbs sampling as well as parameter estimation are derived and implemented into an approximate inference algorithm. Comparative experiments on standard benchmark text corpora reveal the effectiveness of our approach at jointly clustering text documents and unveiling their semantics in terms of coherent topics.**

***Index Terms*—Document Clustering, Topic Modeling, Bayesian Text Analysis**

I. Introduction

Document clustering and topic modeling can be interrelated as components of a unified process [2]. In such a process, each task addresses the limitations of the other one, so that both synergically enhance each other.

Indeed, on one hand, topic modeling uncovers text semantics. This is accomplished by representing the individual documents as mixtures of topics and, in turn, topics as corresponding word rankings. Clustering can be used, in the context of topic modeling, in order to capture and summarize the semantics of portions of text corpora, in terms of the specific topics covered in the respective documents.

On the other hand, document clustering divides a text corpus into cohesive and well-separated groups. Thus, documents within the same group are homogeneous, whereas documents from distinct groups are heterogeneous. However, if homogeneity involves merely syntactic regularities, clusters may lack semantic coherence and, accordingly, be neither intuitively intelligible nor useful in practical applications. Moreover, the adoption of the *bag-of-words* model for processing raw text is likely to further deteriorate cluster quality. Indeed, the latter is affected by the noise arising from the very large dimensionality of the resulting document representation (due to vocabulary size) and its sparsity (which is exacerbated with short text documents). Topic modeling can be used, in the context of document clustering, for conveniently representing text documents in a lower-dimensional semantic space of intelligible topics. This allows for a more effective separation of a text corpus into semantically-coherent clusters, along with a clear explanation of the semantics of both the individual documents and the discovered clusters.

The integration of document clustering and topic modeling is a challenging research issue. As a matter of fact, a reasonable interpenetration and mutuality between both tasks have to be devised. Ideally, the envisaged interaction should ultimately turn each task into an effective expedient for the synergic enhancement of the other one.

In this paper, we present an innovative approach to tightly coupling document clustering and topic modeling. Our approach is built on ideas drawn from Bayesian statistics, probabilistic graphical modeling, generative and latent-factor modeling [5], [19], [25].

In particular, a new generative model of text corpora, called DETECTOR (*DocumEnt clusTErs and topiCs in Text cOrpoRa*), is proposed to explain and reason on document wording from a Bayesian probabilistic perspective. DETECTOR hypothesizes that text documents are the result of a generative process. In the latter, document clustering and topic modeling are seamlessly coupled to act jointly. Essentially, clusters are associated with corresponding topic distributions, which enforce intra-cluster coherence. Moreover, a multinomial probability distribution is used to pick the individual clusters. In this setting, the wording of text documents consists of two steps. At first, the generic text document is assigned to a cluster. Subsequently, its textual content is generated from the topical distribution associated with that cluster.

The generative process envisaged by DETECTOR has two appealing peculiarities.

Firstly, the explicit summarization of document clusters into respective topic distributions provides succinct and clear descriptions of such clusters. In principle, these also allow for querying the DETECTOR model. This is of great practical relevance, while retrieving documents from very large text corpora by pertinency to query documents.

Secondly, uncertainty is dealt with probabilistically. Therein, text documents are treated as the only type of observed

2375-0197/19/$31.00 ©2019 IEEE
DOI 10.1109/ICTAI.2019.00047

data. Instead, all other aspects are viewed as latent factors. In particular, these include the membership of text documents to clusters, the topical distributions of such clusters, the individual topics as well as the topical annotations of document words. The conditional (in)dependencies between observed data and latent factors as well as those among latent factors are defined through probabilistic graphical modeling. Accordingly, the cluster membership of documents is considered as a discrete random variable, drawn from the multinomial distribution over clusters. Likewise, the topical annotations of document words are represented as discrete random variables, sampled from the multinomial distributions over cluster topics. Text documents are sequences of discrete random variables, that are realized by drawing from the multinomial distributions over topic words. Besides, all multinomial distributions are sampled from respective conjugate Dirichlet priors.

The values of all latent random variables under DETECTOR are learnt through the powerful machinery of Bayesian inference [16], [17], [29]. In particular, collapsed Gibbs sampling is exploited to infer the posterior distributions over the involvement of documents in clusters as well as the assignments of topical annotations to document words. Parameter estimation is employed to quantify cluster probabilities along with the topic distributions of the individual clusters and documents. Together, posterior inference and parameter estimation amount to jointly performing document clustering and topic modeling in an interdependent manner.

Comparative experiments are conducted over standard benchmark data sets to study the performance of DETECTOR. The empirical evidence demonstrates its effectiveness in clustering a corpus of text documents and, also, unveiling their semantics in terms of coherent topics.

This paper proceeds as follows. Notation and preliminaries are introduced in Section II. The DETECTOR model is developed in Section III. Collapsed Gibbs sampling and parameter estimation are the focus of Section IV. The empirical evaluation of DETECTOR is covered in Section V. Finally, Section VI concludes and highlights future research.

II. Preliminaries

Let $\boldsymbol{V}$ be a vocabulary of V words, i.e., $\boldsymbol{V} \triangleq \{w_1, \ldots, w_V\}$. A text corpus $\boldsymbol{D}$ is a collection of D text documents, namely, $\boldsymbol{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_D\}$. The generic document $\boldsymbol{d} \in \boldsymbol{D}$ is a sequence of words from $\boldsymbol{V}$. Assume that $n_{\boldsymbol{d}}$ is the length of $\boldsymbol{d}$. The latter is formalized as $\boldsymbol{d} \triangleq \{w_{\boldsymbol{d},1}, \ldots, w_{\boldsymbol{d},n_{\boldsymbol{d}}} | w_{\boldsymbol{d},n} \in \boldsymbol{V} \text{ with } n = 1, \ldots, n_{\boldsymbol{d}}\}$. $\boldsymbol{D}$ is inherently characterized by two latent (or hidden) properties, namely, the underlying topics and the consequent document arrangement into semantically coherent clusters.

The semantics of $\boldsymbol{D}$ is assumed to span T topics $\boldsymbol{\beta}_t$, with $t = 1, \ldots, T$. Such topics are defined in Section III as Dirichlet distributions over $\boldsymbol{V}$. Accordingly, for each topic t and each word w in $\boldsymbol{V}$, $\beta_{t,w}$ is the probability of w in t. $\boldsymbol{\beta} \triangleq \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}$ is the set of all topics in $\boldsymbol{D}$.

Document words are contextualized by topics. To this end, $\boldsymbol{z_d}$ is the sequence of annotations for the words of $\boldsymbol{d}$. More precisely, $\boldsymbol{z_d} \triangleq \{z_{\boldsymbol{d},1} \ldots, z_{\boldsymbol{d},n_{\boldsymbol{d}}}\}$, where $z_{\boldsymbol{d},n}$ is a topic in the interval $[1, T]$, that contextualizes the meaning of the corresponding word $w_{\boldsymbol{d},n}$. Notation $\boldsymbol{Z} = \{\boldsymbol{z_d} | \boldsymbol{d} \in \boldsymbol{D}\}$ indicates the contextualization of the whole text corpus $\boldsymbol{D}$.

Topics are covered to a different extent within K disjoint clusters $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K$. Such clusters partition $\boldsymbol{D}$ into semantically coherent groups, that essentially absorb and reflect the topical regularities across the text corpus. The semantics of each cluster $\boldsymbol{C}_k$ is the topic mixture $\boldsymbol{\theta}_k \triangleq \{\theta_{k,1}, \ldots, \theta_{k,T}\}$, with $\theta_{k,t}$ being the degree to which topic t is covered in $\boldsymbol{C}_k$. Notation $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ groups the semantics of all clusters $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K$.

The generic document $\boldsymbol{d}$ belongs to one cluster among $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K$. The cluster membership of $\boldsymbol{d}$ is denoted by $c_{\boldsymbol{d}}$, that takes on a value in the range $[1, K]$. The cluster memberships of all documents in $\boldsymbol{D}$ are jointly marked as $\boldsymbol{C}$, i.e., $\boldsymbol{C} \triangleq \{c_{\boldsymbol{d}} | \boldsymbol{d} \in \boldsymbol{D}\}$. For any $k = 1, \ldots, K$, π_k is the probability that cluster $\boldsymbol{C}_k$ contains $\boldsymbol{d}$, i.e., $\Pr(c_{\boldsymbol{d}} = k) = \pi_k$. In addition, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ is the probability distribution over all clusters $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K$.

The semantics of documents within each cluster conforms with the semantics of the particular cluster. In this respect, the semantics of $\boldsymbol{d}$ is characterized as the topic mixture $\boldsymbol{\theta_d} \triangleq \{\theta_{\boldsymbol{d},1}, \ldots, \theta_{\boldsymbol{d},T}\}$, with $\theta_{\boldsymbol{d},t}$ being the degree to which $\boldsymbol{d}$ covers topic t. The details on the computation of $\boldsymbol{\theta_d}$ are provided in Section IV.

A. Problem statement

Modeling and clustering a text corpus $\boldsymbol{D}$ involves performing the below tasks jointly and in an interdependent manner:

- *topic modeling*, i.e., the act of learning $\boldsymbol{\beta}_t$ (for each topic $t = 1, \ldots, T$), $\boldsymbol{\theta}_k$ (for each cluster $\boldsymbol{C}_k$ with $k = 1, \ldots, K$), as well as $\boldsymbol{z_d}$ and $\boldsymbol{\theta_d}$ (for each document $\boldsymbol{d}$ in $\boldsymbol{D}$);
- *document clustering*, i.e., the act of determining $\boldsymbol{\pi}$ as well as $c_{\boldsymbol{d}}$ for each document $\boldsymbol{d}$ in $\boldsymbol{D}$.

In Section III, topic modeling and document clustering are jointly modeled as connected causes, that latently determine the wording of text corpora. In Section IV, we derive and implement posterior inference along with parameter estimation to simultaneously uncover such interrelated causes.

III. The DETECTOR Model

DETECTOR (*DocumEnt clusTErs and topiCs in Text cOrpoRa*) is a generative model of text corpora, their underlying topics and clusters. Under DETECTOR, any text corpus $\boldsymbol{D}$ is assumed to be a collection of observations resulting from a hypothetical generative process, that is governed by $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{\pi}$, $\boldsymbol{C}$ and Θ. In such a generative process, all elements of $\boldsymbol{C}$, $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{\pi}$ and Θ are treated as random variables. Moreover, the individual components of $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{\pi}$, $\boldsymbol{C}$ and Θ are regarded as latent factors, i.e., entities whose values are *apriori* unknown and not directly measurable. The conditional (in)dependencies postulated by DETECTOR among the random variables of $\boldsymbol{D}$, $\boldsymbol{C}$, $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{\pi}$ and Θ are shown, through plate notation, in the directed graphical representation of Figure 1. Notice that

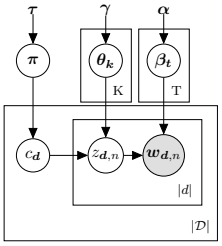

Fig. 1: Graphical representation of DETECTOR

the observed random variables are reported as shaded nodes, whereas the latent random variables appear as unshaded nodes.

In accordance with the conditional (in)dependencies of Figure 1, the generative process under DETECTOR operates by performing the realization of the involved random variables, as described in Figure 2.

1) For each $t = 1, \ldots, T$, draw the topic $\boldsymbol{\beta}_t$, i.e., $\boldsymbol{\beta}_t \sim$ *Dirichlet*$(\boldsymbol{\alpha})$.
2) Draw cluster distribution $\boldsymbol{\pi} \sim$ *Dirichlet*$(\boldsymbol{\tau})$
3) For each $k = 1, \ldots, K$, draw the topical description of cluster $\boldsymbol{C}_k$, i.e., $\boldsymbol{\theta}_k \sim$ *Dirichlet*$(\boldsymbol{\gamma})$.
4) For each document $\boldsymbol{d}$ in $\boldsymbol{D}$
   a) Draw cluster membership $c_{\boldsymbol{d}} \sim$ *Discrete*$(\boldsymbol{\pi})$;
   b) For each $n = 1, \ldots, n_{\boldsymbol{d}}$
      i) draw topic assignment for the n-th word, i.e., $z_{\boldsymbol{d},n} \sim$ *Discrete*$(\boldsymbol{\theta}_{c_{\boldsymbol{d}}})$;
      ii) draw the n-th word, i.e., $w_{\boldsymbol{d},n} \sim$ *Discrete*$(\boldsymbol{\beta}_{z_{\boldsymbol{d},n}})$.

Fig. 2: The generative process envisaged by DETECTOR

Basically, at step 1, the latent topics $\boldsymbol{\beta}_t$ with $t = 1, \ldots, T$ are sampled from a Dirichlet prior with hyperparameter $\boldsymbol{\alpha}$.

At step 2, the distribution $\boldsymbol{\pi}$ over clusters $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K$ is sampled from a Dirichlet prior with hyperpameter $\boldsymbol{\tau}$.

At step 3, the semantics of each cluster $\boldsymbol{C}_k$ with $k = 1, \ldots, K$ is sampled from a related Dirichlet prior with hyperpameter $\boldsymbol{\gamma}$. Interestingly, associating each cluster with its own semantics is expectedly beneficial to enforce intra-cluster coherence. Furthermore, in principle, this allows for querying the individual clusters, which enables applications of information retrieval.

At step 4, the generic document $\boldsymbol{d}$ is eventually generated from the semantics $\boldsymbol{\theta}_{c_{\boldsymbol{d}}}$ of its membership cluster $c_{\boldsymbol{d}}$. More precisely, at step 4a, $\boldsymbol{d}$ is placed into cluster $c_{\boldsymbol{d}}$, which is chosen from $\boldsymbol{\pi}$. Consequently, at step 4(b)ii, each word $w_{\boldsymbol{d},n}$ in $\boldsymbol{d}$ is sampled from the topic $z_{\boldsymbol{d},n}$, that is in turn drawn from the cluster semantics $\boldsymbol{\theta}_{c_{\boldsymbol{d}}}$ at the preceding step 4(b)i. Remarkably, the relationship between $\boldsymbol{\theta}_{c_{\boldsymbol{d}}}$ and $c_{\boldsymbol{d}}$ enables the unified and seamless modeling of document clustering and topic modeling. Besides, the envisaged wording of documents implies the document likelihood $\Pr(\boldsymbol{d}|\boldsymbol{z}_{\boldsymbol{d}}, \boldsymbol{\beta})$, which is defined beneath

$$\Pr(\boldsymbol{d}|\boldsymbol{z}_{\boldsymbol{d}}, \boldsymbol{\beta}) \triangleq \prod_{n=1}^{n_{\boldsymbol{d}}} \prod_{w \in \boldsymbol{V}} \beta_{t,w}^{n_{\boldsymbol{d},t}^{(w)}} \quad (1)$$

with $n_{\boldsymbol{d},t}^{(w)}$ being the number of times that word w occurs in document $\boldsymbol{d}$ under topic t.

Due to the conditional (in)dependencies of Figure 1, the joint distribution over both the observed documents and the latent factors factorizes, under DETECTOR, as follows

$$\begin{aligned}\Pr(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = &\Pr(\boldsymbol{D}|\boldsymbol{Z}, \boldsymbol{\beta}) \cdot \Pr(\boldsymbol{\beta}|\boldsymbol{\alpha}) \cdot \\ &\Pr(\boldsymbol{Z}|\boldsymbol{C}, \boldsymbol{\Theta}) \cdot \Pr(\boldsymbol{\Theta}|\boldsymbol{\gamma}) \cdot \\ &\Pr(\boldsymbol{C}|\boldsymbol{\pi}) \cdot \Pr(\boldsymbol{\pi}|\boldsymbol{\tau}) \quad (2)\end{aligned}$$

where

- $\Pr(\boldsymbol{\pi}|\boldsymbol{\tau}) \triangleq$ *Dirichlet*$(\boldsymbol{\pi}|\boldsymbol{\tau}) = \frac{1}{\Delta(\boldsymbol{\tau})} \prod_{k=1}^{K} \pi_k^{\tau_k - 1}$;
- $\Pr(\boldsymbol{\Theta}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} \Pr(\boldsymbol{\theta}_k|\boldsymbol{\gamma})$ and $\Pr(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \triangleq$ *Dirichlet*$(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) = \frac{1}{\Delta(\boldsymbol{\gamma})} \prod_{t=1}^{T} \theta_{k,t}^{\gamma_t - 1}$;
- $\Pr(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \prod_{t=1}^{T} \Pr(\boldsymbol{\beta}_t|\boldsymbol{\alpha})$ and $\Pr(\boldsymbol{\beta}_t|\boldsymbol{\alpha}) \triangleq$ *Dirichlet*$(\boldsymbol{\beta}_t|\boldsymbol{\alpha}) = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{w \in \boldsymbol{V}} \beta_{t,w}^{\alpha_w - 1}$;
- $\Pr(\boldsymbol{C}|\boldsymbol{\pi}) \triangleq \prod_{k=1}^{K} \pi_k^{n^{(k)}}$, with $n^{(k)}$ being the number of times that cluster k is chosen as document membership;
- $\Pr(\boldsymbol{Z}|\boldsymbol{C}, \boldsymbol{\Theta}) \triangleq \prod_{k=1}^{K} \prod_{t=1}^{T} \theta_{k,t}^{n_k^{(t)}}$, with $n_k^{(t)}$ being the number of times that topic t is treated in cluster k;
- $\Pr(\boldsymbol{D}|\boldsymbol{Z}, \boldsymbol{\beta}) \triangleq \prod_{\boldsymbol{d} \in \boldsymbol{D}} \Pr(\boldsymbol{d}|\boldsymbol{z}_{\boldsymbol{d}}, \boldsymbol{\beta})$, with $\Pr(\boldsymbol{d}|\boldsymbol{z}_{\boldsymbol{d}}, \boldsymbol{\beta})$ being the document likelihood formalized by Equation 1.

IV. Bayesian Inference and Parameter Estimation

DETECTOR interrelates document clustering and topic modeling as latent factors in the generation of text corpora. Uncovering such factors simultaneously involves reversing the generative process under DETECTOR by Bayesian reasoning. To this end, posterior inference is derived for going back to the latent random variables in $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{\pi}$, $\boldsymbol{C}$ and Θ, given a text corpus $\boldsymbol{D}$, through a posterior distribution $\Pr(\boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{C}, \Theta|\boldsymbol{D})$. Unfortunately, as it usually happens with Bayesian models of practical relevance, exact posterior inference is intractable under DETECTOR. For this reason, we resort to Gibbs sampling, namely, a *MCMC* technique for stochastic approximate inference [3], [26], that often enables straightforward approximate inference algorithms, despite a potentially large number of latent random variables [18].

In particular, by taking advantage of the multinomial/Dirichlet conjugacy, we perform stochastic approximate inference via a collapsed Gibbs sampling algorithm, whose pseudo code is outlined in Algorithm 1. Collapsing involves the marginalization of the joint distribution $\Pr(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ of Equation 2 with respect to the random variables $\boldsymbol{\beta}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$. This results in a marginalized joint distribution $\Pr(\boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{D}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, that is formalized below.

Algorithm 1 Collapsed Gibbs sampling

Input: a document corpus $\boldsymbol{D}$;
the number K of underlying clusters;
the number T of latent topics;
the number I of sampling iterations;
Output: the individual random variables in $\boldsymbol{Z}$ and $\boldsymbol{C}$.

1: randomly assign topics to words and documents to clusters;
2: **for all** $i = 1, \ldots, I$ **do**
3: **for all** $\boldsymbol{d} \in \boldsymbol{D}$ **do**
4: **for all** $n = 1, \ldots, n_{\boldsymbol{d}}$ **do**
5: sample $z_{\boldsymbol{d},n}$ via Equation 4 of Figure 3;
6: **end for**
7: sample $c_{\boldsymbol{d}}$ via Equation 5 of Figure 3;
8: **end for**
9: **end for**

$$\begin{aligned}\Pr(\boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{D} | \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \int\int\int \Pr(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{\Theta} \\ &= \frac{\Delta(\boldsymbol{n} + \boldsymbol{\tau})}{\Delta(\boldsymbol{\tau})} \cdot \prod_{t=1}^{T} \frac{\Delta(\boldsymbol{n_t} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \cdot \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{n}_k + \boldsymbol{\gamma})}{\Delta(\boldsymbol{\gamma})}\end{aligned} \tag{3}$$

where

- $\boldsymbol{n} \triangleq \{n^{(k)}\}_{k=1}^{K}$, with $n^{(k)}$ being the count introduced in the definition of the joint probability $\Pr(\boldsymbol{C} | \boldsymbol{\pi})$ over all cluster memberships, that appears on the right hand side of Equation 2;
- $\boldsymbol{n}_t \triangleq \{n_t^{(w)}\}_{w \in \boldsymbol{V}}$, with $n_t^{(w)}$ being the count of the number of times that word w occurs under topic t;
- $\boldsymbol{n}_k \triangleq \{n_k^{(t)}\}_{t=1}^{T}$, with $n_k^{(t)}$ being the count introduced in the definition of the joint probability $\Pr(\boldsymbol{Z} | \boldsymbol{C}, \boldsymbol{\Theta})$ over the topic contextualization, that appears on the right hand side of Equation 2.

Equation 3 explicitly accounts for the inherent uncertainty in the values of the random variables of $\boldsymbol{\beta}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$. Indeed, all possible values of such random variables are suitably considered through integration. Additionally, marginalization is also beneficial to expedite sampling, since only the values of the latent variables in $\boldsymbol{Z}$ and $\boldsymbol{C}$ are to be drawn sequentially and singly from corresponding full conditionals at lines 5 and 7 of Algorithm 1. Instead, unlike conventional Gibbs sampling, the values of the marginalized random variables in $\boldsymbol{\beta}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$ are ignored throughout the whole sampling stage of Algorithm 1 and, lastly, calculated by means of algebraic manipulations for parameter estimation.

As far as the mathematical details are concerned, the full conditional at line 5 of Algorithm 1 is a probability distribution over the corresponding latent random variable $z_{\boldsymbol{d},n}$, given the (values of all) other latent random variables and the whole text corpus $\boldsymbol{D}$. The definition of such a full conditional is provided by Equation 4 of Figure 3, where notation $\boldsymbol{Z}_{\neg(\boldsymbol{d},n)}$ indicates all topics assigned to words but the one corresponding to the n-th word within document $\boldsymbol{d}$. Moreover, α_w is the component of the hyperparameter $\boldsymbol{\alpha}$ corresponding to word w. Similarly, γ_t is the component of the hyperparameter $\boldsymbol{\gamma}$ corresponding to topic t.

Likewise, the full conditional at line 7 of Algorithm 1 is a probability distribution over the respective latent random variable $z_{\boldsymbol{d},n}$, given the values of the remaining latent random variables and the text corpus $\boldsymbol{D}$. Equation 5 of Figure 3 provides a definition of the latter full conditional, in which notation $\boldsymbol{C}_{\neg \boldsymbol{d}}$ denotes all cluster memberships of the individual documents but the cluster membership of document $\boldsymbol{d}$. Besides, $\Delta(\cdot)$ represents the Dirichlet delta function [18]. In addition, τ_k stands for the component of hyperparameter $\boldsymbol{\tau}$ corresponding to cluster k.

Notice that, in Algorithm 1, all counts $n_t^{(w)}$, $n_k^{(t)}$ and $n^{(k)}$ are progressively updated as samples are drawn.

Upon termination of Algorithm 1, such counts allow for the estimation of the values of the latent random variables in $\boldsymbol{\beta}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$. To elaborate, we point out that $P(\boldsymbol{\beta}_t | \boldsymbol{\alpha}, \boldsymbol{D}, \boldsymbol{Z}) \propto \textit{Dirichlet}(\boldsymbol{\beta}_t | \boldsymbol{n}_t + \boldsymbol{\alpha})$. Furthermore, $P(\boldsymbol{\theta}_k | \boldsymbol{\gamma}, \boldsymbol{Z}, \boldsymbol{C}) \propto \textit{Dirichlet}(\boldsymbol{\theta}_k | \boldsymbol{n}_k + \boldsymbol{\gamma})$. Yet, $P(\boldsymbol{\pi} | \boldsymbol{\tau}, \boldsymbol{C}) \propto \textit{Dirichlet}(\boldsymbol{\pi} | \boldsymbol{n} + \boldsymbol{\tau})$. Therefore, by exploiting the mean of the Dirichlet distribution, the values of the arbitrary latent random variables $\boldsymbol{\beta}_t$, $\boldsymbol{\theta}_k$ and $\boldsymbol{\pi}_k$ can be computed through Equations 6, 7 and 8 of Figure 4.

Lastly, we observe that the semantics $\boldsymbol{\theta}_{\boldsymbol{d}}$ of the generic document $\boldsymbol{d}$ can be estimated as reported below

$$\theta_{\boldsymbol{d},t} = \frac{n_{\boldsymbol{d}}^{(t)}}{\sum_{t'=1}^{T} n_{\boldsymbol{d}}^{(t')}}$$

where $n_{\boldsymbol{d}}^{(t)}$ is the number of times that topic t occurs in document $\boldsymbol{d}$.

V. Empirical evaluation

We conducted an experimentation of DETECTOR on standard benchmark text corpora. Tests were aimed to

- assess its effectiveness at clustering a collection of text documents and recovering the underlying topics.
- investigate whether the devised interaction between document clustering and topic modeling is a mutual enhancement between both tasks. Roughly speaking, we inspected whether each task in tandem with the other is more effective than the same task alone;
- understand whether the devised interaction between the two tasks improves their effectiveness, compared to naively pipelining both with no interpenetration.

A. Text corpora

All experiments were carried out over *20-Newsgroups* and *Reuters-21578*. These are standard benchmark document collections in the fields of topic modeling and (un)supervised text classification, that come with ground-truth categories for clustering evaluation.

In particular, *20-Newsgroups*[1] consists of $11,268$ text documents, which are partitioned into 20 groups.

[1] http://qwone.com/~jason/20Newsgroups/

$$P(z_{\boldsymbol{d},n}|\boldsymbol{Z}_{\neg(\boldsymbol{d},n)},\boldsymbol{C},\boldsymbol{D},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\gamma}) \propto \frac{n_t^{(w)}-1+\alpha_w}{\left(\sum_{w'\in\boldsymbol{V}} n_t^{(w')}+\alpha_{w'}\right)-1} \cdot \frac{n_k^{(t)}-1+\gamma_t}{\left(\sum_{t'=1}^{T} n_k^{(t')}+\gamma_{t'}\right)-1} \tag{4}$$

$$P(c_{\boldsymbol{d}}|\boldsymbol{C}_{\neg\boldsymbol{d}},\boldsymbol{D},\boldsymbol{Z},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\gamma}) \propto \frac{\Delta(\boldsymbol{n}_k+\boldsymbol{\gamma})}{\Delta(\boldsymbol{\gamma})} \cdot \frac{n^{(k)}-1+\tau_k}{\left(\sum_{k'=1}^{K} n^{(k')}+\tau_{k'}\right)-1} \tag{5}$$

Fig. 3: The full conditionals used in Algorithm 1

$$\beta_{t,w} = \frac{n_t^{(w)}+\alpha_w}{\sum_{w'\in\boldsymbol{V}} n_t^{(w')}+\alpha_{w'}} \tag{6}$$

$$\theta_{k,t} = \frac{n_k^{(t)}+\gamma_t}{\sum_{t'=1}^{T} n_k^{(t')}+\gamma_{t'}} \tag{7}$$

$$\pi_k = \frac{n^{(k)}+\alpha_k}{\sum_{k'=1}^{K} n^{(k')}+\alpha_{k'}} \tag{8}$$

Fig. 4: Equations for parameter estimation

Reuters-21578[2] includes $21,578$ text documents. These are organized in 90 imbalanced categories and each document can belong to multiple categories. Because of class imbalance and document involvement in multiple categories, we followed a common practice in preprocessing *Reuters-21578*. This consists in sampling the whole corpus, in order to retain only those text documents, that belong to one of the largest categories. Our sample retains $2,189$ text documents belonging to exactly one of the 8 largest categories.

Both text corpora were cleaned. We removed non-alphabetic characters as well as stop words. The alphabetic characters were lower-cased. Besides, we discarded all those words, whose length is below 3 characters. Also, in *Reuters-21578*, all words appearing in less than 5 text documents were ignored. In *20-Newsgroups*, words were ignored if these appear in a number of text documents lower than 10.

B. Competing approaches

We compared the performance of our approach against three classes of competitors.

The first class groups three clustering baselines, i.e., k-means, NMF and HAC [1]. k-means is one the most popular clustering algorithms, that is used in a wide variety of applicative settings. NMF is based on non-negative matrix factorization. HAC stands for Hierarchical Agglomerative Clustering, that is well-known for its effectiveness in processing data from different domains.

The second class regards topic models. We chose the seminal LDA model [7].

The third class contains approaches to the combination of document clustering and topic modeling. One such an approach is MGCTM [30], a state-of-the-art competitor in which both tasks are jointly modeled and simultaneously performed. Our approach differs from MGCTM in several respects. More precisely, from a model design perspective, MGCTM does not explicitly associate a clear description with the individual clusters, whose topical contents are thus neither immediately intelligible nor directly actionable. Also, topics under MGCTM are divided into local and global. This involves the establishment of an appropriate number for both types of topics, which is problematic and calls for non-trivial hyperparameter tuning in the absence of a general criterion. The latter issue is exacerbated by an inherent peculiarity of clusters under MGCTM, i.e., their reliance on respective local hyperparameters. Additionally, no prior distribution is placed under MGCTM on the probability distribution over clusters. On the contrary, our approach uses one undifferentiated set of topics to explicitly characterize the semantics of clusters, which makes them truly understandable and actionable. Besides, the topical descriptions of all clusters rely on one hyperparameter, which eases model tuning and learning. Furthermore, a Dirichlet prior is explicitly placed on the multinomial cluster distribution. Lastly, from an inference viewpoint, MGCTM relies on variational inference, whereas our approach adopts Gibbs sampling. In addition, we considered hybrid approaches to the combination of document clustering and topic modeling, that sequentially perform such tasks as a naive pipeline of baselines from the two above classes. In particular, in LDA $\rightarrow$ k-means and LDA $\rightarrow$ HAC, the text documents are grouped by k-means and HAC on the basis of their topics, which are previously captured through LDA. In k-means $\rightarrow$ LDA and HAC $\rightarrow$ LDA, LDA is exploited to unveil the semantics of the clusters uncovered by k-means and HAC, respectively. In NMF $\rightarrow$ LDA, LDA is used to discover the semantics of the clusters found through NMF. Table I and Table II summarize the two envisaged types of hybrid approaches.

Contrasting our approach with competitors of the first and the second classes provides evidence to empirically substantiate whether the integration of document clustering and topic modeling is actually beneficial in comparison to performing the two tasks separately. Furthermore, the comparative evaluation against baselines of the third class is accomplished for a twofold reason. Firstly, it enables an insightful comparison between our approach and a state-of-the-art competitor, i.e., MGCTM. Secondly, it provides evidence to empirically corroborate whether the integration of document clustering and topic modeling is actually beneficial in comparison to a naive pipeline of the two tasks with no interpenetration.

[2]http://www.daviddlewis.com/resources/testcollections/reuters21578/

Hybrid approach to clustering	Description
LDA → HAC	HAC partitions documents by their semantic topics, captured via LDA
LDA → k-means	k-means clusters documents by their semantic topics, captured via LDA

TABLE I: Pipelines that perform document clustering on topic modeling

Hybrid approach to topic modeling	Description
HAC → LDA	LDA discovers semantic topics in the clusters formed by HAC
k-means → LDA	LDA discovers semantic topics in the clusters formed by k-means
NMF → LDA	LDA discovers semantic topics in the clusters formed by NMF

TABLE II: Pipelines that perform topic modeling on document clustering

C. Document clustering

We assessed the performance of all tested competitors in document clustering, by evaluating their effectiveness at recovering the ground-truth categories in *20-Newsgroups* and *Reuters-21578*.

Accordingly, the number of clusters to find with k-means, DETECTOR and MGCTM was fixed to the number of true categories in *20-Newsgroups* and *Reuters-21578*.

Likewise, in NMF as well as LDA, the number of latent topics was fixed to the number of true categories behind the selected text corpora. Actually, NMF and LDA are not explicitly conceived for document clustering. Nonetheless, both competitors can be still exploited for this purpose, provided that topics are interpreted as clusters and each document is assigned to the most explicative cluster (i.e., to the one topic that best reflects the document semantics).

Following [30], the number of topics under DETECTOR as well as (the LDA stage of) LDA → k-means and LDA → HAC was set to 120 on *20-Newsgroups* and 60 on *Reuters-21578*. As far as MGCTM is concerned, we retained the original settings adopted in [30] on both datasets.

Clustering effectiveness was evaluated in terms of accuracy and normalized mutual information (NMI) [8]. Both are widely-adopted metrics ranging in the interval from 0.0 to 1.0, with higher values denoting better document clustering. The accuracy and NMI of all tested competitors on the selected text corpora are reported in Table III.

It is evident that k-means and HAC perform worse than DETECTOR and MGCTM. This finding corroborates the usefulness of topic modeling to enhance document clustering. Indeed, DETECTOR and MGCTM gain an understanding of the semantics of text documents and, thus, are more effective in their partitioning.

Nonetheless, despite the exploitation of topic modeling, LDA → k-means and LDA → HAC are less effective than DETECTOR and MGCTM. Such an evidence highlights the advantage of performing both tasks jointly, rather than as a naive pipeline of the two tasks with no mutual interaction. Actually, the lack of interplay may be detrimental for clustering effectiveness. In fact, as observed on *Reuters-21578*, LDA → k-means and LDA → HAC do not generally perform better than HAC, k-means and NMF.

Corpus	Competitor	Accuracy	NMI
20-Newsgroups	HAC	0.2580	0.2330
	k-means	0.2662	0.2412
	NMF	0.2474	0.2060
	LDA	0.3036	0.2997
	LDA → HAC	0.3243	0.3388
	LDA → k-means	0.3793	0.3993
	DETECTOR	0.4465	0.4557
	MGCTM	0.4089	0.4112
Reuters-21578	HAC	0.4454	0.1544
	k-means	0.5106	0.4401
	NMF	0.5628	0.4372
	LDA	0.6418	0.4485
	LDA → HAC	0.5756	0.3636
	LDA → k-means	0.4518	0.4461
	DETECTOR	0.7052	0.5172
	MGCTM	0.6533	0.4674

TABLE III: Clustering performance in terms of accuracy and NMI

Corpus	Competitor	SC
20-Newsgroups	LDA	-345.19
	HAC → LDA	-391.49
	k-means → LDA	-394.12
	NMF → LDA	-395.17
	DETECTOR	-332.85
	MGCTM	-341.16
Reuters-21578	LDA	-262.82
	HAC → LDA	-321.79
	k-means → LDA	-319.34
	NMF → LDA	-320.63
	DETECTOR	-246.28
	MGCTM	-257.85

TABLE IV: Topic modeling performance

NMF and LDA are less effective than DETECTOR and MGCTM because of two inherently characteristic limitations, that penalize their clustering performance. Firstly, when used for clustering, NMF and LDA gain a limited understanding of document semantics, being required to summarize the latter through as many topics as the true categories of the underlying text corpora. Secondly, under NMF and LDA, each document is clustered only on the basis of its most explicative topic, without accounting for its cross-topic similarity to the other documents of the same corpus.

Noticeably, DETECTOR outdoes all competitors in clustering effectiveness. Specifically, the overcoming performance with respect to MGCTM substantiates the rationality of the modeling choices behind DETECTOR, i.e., the explicit characterization of cluster semantics and the devised interaction between document clustering and topic modeling.

D. Topic modeling

The performance of our approach in topic modeling was investigated quantitatively and qualitatively.

1) Quantitative evaluation: In principle, various performance criteria can be considered to quantify and compare the performance of our approach in topic modeling, most notably perplexity and held-out likelihood. Although useful to

evaluate the predictive performance of all tested competitors, such criteria do not account for the coherence of the uncovered topics [30]. Therefore, we chose semantic coherence (SC), in order to assess and rank the topic modeling performance of all tested competitors. Basically, semantic coherence SC was defined as the average coherence of the inferred topics, i.e., $SC = \frac{1}{T}\sum_{t=1}^{T} SC^{(t)}$. In turn, LDA, DETECTOR and MGCTM, the coherence $SC^{(t)}$ of the generic topic t was quantified by means of the metric introduced in [24], which satisfactorily matches human coherence assessments. Formally, let $\boldsymbol{w}_{t,1}, \ldots, \boldsymbol{w}_{t,R}$ be the top-R words of topic t (i.e., the R most probable words under t). $SC^{(t)}$ is defined as follows

$$SC^{(t)} = \sum_{r=2}^{R}\sum_{p=1}^{r-1} log\frac{f(\boldsymbol{w}_{t,r}, \boldsymbol{w}_{t,p}) + 1}{f(\boldsymbol{w}_{t,p})}$$

where $f(\boldsymbol{w}_{t,p})$ is the document frequency of $\boldsymbol{w}_{t,p}$ and $f(\boldsymbol{w}_{t,r}, \boldsymbol{w}_{t,p})$ is the co-document frequency of $\boldsymbol{w}_{t,r}$ and $\boldsymbol{w}_{t,p}$ [24].

Instead, under k-means $\rightarrow$ LDA, NMF $\rightarrow$ LDA and HAC $\rightarrow$ LDA, the above definition of $SC^{(t)}$ was suitably adjusted, since these competitors separately infer the individual topics in the context of the input clusters. We adapted $SC^{(t)}$ as follows

$$SC^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\sum_{r=2}^{R}\sum_{p=1}^{r-1} log\frac{f(\boldsymbol{w}_{t^{(k)},r}, \boldsymbol{w}_{t^{(k)},p}) + 1}{f(\boldsymbol{w}_{t^{(k)},p})}$$

where $t^{(k)}$ means topic t in the context of cluster k.

Higher SC scores are indicative of more semantically coherent topics. The SC scores of all competitors are reported in Table IV. Such scores were computed, by maintaining the settings discussed in Section V-C about the number of clusters and topics. Moreover, R was fixed to 20.

DETECTOR and MGCTM infer more semantically coherent topics compared to all other competitors.

The gain in semantic coherence with respect to LDA highlights the benefit of pairing document clustering and topic modeling rather than performing topic modeling alone. Essentially, clustering groups text documents by content homogeneity, which favors the discovery of coherent topics.

k-means $\rightarrow$ LDA, NMF $\rightarrow$ LDA and HAC $\rightarrow$ LDA attain a lower semantic coherence compared to DETECTOR and MGCTM. This is due to the fact that, in such competing pipelines, document clustering is neatly separated from topic modeling. Thus, k-means, NMF and HAC are unable to properly isolate documents by their semantics, which ultimately penalizes the coherence of the inferred topics. This confirms that a suitable integration of document clustering and topic modeling is advantageous with respect to naively pipelining the two tasks with no interplay. Actually, the lack of interplay may be detrimental for semantic coherence. Indeed, on both text corpora, k-means $\rightarrow$ LDA, NMF $\rightarrow$ LDA and HAC $\rightarrow$ LDA unveil less semantically-coherent topics compared to LDA.

The higher semantic coherence achieved by DETECTOR in comparison with MGCTM substantiates the rationality of the devised interplay between topic modeling and document clustering with explicit cluster characterization.

2) Qualitative evaluation: The typical output produced by DETECTOR on real-world text corpora is exemplified through the inspection of its results on *20-Newsgroups*. Table V provides an insight into two of the uncovered clusters. The semantics of each such a cluster is described in terms of the top-2 most inherently-characteristic topics. In turn, the individual topics are summarized by their respective top-10 words. It is evident that DETECTOR uncovers clusters with an intelligible semantics, as suggested by the intuitiveness of the inferred topics, the clarity and specificity of their representative words, as well as the coherence of such words in the context of their respective topics. Interestingly, by looking at Table V, one can also appreciate the intra-cluster coherence enforced by DETECTOR. Therein, the semantics of each cluster is clearly distinguished by discriminative subsets of well-assorted and coherent topics. Indeed, *Cluster 1* is discriminated by *Topic 1* and *Topic 2*. These are distinct, though related topics, that enable the intuitive interpretation of *Cluster 1* as coherently devoted to the more general *guns* theme. Likewise, *Topic 3* and *Topic 4* permit an intuitive interpretation of *Cluster 2* as coherently devoted to *space exploration*.

As a concluding remark, we point out that the explicit summarization of cluster semantics through representative topic distributions is, in principle, helpful to conveniently query text corpora. To elucidate, in the absence of a structured organization of the underlying text corpus, one has to perform a sequential scan, in order to match a query document against the individual text documents. This is especially demanding with (very) large text corpora. Instead, under the unsupervised thematic classification inferred by DETECTOR, one can answer the query document, by issuing the latter only against the most pertinent document clusters. For instance, in the setting of Table V, the retrieval of text documents answering any query document on *Topic 1* or *Topic 2* only involves *Cluster 1*. Analogously, *Cluster 2* is the target portion of the text corpus for all those query documents on *Topic 3* or *Topic 4*.

VI. Conclusions

Document clustering and topic modeling can be interrelated to mutually benefit from each other. We presented an innovative machine-learning approach to coupling document clustering and topic modeling. A Bayesian generative model of document corpora, DETECTOR, was developed to reciprocally and seamlessly interrelate both tasks. Under DETECTOR, topics and clusters are latent factors, ruling document wording. Collapsed Gibbs sampling and parameter estimation were mathematically derived and implemented into an approximate inference algorithm, in order to perform the two tasks simultaneously and in an interdependent manner.

An intensive experimentation over standard benchmark datasets demonstrated the effectiveness of our approach in jointly clustering text documents and unveiling their semantics in terms of coherent topics. A qualitative analysis of its behavior on real-world text corpora was also provided.

	Topic 1	Topic 2
Cluster 1	right	weapons
	people	control
	gun	killed
	state	law
	think	power
	government	bill
	peace	gun
	law	peace
	against	problem
	weapons	article
	Topic 3	**Topic 4**
Cluster 2	space	research
	nasa	data
	gov	earth
	technology	science
	program	system
	center	technology
	science	space
	project	government
	cost	development
	high	answer

TABLE V: Two clusters from *20-Newsgroups* and their most inherently characteristics topics

There are several promising avenues for future research. An appealing line of further work involves the exploitation of extra-knowledge from external reference corpora under the form of word vectors [4], [15], [23]. This would allow for the exploitation of the syntactic and semantic relationships among words, which is expected to further improve the effectiveness of both tasks. Additionally, variational inference [6] can be investigated as an alternative to *MCMC* sampling [3], [26] in the attempt to further expedite approximate posterior inference on very-large text corpora. Besides, it is also interesting to study adaptations of DETECTOR for the domains of short text documents [20], [31], [32] and author-topic models [21], [22], [27], [28]. Finally, we planned to explore the integration of document clustering and topic modeling for semi-structured data analysis [10], [11], [14] as well as user profiling in social recommendation [9] and network role analysis [12], [13].

REFERENCES

[1] C. Aggarwal and C. Zhai. A survey of text clustering algorithms. In C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 77 – 128. Springer, Boston, MA, 2012.

[2] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez, and K. Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. In *arXiv preprint arXiv:1707.02919*, 2017.

[3] C. Andrieu, N. De Freitas, A. Doucet, and M.I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5 – 43, 2003.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137 – 1155, 2003.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859 – 877, 2017.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993 – 1022, 2003.

[8] D. Cai, X. He, and J. Han. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):902 – 913, 2011.

[9] G. Costa and R. Ortale. Model-based collaborative personalized recommendation on signed social rating networks. *ACM Transactions on Internet Technology*, 16(3):20:1–20:21, 2016.

[10] G. Costa and R. Ortale. Xml clustering by structure-constrained phrases: A fully-automatic approach using contextualized n-grams. *International Journal on Artificial Intelligence Tools*, 26(1):1 – 24, 2017.

[11] G. Costa and R. Ortale. Machine learning techniques for xml (co-)clustering by structure-constrained phrases. *Information Retrieval Journal*, 21(1):24 – 55, 2018.

[12] G. Costa and R. Ortale. Marrying community discovery and role analysis in social media via topic modeling. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2018.

[13] G. Costa and R. Ortale. Mining overlapping communities and inner role assignments through bayesian mixed-membership models of networks with context-dependent interactions. *ACM Transactions on Knowledge Discovery from Data*, 12(2):18:1 – 18:32, 2018.

[14] G. Costa and R. Ortale. Mining cluster patterns in xml corpora via latent topic models of content and structure. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 237 – 248, 2019.

[15] R. Das, M. Zaheer, and C. Dyer. Gaussian lda for topic models with word embeddings. In *Proc. of the Meeting of the Association for Computational Linguistics*, pages 795 – 804, 2015.

[16] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, and D.B. Dunson. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2009.

[18] G. Heinrich. Parameter estimation for text analysis. Technical report, University of Leipzig, 2008. Available at http://www.arbylon.net/publications/text-est.pdf.

[19] D. Koller and N. Friedman. *Probabilistic Graphical Models. Principles and Techniques*. The MIT Press, 2009.

[20] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proc. of Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 165 – 174, 2016.

[21] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: Joint models of topic and author community. In *Proc. of the Int. Conf. on Machine Learning*, pages 665 – 672, 2009.

[22] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249 – 272, 2007.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of Int. Conf. on Neural Information Processing Systems*, pages 3111 – 3119, 2013.

[24] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*, pages 262 – 272, 2011.

[25] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[26] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[27] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):4:1 – 4:38, 2010.

[28] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 306 – 315, 2004.

[29] R. Winkler. *An Introduction to Bayesian Inference and Decision*. Probabilistic Publishing, 2003.

[30] P. Xie and E.P. Xing. Integrating document clustering and topic modeling. In *Proc. of Int. Conf. on Uncertainty in Artificial Intelligence*, pages 694 – 703, 2013.

[31] J. Yin and J. Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 233 – 242, 2014.

[32] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong. Topic modeling of short texts: A pseudo-document view. In *Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 2105 – 2114, 2016.