Front. Comput. Sci. China 2010, 4(2): 280–301
DOI 10.1007/s11704-009-0062-y

REVIEW ARTICLE

Knowledge discovery through directed probabilistic topic models: a survey

Ali DAUD (✉)[1], Juanzi LI[1], Lizhu ZHOU[1], Faqir MUHAMMAD[2]

1 Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
2 Department of Mathematics and Statistics, Allama Iqbal Open University, Sector H-8, Islamabad 44000, Pakistan

Abstract Graphical models have become the basic framework for topic based probabilistic modeling. Especially models with latent variables have proved to be effective in capturing hidden structures in the data. In this paper, we survey an important subclass Directed Probabilistic Topic Models (DPTMs) with soft clustering abilities and their applications for knowledge discovery in text corpora. From an unsupervised learning perspective, "topics are semantically related probabilistic clusters of words in text corpora; and the process for finding these topics is called topic modeling". In topic modeling, a document consists of different hidden topics and the topic probabilities provide an explicit representation of a document to smooth data from the semantic level. It has been an active area of research during the last decade. Many models have been proposed for handling the problems of modeling text corpora with different characteristics, for applications such as document classification, hidden association finding, expert finding, community discovery and temporal trend analysis. We give basic concepts, advantages and disadvantages in a chronological order, existing models classification into different categories, their parameter estimation and inference making algorithms with models performance evaluation measures. We also discuss their applications, open challenges and future directions in this dynamic area of research.

Received May 30, 2009; accepted September 5, 2009

E-mail: ali_msdb@hotmail.com, ljz@keg.cs.tsinghua.edu.cn, dcszlz@tsinghua.edu.cn, aioufsd@yahoo.com

Keywords text corpora, Directed Probabilistic Topic Models (DPTMs), soft clustering, unsupervised learning, knowledge discovery

1 Introduction

Broadly graphical models can be divided into two main categories: "Directed" and "Undirected" graphical models. These types can be further classified into "Parametric" and "Non-Parametric" models (please see Fig. 1). Latent topic layer based graphical models have gained a lot of success in recent years by capturing the hidden patterns present in the data. Automatic extraction of topics from text is performed in Refs. [1,2] to cluster documents into groups based on similar semantic content. These models provide a good way of documents classification, but they are inherently limited by the fact that each document is only associated with one cluster. For this reason, soft clustering models are required, which can allow documents composed of multiple topics to relate to more than one cluster on the basis of hidden topics. Consequently, a research area of unsupervised learning with soft clustering abilities "Directed Probabilistic Topic Models (DPTMs)" came into being and attracted a lot of interest from researchers in both academic and industrial fields. Initially, Probabilistic Latent Semantic Analysis (PLSA) [3] was proposed as a probabilistic alternative to projection and clustering methods which can assign documents to different clusters by using maximum likelihood principle. PLSA was followed by the state-of-the-art Latent Dirichlet Allocation (LDA) and many of its extensions. Several challenging questions can be answered by applying topic models, e.g., How to do document modeling and classification by exploring their hidden patterns? How to perform collaborative filtering? How to explore authors interests and find experts of a specific area? How to find hidden associations between researchers or group of people? How to find collaborators for projects? How to find temporal trends in documents to analyze emerging fields? How to find roles of persons in social networks? and how to do document summarization and indexing?

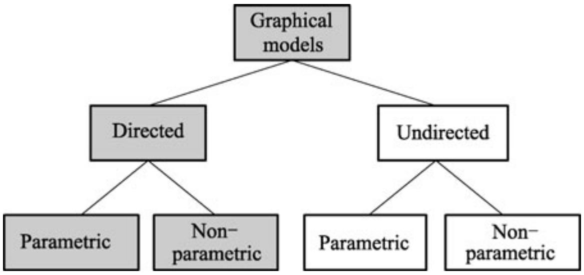

Fig. 1 Graphical models

Previously, two very useful introductory works about probabilistic topic models and their parameter estimation [4,5] skip details of classification, relative need, advantages and disadvantages of up-and-coming topic models. They also did not discuss topic models applications for different problems and thrust of future research in detail.

In this paper, we provide DPTMs classification on the basis of their key functionalities. We also investigate the problems of modeling text corpora and review important existing methods to solve these problems. Our contributions in this paper includes:

(1) a chronological study of important models by explaining their motivation, advantages and disadvantages;

(2) a classification of important models on the basis of their main functionality with explanation of the general framework of the selected models;

(3) a summarization of different problem domains in which these models are applied; and

(4) our insights about open challenges in topic modeling by discussing about probable solutions.

To the best of our knowledge, this is the first work that provides a detailed investigation about DPTMs by investigating such a long literature review, with classification and implementation details of important models and their advantages and disadvantages. We focus on Directed Probabilistic Topic Models (Bayes net) in this paper, so we will not discuss Undirected Probabilistic Topic Models "Harmoniums" (Markov Random Field) [6,7]. In the probabilistic modeling domain, "parametric" normally means number of topics for an approach are fixed in the beginning and they will not change during the process [8–10], while "non-parametric" normally means the number of topics are not fixed in the beginning, so they will be automatically optimized by the approach according to their fitness of dataset during the process [11].

The rest of the paper is organized as follows. In Section 2, we give concepts and terminologies related to DPTMs. Section 3 discusses the limitations of old models and the need for emerging models in a sequential way. Later in this section, we provide a general framework of selected models from different categories. Section 4 provides parameter estimation and inference making algorithms. Performance evaluation measures are explained in Section 5. Section 6 summarizes the applications of models in different problem domains. In Section 7, we discuss research issues and related future directions. Finally, we conclude this paper in Section 8.

2 Concepts and terminologies

In this section, we provide basic, intermediate and high level concepts and terminologies related to topic models.

2.1 Basic concepts and terminologies

2.1.1 Document

A document usually consists of many words, terms (multiple words), symbols, diagrams or tables, gathered under a representative title. Research papers and news articles are common examples of documents in topic modeling. Symbolically, for a document d: $\boldsymbol{w}_{di} = \{w_1 + w_2 + w_3 + \cdots + w_n\}$, where w_i is a word in a document. The total number of documents D is denoted by $\boldsymbol{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n\}$, where $\boldsymbol{w}_1$ is a word vector of document d_i .

2.1.2 Text corpora

A large collection of documents is called text corpora. For example, some famous datasets used in topic modeling are "NIPS proceedings" and "Cite seer". Both consist of a very large number of scientific publications which are used for different knowledge discovery tasks. "TREC AP" newswire articles corpus and the "Reuter's" news articles corpus are some famous datasets for news mining.

2.1.3 Topics

Different communities have different definitions for topics. In this paper and in topic modeling literature, topics are referred to hidden patterns or short descriptions of documents in a text corpus. Technically “topics are semantically related probabilistic clusters of words” which are used as a bridge between words and entities (e.g., documents or authors) to find hidden associations between them.

A topic is formally defined as “a probability distribution over words or terms in a vocabulary” and informally defined as “an underlying semantic theme; a document consisting of a large number of words might be concisely modeled as deriving from a smaller number of topics” [12]. Table 1 shows an example of topics Arts and Education. Each topic is shown with the top fifteen words and corresponding probabilities. The titles are our interpretation of the topics.

Table 1 Topic examples

Arts		Education	
Word	Probability	Word	Probability
New	0.03741	School	0.07344
Film	0.03626	Students	0.05702
Show	0.02753	Schools	0.04136
Music	0.02151	Education	0.02605
Movie	0.01854	Teachers	0.02465
Play	0.01124	High	0.02122
Musical	0.01109	Public	0.02026
Best	0.00989	Teacher	0.02006
Actor	0.00966	Bennett	0.01766
First	0.00899	Manigat	0.01746
York	0.00895	Namphy	0.01478
Opera	0.00870	State	0.0143
Theater	0.00854	President	0.01359
Actress	0.00817	Elementary	0.01219
Love	0.00806	Haiti	0.01211

2.1.4 Bag of words, sentences and documents assumptions

Bag of words assumption means that the order of words is ignored in the documents. So the important information for the model is the number of times words appeared in the document, not the order of words. Similarly, bag of sentence assumption means that the order of sentences is ignored in the documents and bag of document assumption means that the order of documents is ignored in the corpus.

2.1.5 Topic models and modeling

Topic models are based on the idea that the documents can be represented as a mixture of topics, where a topic is a probability distribution over words and these documents can be generated by the simple probabilistic procedure of generative models [4].

Generally speaking, the process for finding latent topics from text corpora by using topic models is called topic modeling. Technically speaking, it is the process of finding a topic z in a document d with defined probability distribution of words in a vocabulary V by using topic models.

2.1.6 Synonymic terms and notations

The history of topic modeling is not new, so different researchers have used different terminologies to represent synonymic terms. Terminologies used interchangeably in this paper as well as in the topic modeling literature are given in the following. Text corpora, corpus, and large collections of documents are used interchangeably. Information discovery, statistical analysis, human language learning and processing, modeling text corpora, statistical modeling of language, and statistical language learning are used interchangeably. Topics, hidden topics, hidden patterns, latent topics, latent aspects, buried patterns, latent structure, and short descriptions are used interchangeably. Different researchers used different notations while explaining the structure of topic models. We used similar notations for all models discussed here for readability. Table 2 summarizes the notations used throughout this paper.

2.2 Intermediate concepts and terminologies

2.2.1 Exchangeability of topics

Exchangeability of topic means that there is no fixed order of topics for different runs of the algorithm. For example, a topic z_i in the first run of the algorithm is not theoretically considered to be similar to topic z_i in the second run of the algorithm.

2.2.2 Topic optimization

For a topic model, finding the best number of topics is very important because usually the choice of topics can

Table 2 Notations

Symbol	Description
$\boldsymbol{D}$	Number of documents
N	Number of words
T	Number of topics
A	Number of unique authors
V	Number of unique words
N_d	Number of word tokens in document d
$\boldsymbol{w}_d$	Vector form of document d
$\boldsymbol{a}_d$	Vector form of authors in document d
w_{di}	The ith word token in document d
z_{di}	Topics assigned to word token w_{di}
x_{di}	The author associated with w_{di}
y_{di}	The timestamp associated with token w_{di}
θ_d	Multinomial distribution over topics with parameter α
Φ_z	Multinomial distribution of words specific to z with parameter β
Ψ_z	Time specific Beta distribution of topic z
α	Dirichlet distribution associated with topic z
β	Dirichlet distribution associated with word w_{di}
ε	Binomial Distribution associated with transition Ω_i
rn	Root Node (or root topic)
R	Response variable used as observed value in supervised topic models
L	Link between documents
d	Source document
d'	Target document
τ	Link value between documents
γ	Dirichlet distribution associated with link τ
λ	Multinomial distribution for link generation between documents
$\boldsymbol{C}$	Class of word, e.g., Noun Phrase (NP), Not Noun Phrase (NNP)

affect the interpretability of the results. A solution with a small number of topics usually results in very general topics; conversely a solution with a large number of topics usually results in un-interpretable topics that pick out idiosyncratic word combinations. Additionally, topic optimization is usually dependent on the number of documents in a dataset, as a small dataset will usually be optimized at a small number of topics, as compared to a large dataset. For example, in Ref. [8] the optimized number of topics for 16333 newswire articles are 100, while the optimized number of topics for 5225 scientific abstracts are 50.

A number of methods are used for topic optimization. First, optimal numbers of topics are found based on a model's generalization performance on the unseen dataset. For example, a model is first estimated on a subset of dataset and then used for inference making on the word choice in the remaining set of documents. Perplexity is used for accessing the generalization power of text models on subsets of documents [8,9]. Second, a Bayesian model selection approach [10] can also be used to estimate the posterior probability of the model while integrating over all possible ways to assign words to topics. The number of topics is then based on the model that leads to the highest posterior probability. Finally, Teh et al. [11] proposed a solution that can be used for topic optimization.

2.2.3 Polysemy with topics

Polysemy is a very important language characteristic discussed in natural language processing, in which words have multiple meanings and ambiguity can be handled by using other words in the context. Topic models play an important role to resolve this polysemy issue. For example, Table 3 shows two topics "Sports" and "Entertainment" with top ten words in which the word **Play** is present. The context in which the word play is used in the sports topic is different from the context in which it is used in the entertainment topic, which helps to deal with polysemy of words. By using the other words in that, e.g., in document classification task, a document that has more words related to the entertainment topic will see play in a different context as compared to the sports topic. Topics titles Sports and Entertainment are just our interpretation of the topics.

Table 3 Polysemy with topics

Sports	Entertainment
Game	Art
Play	Music
Ball	**Play**
Team	Part
Playing	Sing
Games	Like
Football	Poetry
Baseball	Band
Field	World
Sports	Rhythm

2.2.4 Dirichlet distribution

The Dirichlet distribution often denoted Dir(α), is a family of continuous multivariate probability distributions parameterized by the vector α of positive real's. It is the multivariate generalization of the beta distribution, and the conjugate prior of the categorical distribution and multinomial distribution in Bayesian statistics. That is, its probability density function returns the belief that the probabilities of E rival events are e_i given that each event has been observed $\alpha_i - 1$ times[†].

2.2.5 Multinomial distribution

The multinomial distribution is a generalization of the binomial distribution. The binomial distribution is the probability distribution of the number of "successes" in n independent Bernoulli trials, with the same probability of "success" on each trial. In a multinomial distribution, the analog of the Bernoulli distribution is the categorical distribution, where each trial results in exactly one of some fixed finite number k of possible outcomes, with probabilities $p_1,..., p_k$ (so that $p_i \geqslant 0$ for $i = 1,..., k$ and $\sum_{i=1}^{k} p_i = 1$), and there are n independent trials. Then let the random variables X_i indicate the number of times outcome number i was observed over the n trials. The vector $U = (U_1,..., U_k)$ follows a multinomial distribution with parameters n and p, where $p = (p_1,..., p_k)$[†].

2.2.6 Beta distribution

The beta distribution is a family of continuous probability distributions defined on the interval [0,1] parameterized by two positive shape parameters, typically denoted by α and β. It is the special case of the Dirichlet distribution with only two parameters. Since the Dirichlet distribution is the conjugate prior of the multinomial distribution, the beta distribution is the conjugate prior of the binomial distribution. In Bayesian statistics, it can be seen as the posterior distribution of the parameter p of a binomial distribution after observing $\alpha - 1$ independent events with probability p and $\beta - 1$ with probability $1 - p$, if the prior distribution of p was uniform[†].

2.3 High-level concepts and terminologies

2.3.1 Graph plate notations

Graph plate notations are a visual presentation to interpret topic models. In graph plate notations, shaded and unshaded variables indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and plates indicate repeated sampling with the number of repetitions given by the variable in the bottom. Symbols of graph plate notations are shown in Fig. 2. For additional details please see Ref. [13].

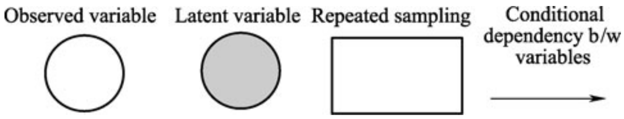

Fig. 2 Graph plate notations symbols

2.3.2 Generative models

Formally, a generative model is a model for randomly generating observable data, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Generative models are used in machine learning for either modeling data directly (i.e., modeling observed draws from a probability density function), or as an intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through the use of Bayes' rule[†].

Informally, generative models are simple probabilistic sampling rules that describe how words in a document might be generated on the basis of latent variables (such as topics). The goal is to find the optimized set of latent variables (topics) that can explain the words in documents. Usually in generative models it is assumed that the data are generated by the model [4].

2.3.3 Bayes network

A Bayesian network, belief network or directed graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between crops and rain. Given rain, the network can be used to compute the probabilities of the quantity of different crops[†].

3 Directed Probabilistic Topic Models (DPTMs)

In this section, we will provide development of topic models in a chronological order with classification into five categories based on their main functionalities

[†] Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Generative_model

observed during analysis. Discussing advantages and disadvantages in a chronological way provides us with a better way to understand problems of old models and their extensions which have overcome old models problems. Later at the end of each category sub-section, we describe a basic intuition about each category and framework of selected topic models.

Careful analysis of historical paradigm of topic models provides us with several interesting trends. One can clearly see from Table 4, that LDA is the unanimously leading model in the history of probabilistic topic modeling due to its numerous extensions for modeling text corpora. By analyzing Table 4 deeply, we uncover additional interesting trends, such as the topic models are not only limited to bag of words assumption, in addition they also considered inter document link dependencies, intra document Markov dependencies, temporal trends and labeled data. Some models have multiple applications, e.g., the Author-Topic model [14], which is mainly used for authors interests and association finding with respect to topics, but additionally it can be used for temporal topic trend finding. Continuous-Time model [15] has a mixture of functionalities, as it can use both temporal information and Markov dependencies in the documents at the same time to discover temporal topic trends.

In Table 4, directed links (→) show that model mentioned in bold are extended to the model mentioned in the normal font. Here the latest models are improvements over the old ones and can be considered best for the problems they solved in their category.

3.1 Basic DPTMs (BDPTMs)

We begin with Probabilistic Latent Semantic Analysis (PLSA) [3] which has a latent layer that can be used for finding latent topics in text corpora and information retrieval tasks. PLSA is a useful step toward probabilistic modeling of text, but it considers that each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers or documents. This leads to two major drawbacks. First, the number of parameters in the model

Table 4 Historical paradigms of PDPTMs from 1999–2009

Year/Type	Basic PDPTMs	Inter-Document Correlated PDPTMs	Intra-Document Correlated PDPTMs	Temporal PDPTMs	Supervised PDPTMs
1999	PLSA				
2000					
2001		**PLSA**→A Joint Probabilistic Model			
2002	A probabilistic Approach				
2003	LDA, A Topic Model				**LDA** → Corr-LDA
2004	Discrete PCA	**LDA** →Mixed Membership Models, **LDA** →Author-Topic Model, **LDA**→ **Author-Topic Model** →ART			
2005			**LDA** → HMM-LDA		**LDA** →LLDA
2006		**LDA** → PAM, **LDA** → CTM, **LDA** →Statistical Entity-Topic Models	Bigram Topic Model, **PLSA** → CPLSA	**LDA** → TOT, **LDA** → DTM, (**PAM, TOT**) → Continuous Time Model	
2007		**LDA** → GWN-LDA, **LDA** → Citation Influence Model	**LDA** →HTMM, **LDA** → TNG	MTTM	**LDA** → sLDA
2008		**LDA** →LTHM, **LDA** → **Author-Topic Model** → ACT Model **(A Joint Probabilistic Model, LDA)** → Link-PLSA-LDA		**LDA** → **DTM** →cDTM	
2009		**LDA** → Generalized LDA, **LDA** → **Author-Topic Model** → **ACT** à Generalized ACT		**LDA** → **Author-Topic Model** → TAT	

grows linearly with the size of the corpus, which leads to serious problems of over fitting, and second, it is not clear that how to assign probability to a document outside of the training set [8]. Simply speaking it is generative at the words level but not at documents level. A model based on the unigram model was proposed called Mixture of Unigrams [16]. In unigram model, the words of every document are drawn independently from a single multinomial distribution. If the unigram model is augmented with a discrete random topic variable z, we obtain a mixture of unigrams model. Mixture of Unigrams is based on the supposition that each document exhibits only one topic, which was too limited to effectively model text corpora, a limitation discussed in Ref. [8].

One cannot disagree with the usefulness of PLSA for an information retrieval tasks, however in order to overcome its limitations a generative probabilistic topic model Latent Dirichlet Allocation (LDA) was proposed [8]. LDA assumes that each word in the document is generated by a hidden topic and explicitly models the words distribution of each topic, as well as the prior distribution over topics in the document. Given these parameters, topics of all words in the same document are assumed to be independent. In LDA, a document can generate more than one topics and it is possible to assign probability to documents outside the corpus by using variational inference algorithm and Gibbs sampling. It is generative at both words and documents level. LDA is computationally efficient than PLSA due to not having the problem of large parameters growth with the scale of input data. Various extensions of LDA (shown in Table 4) have discussed its limitations and proposed enhanced models with improved performance in different problem domains. A similar kind of topic modeling approaches (A probabilistic Approach and a Topic Model) [17,18] were proposed, both of them capturing the probabilistic relationships between words and documents to effectively capture the semantics of words. These approaches were also based on the ideas that documents are mixture of topics, where a topic is a probability distribution over words in the documents.

LDAs' influence is everywhere on modeling text corpora. However, Ref. [19] came with a generalization of previous models in the form of discrete Principle Component Analysis (discrete PCA) for analyzing large collections of data. They argued that PLSA, LDA and Expectation-propagation [3,8,20] are similar approaches, ignoring methodology and notations. Thus, there was a need to translate notations; they jointly called the methods for translating these notations as discrete PCA. These methods proved useful for modeling text corpora problems with improved generalization capabilities on unseen data.

BPDPTMs mainly discover hidden topics on the basis of semantic-based text information with bag of words assumption. They consider that there are hidden relationships present between entities (e.g., documents, authors, conferences, product or users) that can be exploited by using the structures of words present in documents. They simply exploit the semantics based similarity of unigram words to correlate entities on the basis of latent topic layer. The framework of selected topic models for BPDPTMs category is explained below.

3.1.1 Probabilistic Latent Semantic Analysis (PLSA)

PLSA [3] can be considered as the first probabilistic methodology with a latent layer and a strong statistical foundation, which is used for topic modeling. The core of PLSA is a statistical model which is called the aspect model [21]. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in \boldsymbol{T} = \{z_1,...,z_t\}$ with each observation.

A joint probability model over $d \times w$ is defined by the mixture,

$$p(d,w) = p(d)P(w|d),$$

$$\text{where } p(w|d) = \sum\nolimits_{z \in T} p(w|z)p(z|d). \tag{1}$$

Each pair (d,w) is assumed to be generated independently, corresponding to bag of words assumption. The words w is generated independently of the specific document d conditioned on topic z; simply it can be called a generative model at word level but not at document level. The corresponding graphical model representation is depicted in Fig. 3. The standard procedure for maximum likelihood estimation in latent variable models Expectation Maximization (EM) algorithm is used for parameter estimation.

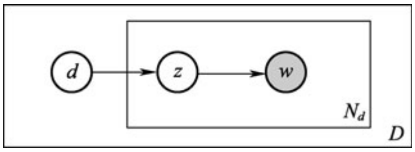

Fig. 3 PLSA

3.1.2 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model for modeling text corpora [8,10], which has overcome the limitations of PLSA by providing a generative model at words and documents level. The basic idea of LDA implies that documents are represented as random mixtures over latent topics, where each topic is defined by a distribution over words. LDA captures implicit correlations between words via topics and can also assign a probability to the unseen documents by using the variational inference algorithm.

Graphical representation of smoothed LDA is shown in Fig. 4. Principally LDA is a three-level Bayesian network that generates a document using a mixture of topics. First, for each document *d*, a multinomial distribution θ_d over topics is randomly sampled from a Dirichlet with parameter α (where α is commonly set as 50/*T*). Second, for each word w_{di}, a topic z_{di} is chosen from this topic distribution. Finally, the word w_{di} is generated by randomly sampling from a topic-specific multinomial distribution Φz_{di} from a Dirichlet with parameter β (where β is set as 0.1, an increase in the value of β will result in sparse topics while decrease will result in dense topics) [10]. Therefore, the generating probability of word *w* from document ***D*** is given as

$$p(w|d,\theta,\Phi) = \sum\nolimits_{z=1}^{T} p(w|z,\Phi_z)p(z|d,\theta_d). \qquad (2)$$

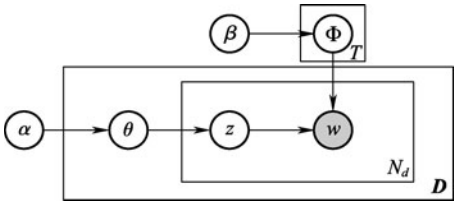

Fig. 4 Smoothed LDA

For parameter estimation and making inference in LDA, a simple convexity-based variational EM algorithm and Gibbs sampling algorithm is used [8,10], respectively.

3.2 Inter-Document Correlated PDPTMs (IrCPDPTMs)

PLSA takes into account the semantic structure present between the words of documents on the basis of latent topics, while natural links between research papers through citations can be useful for finding correlations between them additionally with topics. A model based on these natural links is A Joint Probabilistic model [22], which is an extension of PLSA [3]; this joint model simultaneously models the topic specific influence of documents and it is different from PLSA in the sense that it defines a generative process not only for text but also for citations. However, it suffers from a large number of parameters and local maxima problems similar with PLSA [3]. Implicit dependencies between the words of a document and link based associations (e.g., citations) between documents were already considered important in the past by Mixture of Unigrams and Joint Probabilistic model [16,22]. However Markov dependencies between hidden topics are also important to know about the semantics of topics. Based on this idea, the aspect HMM (AHMM) model [23] was proposed. AHMM considers Markov dependencies for unstructured data streams. However in AHMM, documents were inferred using heuristics, which assume that each document exhibits only one topic, a similar limitation as of mixture of unigrams.

A Joint Probabilistic model [22] has shown the importance of citation links between documents for information discovery in a corpus with the above discussed limitations. To overcome these in link based topic models, fully link based generative models named Mixed-Membership models [24] were proposed. They can be considered as an extension of LDA with embedding link information for references or citations. Their limitation is that they cannot exploit the topical relationships between the papers on either side of links [25].

The basic idea of topic modeling, that words and documents can be modeled by considering latent topics became the intuition of modeling words and authors of documents by considering latent topics [14] for discovering authors' interests. In the Author-Topic model, added information of authors it can be used to find authors with respect to their associations with latent topics. This model can also be used to analyze topic trends over time, find authors who are likely to write on some specific topics, and find the most unusual writings of authors. Author-Topic model is very useful in documents and authors context, but it was not useful in finding directed relationships and interactions of persons in email social networks. Consequently, a probabilistic model of words in a generated message given their authors and recipients, Author-Recipient-Topic (ART) model was proposed [26]. ART model is almost similar to Author-Topic model but with an important enhancement of conditioning per message topic distribution jointly on both the author (sender) and individual recipients of email. Thus discovery of topics and roles are fully dependent on the social structure of senders and receivers in an email network.

LDA, as well as other models, failed to explicitly model correlation between topics. While in most collections of documents it is quite natural that subsets of underlying latent topics will also be correlated, this thinking became the base of Correlated Topic Model (CTM) [27]. It uses flexible distribution for topic proportions that allows for considering direct correlation between topics. Its logistic normal distribution parameters include a covariance matrix, in which each entry indicates the correlation only between a pair of topics. In CTM [27] topics are not independent, only pair wise correlations are modeled, and number of parameters in the covariance matrix grows as the square of the number of topics. These inabilities of CTM was discussed in Ref. [28], and the flexible Pachinko Allocation Model (PAM) which captures arbitrary, nested, and probably sparse correlations between topics using a Directed Acyclic Graph (DAG) was proposed. In PAM, Individual words in the vocabulary are represented by DAG, while correlation among children is represented by interior nodes.

Topics were discovered from documents so far, but what about relationships between the entities present in the text of documents? The answer to this question was the Statistical Entity-Topic models [29]. These models directly learn relationships between the discussed topics and the mentioned entities by using word topics including a mixture of entity topics (not individual entities), in the news articles to provide a better view of entity topic correlations.

In the Author-Topic Model [14], author groups with respect to topics are discovered by using latent topics of documents. A very similar kind of problem was discussed [30] for discovering probabilistic community profiles in social networks and an extension of LDA named Generic Weighted Network-Latent Dirichlet Allocation (GWN-LDA) was proposed. GWN-LDA can model communities as latent variables, and latent variables are considered as a distribution over all actors of a social network. The effectiveness of GWN-LDA for discovering community structures in social networks was shown in comparison with distance based clustering measures, which cannot consider latent structures of the documents.

In case of link based (e.g., citations) relationships, it is quite natural to assume that if a paper is cited by another paper, they can be topically related and there is a chain of papers based on this assumption. This is why links utilization between papers in topic modeling has already been started to develop slowly in the form of A Joint Probabilistic model and Mixed Membership Models [22,24] to discover better latent topics and correlations between documents by influencing additional information of links. Recently, a Citation Influence model was proposed to predict the citations influences of the documents on citing documents [31]. In this model a DAG is used for modeling the particular structure of paper citations. The citation influence model only considers general influences of citations, and ignores explicit topics specific influences of citations. To address this problem Link-PLSA-LDA was proposed [25]. The Link-PLSA-LDA model can exploit the relationships on either side of the citations, (between cited and citing documents), by considering the cited documents as bins to be filled by words. Its limitations are that, citing and cited documents are generated separately; as a result a single document cannot have both citations and be cited. Secondly, topical distribution of topics is defined by the model from a fixed number of documents, which means that the model is only generative at word level but not at document level.

The limitations of Mixed-Membership models and Citation Influence model [24,31] (that they cannot exploit the topical relationships between the papers on either side of links) are discussed in Ref. [32], and a new model Latent Topic Hypertext Model (LTHM) was proposed. The LTHM model can directly model a hypertext corpus in which document to same document or document to all other documents links can exists. The effectiveness of the LTHM model on webkb and Wikipedia datasets is shown in comparison to other models, due to LTHMs ability to have a smaller number of parameters.

Author-Topic model was used to model the documents and authors simultaneously to discover writing habits of authors. Tang et al. [33] discussed that conferences and authors are interdependent and should be modeled together. Consequently, a unified topic modeling approach called Author-Conference-Topic (ACT) was proposed, which can discover academics social networs on the basis of semantic structure of words and authors by considering conferences information. A variation of ACT [33] was proposed for expert finding problem named Semantics and Temporal Information based Maven Search (STMS) [34] which can also be called Generalized ACT. It is based on the ideas that (1) authors publishing in the world class conferences are probably the mavens (experts) of a specific area of research and there correlations are highly influential and (2) papers submitted to world class conferences are carefully judged for relevance to the sub-topics, so papers are more typical (strongly semantically related). STMS considers a collection of all papers and authors in a conference as a virtual document and exploits the semantics-based structure of words, authors' correlation and time between conferences (instead of single documents without time in ACT [33]) to normalize the time effect and include the influence of conferences.

Daud et al. discussed that semantics at subgroup-level (document) are poorer than the semantics at group-level (conference) and proposed the Conference Mining approach "Generalized LDA" (GLDA) [35]. GLDA considers a collection of all papers in a conference as a super-document and exploits the semantics-based structure of words presented in the conferences without considering authors information. It performs better than the ACT model [33] for conference ranking and conference correlations because it produced dense topics (Less Entropy).

IrCPDPTMs mainly make use of the links (e.g., citations or co-authorships in case of research papers) between the documents. They consider that directed links between entities are important and should be used in addition to the structure of words present in documents. The main intuition behind them is that the structure of words and directed relationships between entities should be modeled together to find more realistic relationships between entities by influencing latent topic layer with directed links. The framework of selected topic models for IrCPDPTMs category is explained below.

3.2.1 Author-Topic Model

The Author-Topic Model [14] models documents and authors of document together by extending LDA. LDA's basic idea that words of documents can be modeled to find document correlations by using a latent topic layer motivated the modeling of words and authors of documents together to find authors, interest. In the real world, when an author decides to write a document, initially he selects some topic(s) and then generates words of document related to the topic(s). The Author-Topic model can successfully discover topical authors' interests.

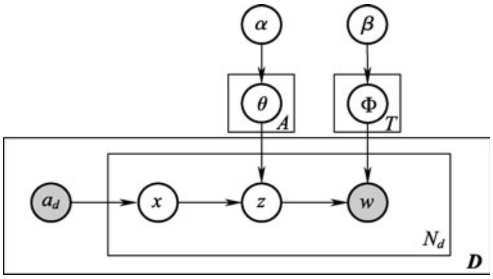

Fig. 5 Author-Topic Model

The graphical representation of the Author-Topic model is shown in Fig. 5. In this model, each topic is associated with a multinomial distribution Φ over words. Each author from a set of A authors is associated with a multinomial distribution θ over topics. Both θ and Φ have a symmetric Dirichlet prior with hyper parameters α and β. For each word in the document, an author x is uniformly sampled from a set of coauthors a_d, then topic z is sampled from the multinomial distribution θ associated with author x, and word w is sampled from the multinomial topic distribution Φ associated with topic z.

$$p(w|a,d,\Phi,\theta) = \sum\nolimits_{z=1}^{T} p(w|z,\Phi_z)p(z|a,\theta_a). \tag{3}$$

For the parameter estimation, a Monte Carlo Markov Chain MCMC technique Gibbs sampling is used in Author-Topic model.

3.2.2 Latent Topic Hypertext Model (LTHM)

LTHM [32] explicitly models the generation of links for hypertext document collections. It makes use of links together with exploiting information provided in the text of the linked documents to provide better topics and most related links to it (links with high probability to be generated from the topic).

Figure 6 shows two scenarios (a) scenario in which LTHM generates links from target document d' to source document d and (b) scenario in which LTHM model generates links from document d' to any other document in the collection of $\boldsymbol{D}$ documents.

The generative model is based on two steps. In the first step, a similar generative process to LDA is adopted. In the second step, links are created for already generated words of the documents. Here, a multinomial distribution λ with Dirichlet parameter γ is used to create a link τ from word w_i to document τ_i. LTHM models the generation of links from a word w to a document d, depending on how frequent the topic of word w is in the document d in addition to the in degree of document d. The probability of the generation of a link from a source document d to target

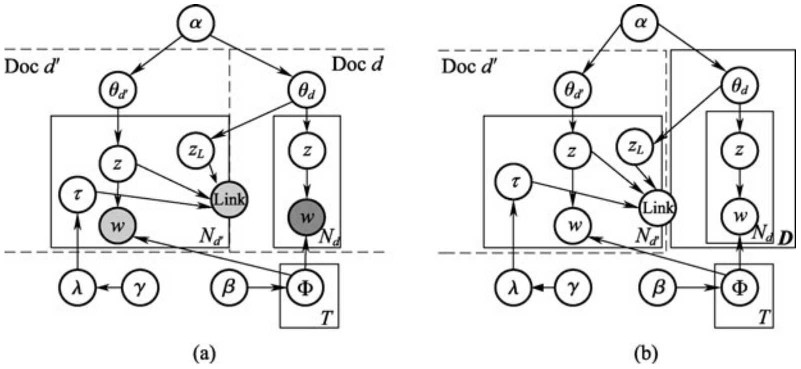

Fig. 6 LTHM Model

document *d'* depends on the topic of the link originating word, in degree of document *d'*, and topic mixtures of target document *d'*. For parameter estimation and inference making in LTHM an expectation-maximization (EM) procedure is adopted.

3.3 Intra-Document Correlated DPTMs (IcCDPTMs)

In the past, the AHMM model [23] considered Markov dependencies between the words of documents, but its basic idea was too limited in the sense that one document could generate only one topic. Consequently, Ref. [36] presented the HMM-LDA model for text documents to consider Markov dependencies with no limitations of topics per each document. In the HMM-LDA, hidden Markov model (HMM) determines when to generate a word from a model. As a result the sentence is factorized into function words handled by HMM, and content words are handled by LDA. HMM-LDA treated only the latent variables of syntactic classes as a sequence with local dependencies while latent assignments of topics were treated in a similar way to LDA. Therefore, their limitation is that topic extraction cannot benefit from the additional information conveyed in the structure of words [37].

The problem of ignoring Markov dependencies in previous topic models was discussed in Ref. [38] and a Bigram Topic model was proposed. In this model, *N*-Gram statistics are combined with latent topic variables to model consecutive relationships. The model predicts each word based on the immediately preceding words. This was a good step to move away from the bag of words assumption and give topic models more predictive power which really happened, but words only depending on the previous words becomes too limiting in some other real world datasets.

Both HMM-LDA and Bigram Topic model [36,38] discussed correlations between the short-range syntactic dependencies or long-range semantic dependencies between words of the documents, Ref. [39] proposed a Contextual Mixture (CPLSA) model based on Latent Semantic Analysis (LSA) [40]. It is an extension of PLSA [3], which introduces context variables to model the syntactic dependencies of a document to explore temporal correlations between topics and entities.

Later, Ref. [37] came up with the idea to consider text dependencies in the form of Hidden Topic Markov Models (HTMM). They made an assumption that all words in a sentence belong to the same topic, and consecutive sentences are more likely to belong to the same topic also. By using HMM they incorporated this dependency and showed that HTMM can learn better topics, and can do improved disambiguation of words in comparison with LDA. In the same year, a Topical *N*-Gram (TNG) model [41] was proposed, which also used Markov dependencies. It discovers semantically related arbitrary length phrases instead of discovering semantically related unigram words like the Bigram Topic model [38].

IcCPDPTMs mainly make use of Markov dependencies within the text of the document. They consider that syntactic dependencies between words in a document, e.g., current word is dependent on the previous word is important and should be considered instead of just considering the document as a bag of words. One can consider the document as a bag of Bigrams, bag of sentences, bag of paragraphs or bag of *N*-Grams depending on different situations. The main intuition behind them is natural language processing and sequential labeling techniques which have shown improvements by embedding syntax dependencies. More clear and representative topics can be found by considering within document syntax dependencies, which can result in a better performance of models. The framework of selected topic models for IcCPDPTMs category is explained below.

3.3.1 A composite model (HMM-LDA)

A composite model [36] also known as HMM-LDA, is a model in which HMM is the syntactic component and LDA is a semantic component. Its syntactic component HMM captures the short-term dependencies between the words of a document, while its semantic component topic model (LDA) captures long term dependencies within a document. HMM-LDA assumes that words in one document are related to same topics. HMM is comprised of different states that correspond to different syntactic word classes. One of its special states is used to host LDA to divide content words into different topics.

A graphical representation of HMM-LDA is shown in Fig. 7. Formally, it is composed of a sequence of words $\boldsymbol{w} = \{w_1, w_2,\ldots, w_n\}$ with each w_i being one of the words $\boldsymbol{w}$, a sequence of topics $\boldsymbol{T} = \{z_1, z_2,\ldots, z_n\}$ with each z_i being one of the topics $\boldsymbol{T}$, and a sequence of classes $\boldsymbol{C} = \{c_1,c_2, \ldots,c_n\}$ with each c_i being one of the classes $\boldsymbol{C}$. When one class $c_i = 1$, it is assigned a semantic class and the zth topic is associated with a distribution over words Φ_z. When class $c_i \neq 1$ it is associated with a distribution over words Φ_C because it does not carry any meaningful information. Here, each document has a distribution over topics θ_d and a distribution $\pi_{c_{i-1}}$ is used for transition between two classes' c_{i-1} and c_i. Given $\boldsymbol{w}$ words, $\boldsymbol{C}$ class assignments, other topic assignments z_{-i}, and the hyper-parameters, each z_i is drawn as

$$p(z_i|z_{-i},c,\boldsymbol{w}) \propto p(z_i|z_{-i})p(w_i|z,c,w_{-i})$$

$$\propto \begin{cases} n_{zi}^{(di)} + \alpha, & c_i \neq 1; \\ (n_{zi}^{(di)} + \alpha)\dfrac{a_{wi}^{(zi)} + \beta}{n^{(zi)} + V\beta}, & c_i = 1, \end{cases} \qquad (4)$$

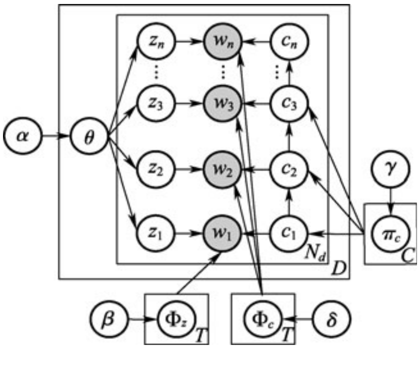

Fig. 7 HMM-LDA Model

where $n_{zi}^{(di)}$ the number of words in a document d_i is assigned to a topic z_i, and $n_{wi}^{(zi)}$ is the number of words assigned to a topic z_i. Case i is excluded and all counts include only the words for which $c_i = 1$. By using a conjugate of the Dirichlet and multinomial distributions, conditional distributions are obtained to integrate out the parameters θ and Φ. For performing Bayesian inference a Monte Carlo Markov Chain (MCMC) approach Gibbs sampling is used in HMM-LDA.

3.3.2 Hidden Topic Markov Model (HTMM)

HTMM [37] models the topics of words in a Markov chain. HTMM assumes that all words in the same sentence have the same topic, and successive sentences are more likely to have the same topics. Hence, the order of words in sentences is considered important. HTMM proves that by incorporating this dependency, better topics can be learned and that words can be differentiated better with respect to different topics as compared to LDA [10].

Figure 8 shows topics in a document forming a Markov chain with a transition probability that depends on θ and a topic transition variable Ω_n with Binomial distribution ε. When $\Omega_n = 0$ the topic of the nth word is the same as the previous one and for $\Omega_n = 1$, a new topic is drawn from θ. It is assumed that topic transitions can only occur between sentences, so Ω_n may only be nonzero for the first word in the sentence. T denotes the number of topics and N_d is the length of the document. The generating probability computed for each sentence in a document is given as

$$p(z_n,\Omega_n|d,w_1 \cdots w_n;\theta,\Phi,\varepsilon). \qquad (5)$$

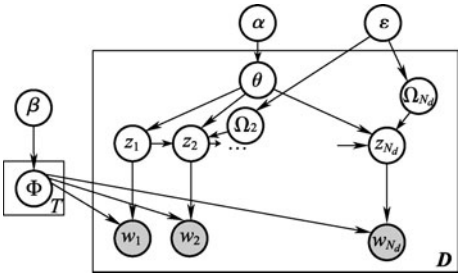

Fig. 8 HTMM Model

For model parameter estimation standard HMM tools are used, namely Expectation-Maximization and the Forward-backward algorithm, please see Ref. [42] for details.

3.4 Temporal DPTMs (TDPTMs)

Document topics in a text collection evolve over time, and it is interesting to explicitly model the dynamics of the underlying topics. For this purpose an extension of LDA, named Dynamic Topic Model (DTM) [43] was proposed, which captures the evolution of topics in a sequentially organized corpus. However, DTM ignores the natural term drift by time discretization, which can explicitly capture the rise and fall in the popularity of topics. One problem with DTM is that normal distribution was considered as a conjugate to the multinomial distribution. Consequently, there is no simple solution to the problems of inference and estimation. So, an alternative to DTM is the Multiscale-Topic Tomography model (MTTM) [44], which was more natural to sequential modeling of counts data. MTTM provide a better solution to the evaluation of topics by using conjugate priors on the topic parameters, and also provide topics evaluation at various resolutions of the time scale. Recently, an extension of DTM [43] named Continuous Time Dynamic Topic model (cDTM) [45] was proposed, to solve the problem of time discretization, which affects the memory requirements and computational complexity of posterior inference in the case of DTM [43]. For a given sequence of documents, they consider time to be continuous by using Brownian motion [46] to model continuous-time topic evolution.

A Topics over Time (TOT) model was proposed [47], which uses temporal information. This model represents time-stamps of documents as observed continuous variables. It has discretizes time for documents at the year level; meaning all the words in one document to have the same time stamp. Each topic is associated with a continuous beta distribution over time, and topics simply generate both words and observed time-stamps, while ignoring temporal patterns in their co-occurrences. TOT does not capture arbitrary, nested, and probably sparse correlations between topics. Thus, it was extended to Continuous-Time model [15], by adding these abilities by using a directed acyclic graph (DAG). This model overcomes the drawbacks of TOT and discovers correlations among topics and their changes over time simultaneously, in research paper corpus.

TOT only models the changing trends of documents while ignoring authors' interests. Intuitively authors' interests change with respect to time and there should be different authors related to a topic for different years. In order to model changing trends of document and authors together, the Temporal-Author-Topic (TAT) approach [48] was proposed. It is an extension of the Author-Topic models that add a multinomial distribution for temporal information. TAT can rank authors for different years and topics. It can also be used to find semantics based correlations of authors for different time periods.

TPDPTMs mainly discover hidden topics with respect to temporal trends by using time stamps of the documents. They consider that the time feature of a document is important and should be used in addition to the structures of words present in the text of documents. The main intuition behind them is that the structure of words and relationships between entities are not the same every year. Consequently there is a need to exploit structure of words with changing time to find word trends, topic trends and topically related entities for different years on the basis of latent topic layer. The framework of selected topic models for TPDPTMs category is explained below.

- Continuous-Time Model

Continuous-Time Model [15] combines the advantages of two previous models PAM [28] and (TOT) [47] to capture temporal patterns in individual topics as well as temporal patterns in their co-occurrences. It can capture both time-localized changes in topic occurrence with continuous time stamps and arbitrary topic correlations to show interesting correlations among topics and their change over time.

A graphical representation of the continuous-time model is shown in Fig. 9. Here $\boldsymbol{T} = \{z_1, z_2, \ldots, z_n\}$, a set of topic nodes. Each of them captures some correlation among words of topics. Special node is called root node and denoted by *rn*. It has no incoming links and every topic path starts from it. A DAG consists of nodes in *V* and

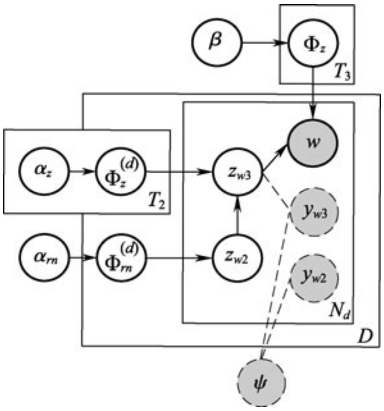

Fig. 9 Continuous-Time model

T. The topic nodes ***T*** occupy the interior levels and the leaves are words in *V*. *G*: $\{g_1\ (\alpha_1), g_2\ (\alpha_2), \ldots, g_z(\alpha_z)\}$: g_i parameterized by α_i, is a Dirichlet distribution associated with topic z_i. α_i is a vector with the same dimension as the number of children in z_i, specifying the correlation among them.

In this model, for each word ***w*** in a document *d*, a topic path z_w is sampled based on the multinomial distribution $\Phi_{z_1}, \Phi_{z_1}, \ldots, \Phi_{z_n}$. Simultaneously, a stamp y_{wi} is sampled from each topic z_{wi} on the path based on the corresponding Beta distribution $\psi_{z_{wi}}$, where the parameters of a Beta distibution for topic *z* are denoted by Ψ_z. Finally, the probability of generating a corpus with timestamps is the product of the probability for every document:

$$p(d,y|\alpha,\psi) = \Pi_d p(d,y^{(d)}|\alpha,\psi), \tag{6}$$

which is the product of the probability for every document. The aforementioned generative process associates each word with multiple timestamps that are sampled for different topics. Initially each time stamp is shared by all the words of the training document; however, in the generative process of continuous-time model, different time stamps can be generated for every word in a document. Fig. 5 shows a four-level hierarchy consisiting of one root topic *rn*, z_2 (super-topic) at the second level, z_3 (sub-topics) at the third level and words at the bottom. Here, the root is connected to all super-topics, super-topics are fully connected to sub-topics and sub-topics are fully connected to the words at the bottom. For parameter estimation and approximate inference a Gibbs sampling algorithm is used in Continuous-Time Model.

3.5 Supervised PDPTMs (SuPDPTMs)

LDA and its extensions were proven as a major revolution in the history of topic modeling for discovering information in the world of unsupervised learning. It was extended to Correspondence LDA (Corr-LDA) [49], which is based on a mixture of Gaussian multinomial mixture model and Gaussian multinomial LDA. The mixture model (Corr-LDA) is a solution to the problem of modeling annotated data with multiple types where the instance of one type such as a caption serves as a description of the other type such as an image. Corr-LDA was proven as a useful automatic annotation and text-based image retrieval approach.

Corr-LDA [49] was proposed for modeling annotated data of images and making predictions. A similar kind of probabilistic model, Labeled LDA (LLDA) [50] was proposed for clustering genes with soft clustering abilities, allowing one gene to belong to more than one cluster, to depict a more functional classification of genes. The LLDA model can also incorporate the annotation of known genes, which make it a labeled or supervised topic model that has additional predictive power.

There was a boom of unsupervised topic models. At that time the inability of unsupervised topic models when prediction is ultimate objective was discussed [12] in comparison with supervised topic models. Consequently, Supervised LDA (sLDA) was proposed, which can be considered as a statistical model of labeled documents with prediction as its main goal. In sLDA every document was paired with a response, which is used to infer latent topics that are predictive of that response. sLDA was used to effectively predict movie ratings from the reviews about movies, and also to predict web page popularity from the text description about the web pages.

SuPDPTMs mainly make use of labeled data (e.g., movies rating by users by giving 3 or 5 stars) for better predictions. They consider that sometime we have important labeled data that should be used in addition to the structure of words present in documents. The main idea behind them are other semi-supervised learning techniques that can result in a better performance by using some supervised or labeled information. They exploit the structure of words and label information together to make better predictions. The framework of selected topic model for SuPDPTMs category is explained below.

- Supervised LDA (sLDA)

sLDA [9] models the documents by taking into account additional labeled information in the form of labeled documents. In sLDA each document is paired with a response variable that differentiates it from previous methods. The response variable can be thought of as a number of stars given to a movie on the basis of its likeliness, as supervised information. sLDA can also be called a statistical model of labeled documents, which performs better than unsupervised topic models such as LDA when prediction is an ultimate goal. sLDA can accommodate various types of responses; such as unconstrained real values, real values constrained to be positive, ordered or unordered class labels, non-negative integers and other types.

Figure 10 shows a family of distributions corresponding

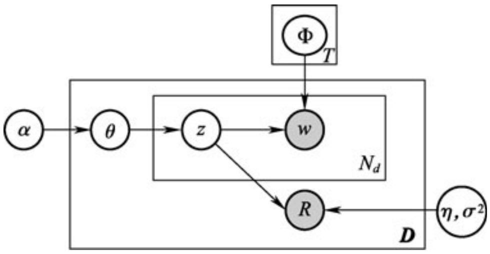

Fig. 10 sLDA model

to the generative process of sLDA. For the case $r \in \boldsymbol{R}$ fix for a moment the model parameters: the $\boldsymbol{T}$ topics $\Phi_{1:n}$ (each Φ_z a vector of term probabilities), the Dirichlet parameter α, and the response parameters η and σ^2. In the generative process of sLDA, the document d is generated first, under exchangeability assumption, and then the response variable r is generated based on the previously generated document. Thus, the response depends on the topic frequencies that actually occurred in the document, rather than the mean of the distribution of generating topics. sLDA treats α, Φ, η and σ^2 as unknown constants to be estimated rather than random variables. For approximate maximum likelihood estimation of sLDA model, the variational Expectation-Maximization (EM) procedure is adopted.

4 Parameter estimation and inference making algorithms

The goal of parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data, and the goal inference making is to estimate the values (hidden topics) of unseen documents by using the values (hidden topics) of the observed documents. If roots are observed documents in a model, and we try to predict the leaves, this is called prediction, or top- down reasoning. If leaves are observed documents in a generative model, and we try to infer the hidden roots, this is called diagnosis, or bottom-up reasoning. Bayesian networks can be used for both aforementioned tasks [51]. In basic topic modeling, the topic-word distribution Φ and the topic distribution θ are the main variables of interest for each document. Here, we provide a brief summary of some popular parameter estimation algorithms for estimating Φ and θ and inference making.

Expectation Maximization (EM) is an exact parameter estimation approach [52] that can be used to directly estimate Φ and θ variables. In PLSA [3], an EM approach was used to estimate parameters directly; however the EM approach suffers from the problems of a large number of parameters growth with scale of input data and local maxima of likelihood function. Therefore, variational methods are used [8,53] which are a special case of EM methods, when computation of the posterior distribution of the hidden variables in a given document is intractable for exact inference. Variational methods provide tight lower and upper bounds to directly estimate the posterior distribution of the hidden variables given the document, instead of directly estimating Φ and θ variables. Variational methods such as Expectation-propagation [20], Variational Extensions to EM [54], and Variational Expectation Maximization (VEM) [8] are used when direct inference is intractable.

Markov Chain Monte Carlo (MCMC) [55,56] has a set of approximate iterative approaches, which are very useful and highly efficient for sampling values in case of high dimensional distributions. Gibbs sampling is one of the important approaches of MCMC, and can be applied to construct a Markov chain that converges to the posterior distributions on topic z. The results are then used to infer Φ and θ variables indirectly; it provides an easy implementation to discover hidden topics from a large collection of documents [10]. It also provides more efficient estimation procedures than variational approximation methods in most of the cases [5,10,19,57]. Details and applications of all aforementioned estimation and inference making approaches can be found in the references mentioned above.

Variational and MCMC methods are commonly used within an EM framework. Because the applications are large scale, the two important procedures, variational Bayes and Gibbs sampling, are not computationally efficient. Consequently, the collapsed variational Bayesian (CVB) inference algorithm [58] is proposed for LDA. It is computationally efficient, easy to implement and significantly more accurate than standard variational Bayesian inference. Gaussian approximation is used for achieving computational efficiency, which was so accurate that exact summation is not needed. The main idea of CVB is that instead of assuming parameters to be independent from latent variables, their dependence on the topic variables is treated in an exact manner. The factorization assumptions made by CVB are more relaxing than those made by variational Bayes; as a result the approximation is more precise.

5 Performance evaluation measures

There are a number of ways that can be used to analyze the performance of topic models. Most important,of which is perplexity [8], which is also used to find the number of topics by showing the generalization power of model on unseen data. It does not require a priori categorization, and was originally used in language modeling [59]. It is used to estimate a model on a subset of a corpus and then the estimated model is used for prediction on an unseen or held out dataset. Lower values of perplexity indicate better generalization power of the model on the words of test documents by the trained topics. For a test set of M documents the perplexity is given in Eq. (7):

$$perplexity(D_{test}) = \exp\left\{\frac{\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M}N_d}\right\}. \quad (7)$$

Entropy (under root of perplexity) can be used to measure the quality of discovered topics, which reveals the purity of topics. Entropy is a measure of the disorder of the system, less intra-topic entropy is usually better. Alternatively, Symmetric KL (*sKL*) divergence [9] can also be used to measure the quality of topics, in terms of inter-topic distance. *sKL* divergence is used here to measure the relationship between two topics, higher inter-topic *sKL* divergence (distance) is usually better. Low entropy or higher *sKL* divergence means less sparse topics (is equivalent to the high generalization power of the model) which can result in better performance of topic model when used for object ranking in information retrieval. That is discussed for expert finding issue in Ref. [34].

$$Entropy\ of\,(Topic) = -\sum_z P(z)\log_2[P(z)]. \quad (8)$$

$$sKL(Topic\ i,Topic\ j)$$
$$= \sum_{z=1}^{T}\left[\theta_{iz}\log\frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz}\log\frac{\theta_{jz}}{\theta_{iz}}\right]. \quad (9)$$

$$Precision = \frac{Correct\ Answers}{Answers\ Produced}. \quad (10)$$

$$Recall = \frac{Correct\ Answers}{Total\ Possible\ Correct}. \quad (11)$$

Perplexity, Entropy and *sKL* divergence can be used to evaluate topic models in terms of their better generalization ability on unseen data and producing better topics (clusters), respectively. They are not statistically significant measures when they are used for object ranking in information retrieval [60]. Recently, Ref. [34] showed the relationship between low perplexity (2^{entropy})† and object ranking in information retrieval. Nevertheless one can use labeled data to evaluate model performance in terms of precision and recall [59], which is demonstrated for the expert finding problem in Ref. [60].

6 Applications

Topic models have been applied to solve diverse kinds of problems in modeling text corpora, such as topic discovery, document classification and indexing, entities relationship discovery, temporal topic trends and community discovery etc. Table 5 provides a concise summary about the applications of topic models in several problem domains.

The topic discovery problem domain is aimed at finding hidden topics of documents that are their representatives in addition to the given titles of documents. Bag of words assumption of models helps us to capture implicit relationships between the words by considering their polysemy. In the past, many efforts have been made to find latent topics [8,16–19,50] by considering implicit dependencies to provide better understanding of semantics-based information hidden in the text of documents. Some efforts have been made by taking into account the Markov dependencies [37,38,41] to discover latent topics from corpora, while Ref. [27] captured topic correlations, in addition to implicit dependencies to discover semantically related topics.

In text classification problem the purpose is to classify the documents into two or more mutually exclusive classes or clusters. In document indexing the focus is on finding the most related documents to a query by ranking them in order. Models such as Refs. [3,8,19,28,36,49] have proved their effectiveness for document classification and indexing tasks on datasets from different domains. They captured the hidden structures of the documents on the basis of implicit relationship between the words of documents, while most of the clustering methods other than topic models usually use distance measures such as Euclidian distance. As a result they are unable to capture the semantics-based information present between the words of a document.

† http://w3.msi.vxu.se/~nivre/teaching/statnlp/langmodel/entropy.html

Table 5 Summary of DPTMs applications

Models	Type	Parameter Estimation and Inference Making Algorithms	Problem Domain (s)	Dataset (s)
PLSA	BDPTMs	EM	Ranking (automatic document indexing)	LOB corpus, MED abstract dataset, CRAN abstracts dataset, CACM abstracts dataset, CISI abstracts dataset
A Joint Probabilistic Model	IrCDPTMs	EM	Document Classification, Relationship between Topics and Links	Webkb web pages dataset (http://www.cs.cmu.edu/~webkb/), Cora abstracts dataset (http://www.cora.justresearch.com)
A probabilistic Approach	BDPTMs	Gibbs Sampling	Topic Discovery (semantics of words)	TASA corpus “a collection of children reading”
LDA	BDPTMs	Variational EM	Topic Discovery, Document Classification, Collaborative Filtering	TREC AP newswire articles corpus, Reuters news articles dataset (http://www.daviddlewis.com/resources/testcollections/reuters21578/), C Elegants Literature (http://elegans.swmed.edu/wli/cgcbib), EachMovie collaborative filtering dataset
A Topic Model	BDPTMs	Gibbs Sampling	Topic Discovery (semantics of words)	TASA corpus “a collection of children reading”
Corr-LDA	SuDPTMs	Variational EM	Automating Annotation, Text-based Image Retrieval	Corel images and caption dataset
discrete (PCA)		Gibbs Sampling	Text classification, Information Retrieval	20 Newsgroup dataset (http://www.ai.mit.edu/_jrennie/20Newsgroups/), Reuters news articles dataset (http://www.daviddlewis.com/resources/testcollections/reuters21578/)
Mixed-Membership Models	IrCDPTMs	EM	Topic Discovery, Document Classification	PNAS scientific articles dataset (http://www.pnas.org)
Author-Topic Model	IrCDPTMs	Gibbs Sampling	Entities and Topics Correlations, Topics Evolution over Time	Cite seer dataset (http://citeseer.ist.psu.edu/oai.html)
ART Model	IrCDPTMs	Gibbs Sampling	Topic and Role Discovery	Enron email dataset (http://www.cs.cmu.edu/~enron/), Researchers email achieve
A Composite Model (HMM-LDA)	IaCDPTMs	Gibbs Sampling	Document Classification, Part-of-Speech Tagging	Brown and TASA corpus “a collection of children reading” datasets, NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html)
LLDA Model	SuDPTMs	Variational EM	Topic Discovery	Microarray dataset (http://genomics.lbl.gov/~patrickf/llda.html)
CTM	IrCDPTMs	Variational EM	Topics Correlations	JSTOR science articles dataset (http://www.jstor.org)
DTM	TDPTMs	Variational Kalman Filtering	Topics Evolution over Time	JSTOR science articles dataset (http://www.jstor.org)
Statistical Entity-Topic Models	IrCDPTMs	Gibbs Sampling	Entities and Topics Correlations	New York Times dataset (http://www.ldc.upenn.edu), Foreign broadcast information service FBIS dataset (http://www.fbis.gov)
Bigram Topic Model	IaCDPTMs	Gibbs EM	Topic Discovery	Psychological review abstracts dataset (http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm), 20 News group dataset (http://people.csail.mit.edu/jrennie/20Newsgroups/)
PAM	IrCDPTMs	Gibbs Sampling	Super and Sub Topic Discovery, Document Classification	NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html) , 20 Newsgroup dataset (http://www.cs.cmu.edu/~textlearning/), Rexa research paper search engine (http://Rexa.info)
TOT Model	TDPTMs	Gibbs Sampling	Topics Evolution over Time	State of the Union Addresses dataset (http://www.gutenberg.org/dirs/etext04/suall11.txt), Researchers Email Achieve, NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html)

(*Continued*)

Models	Type	Parameter Estimation and Inference Making Algorithms	Problem Domain (s)	Dataset (s)
Continues-Time Model	TDPTMs	Gibbs Sampling	Topics Evolution over Time and their Correlations	Rexa research paper search engine (http://Rexa.info)
CPLSA	IrCDPTMs	EM	Temporal (Entities-Topic) Correlations, Topics Evolution over Time, Event Impact Analysis	Abstracts of 282 papers of two Data Mining researchers, from ACM Digital library, MSN Space documents, Abstracts of 28 years' SIGIR conferences from ACM Digital Library
HTMM	IaCDPTMs	EM and Forward-backward algorithm	Topic Discovery	NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html), used dataset (http://www.cs.huji.ac.il/~amitg/htmm.html)
MTTM	TDPTMs	Variational EM	Topics Evolution over Time	JSTOR science articles dataset (http://www.jstor.org)
sLDA Model	SuDPTMs	Variational EM	Ranking Movies and Web Pages	News paper movie reviews dataset (http://www.cs.cornell.edu/people/pabo/movie-review-data/), Digg Links (digg.com)
Citation Influence Model	IrCDPTMs	Gibbs Sampling	Citation Influence	Cite seer dataset (http://citeseer.ist.psu.edu/oai.html)
GWN-LDA Model	IrCDPTMs	Gibbs Sampling	Entities and Topics Correlations	NanoSci articles dataset (2000-2006) taken from (http://scientific.thomson.com/products/sci/), Cite seer dataset (http://citeseer.ist.psu.edu/oai.html)
TNG Model	IaCDPTMs	Gibbs Sampling	Topic Discovery, Information Retrieval	TREC dataset, NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html),
Link-PLSA-LDA	IrCDPTMs	Variational EM	Blogs Influence	Nielsen Buzz metrics blogs postings dataset (http://www.nielsenbuzzmetrics.com)
cDTM	TDPTMs	Variational Kalman Filtering	Topics Evolution over Continuous Time	TREC-1 AP newswire articles corpus, "Election 08" dataset (digg.com)
LTHM	IrCDPTMs	EM	Relationship between Topics and Links	Webkb web pages dataset (http://www.cs.huji.ac.il/~amitg/lthm.html), Wikipedia (http://www.cs.cmu.edu/~webkb/)
TAT	TDPTMs	Gibbs Sampling	Temporal Authors Interests and Correlations	Computer science research papers taken from http://www.informatik.uni-trier.de/~ley/db/
ACT	IrCDPTMs	Gibbs Sampling	Expertise Search in Academics Social Network	Computer science research papers taken from http://www.arnetminer.org/
STMS	IrCDPTMs	Gibbs Sampling	Expert Finding	Computer science research papers taken from http://www.informatik.uni-trier.de/~ley/db/
GLDA	IrCDPTMs	Gibbs Sampling	Conference Mining	Computer science research papers taken from http://www.informatik.uni-trier.de/~ley/db/

Relationships between the entities in a network exist which build up communities in social networks, as information is hidden in the latent structure of documents and links of complex network structures. It is a challenging task to capture those relationships between entities. Topic models such as Refs. [14,22,29,30,33,35] have been used to discover entities and communities with respect to their relationships with topics by considering the latent structures of the documents.

The ART model [26] discovered the role of entities in an organization network with respect to their jobs on the basis of email messages text, and also by exploiting senders and receivers directed relationships, while in the past this kind of roles were discovered only on the basis of directed links without paying attention to the text based semantics of the email messages. Relationships between documents are considered for citation suggestion on the basis of citations given in research papers by using the Citation Influence model [31]. Sometime a model is dependent on the dataset as in the case of the ART model [26] which needs email messages with senders and receivers email address plus text sent or received or the Citation Influence model [31] which needs citations of papers to model the relationships between papers. It is not applicable to a dataset which has no citation information with text e.g. it cannot be applied to JSTOR science articles because they have no citations or references information. Daud et al. [34] used semantics and temporal information based topic modeling approach for expert finding in academics social networks. In STMS, influence of conferences and time factor is modeled together, both of which are important aspects of expertise modeling. As authors publishing in high class conferences are more appropriate to be found as experts and time factor modeling is needed to get the experts of different years for all topics. While without modeling time, when modeling is done for each year independently, model faces the problem of topic exchangeability.

Discovery of temporal trends or the evolution of topics is an interesting problem discussed to become familiar with the evolution of research trends in a research community, as well as to identify the changing tastes of users with respect to their eating and clothing habits. In a research community, temporal author topic relationships are modeled and effects of events are analyzed [39]. Specifically the problem of topic evolution over time is investigated [14,43–45,47] by providing a diverse type of solutions. Continuous-Time model [15] modeled topic relationships by also considering the evolution of topics over time to provide better insights of temporal topic trends in a researcher's community. TAT is used to rank authors for different years according to their interests [48]. Models proposed for the evolution of topics are dependent on the dataset because they need timestamps (e.g., month or year) of documents for temporal modeling. Finding blogs influence [25], document summarization [61], multi-document summarization [62] and web spam filtering [63] problems are also investigated by using these models.

Other than text based modeling, topic models are also applied to content-based image clustering [64], object recognition [65,66], hand written character recognition [67] and other applications of computer vision in general. Microarray, genes and cells data analysis [50] in biological data are some other applications of these models.

7 Research issues and future directions

In this section, we will discuss research issues, open challenges and future directions in the field of topic modeling. We categorized the challenges of modeling text corpora into four types.

The first type of challenges comes from the need for finding hidden structures of data by considering Markov relations between the contexts of the documents, which can provide a better understanding of the topics hidden in the text corpora. According to Ref. [37] incorporation of HMM into the LDA model is related to other extensions of LDA such as Refs. [14,36]. Hence it can be interesting to combine different extensions of models to form a better text corpora model. In the Author-Topic model [14] stylistic features of the text contents for authors of the documents are not considered. By successfully combining stylistic features with topics, a more realistic author's classification can be achieved. In fact Markov dependencies can be more useful but they demand a deep understanding of statistical language processing to explore highly effective new solutions.

The second type of challenges comes from the usage of explicit links (e.g., citations) between the documents to find better associations between documents, researchers and social networks. Undoubtedly Refs. [25,37] have given useful insights to cope with documents correlations, but the Link-PLSA-LDA [25] approach faces the problem of large number of parameters growth just like PLSA [3] and the HTMM [37] approach cannot take into account the text content of linked documents. One can solve these two problems in a new model to provide an effective solution for the usage of links between documents.

The third type of challenges comes from the usage of time stamps to find temporal trends in documents and other entities together with considering contents of documents. Continuous time modeling is important and one of the new solutions to this is in Ref. [43], which does not care for syntax dependencies explicitly. However, it is somehow important to capture multiple meanings (polysemy) of words and also to deal with the discretized time, which has to be modeled as continuous. Continuous-Time Model [15] has considered time and syntax dependencies by using directed acyclic graph, which is not the case in the real world. There is a need to explore models with consideration of both time and syntax dependencies but not with the limitations of directed acyclic graph. In addition, explicit relationships between documents through the utilization of topic correlations of time stamps can be a useful step forward.

The fourth type of challenges comes from the usage of labeled documents by also considering implicit information in the context of the documents to make better predictions. sLDA model [12] is an effective move to handle these types of problems by adding responses; however, sLDA is unable to consider the semantic information of the syntax. One can integrate explicit semantic information of syntax in the sLDA model to provide a better supervised model of labeled documents for making more useful predictions, and if temporal trends analysis is the case then time stamps can also be modeled in the extended model.

From an application point of view, different solutions for the above mentioned four challenges can be integrated into mixture models to provide more practical solutions which can obtain superior results, as it was the case in the past; however deeper insights are required for developing the understanding of statistical language learning. There are two main objectives. One is how to choose and integrate different approaches to make their advantages together in one compact solution. The other is to define new DPTMs, by investigating the opportunities to which new models can be applied to attain optimal results.

From a theoretical point of view, parameter estimation and inference making procedures are a very important part of topic models, and new procedures like collapsed variational Bayesian inference algorithm [58] have been recently proposed to increase the computational efficiency of LDA. Model Convergence and choosing suitable hyper-parameters for DPTMs still need to be investigated in more detail. How to propose specific purpose models like ART and Citation Influence [31,26] and interpret results obtained by using different models are also some of the areas that demand future investigation.

8 Conclusions

In this paper, we discussed the important DPTMs for modeling text corpora by providing their classification into five categories with respect to their main functionalities and a general framework of selected models from all categories is explained. We discussed parameter estimation and inference making algorithms for topics extraction, ploysemy with topics and performance measures for topic models. We discussed the applications of topic models for modeling text corpora. We investigated several open problems and future directions for modeling text corpora. As a whole in this paper, we presented a snapshot of research work done in the last decade about PDPTMs, their applications and future challenges around the time of its writing. However, we do believe that the core information and models presented here will be useful for the researchers in this area of research now and in future too. As a future work, one can write a comprehensive survey about undirected or probabilistic topic models.

Acknowledgements The work was supported by the National Natural Science Foundation of China (Grant Nos. 90604025, 60703059), Chinese National Key Foundation Research and Development Plan (2007CB310803) and Higher Education Commission (HEC), Pakistan. We are thankful to Jie Tang, Jing Zhang, Feng Wang, Bo Wang, Liu Liu, Zi Yang and Jun Li for their valuable discussions and suggestions. Especially we are thankful to Wim De Smet for helping us to improve English writing and anonymous reviewers for their valuable suggestions, which has really improved the contents and structure of the paper to a high extent.

References

1. Popescul A, Flake G W, Lawrence S, Ungar L H, Giles C L. Clustering and identifying temporal trends in document databases. IEEE ADL, 2000, 173–182
2. McCallum A, Nigam K, Ungar L H. Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the 6th ACM SIGKDD, 2000, 169–178
3. Hofmann T. Probabilistic latent semantic analysis. In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, July 30-August 1, 1999
4. Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer T, Mcnamara D, Dennis S, Kintsch W (Eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2007
5. Heinrich G. Parameter Estimation for Text Analysis. Technical report, Version 2, February 2008
6. Smolensky P. Information processing in dynamical systems: foundations of harmony theory. In: Rumehart D E,McClelland J L (Eds), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. McGraw-Hill, New York, 1986
7. Welling M, Rosen-Zvi M, Hinton G. Exponential family harmoniums with an application to information retrieval. In: Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2004
8. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022
9. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI), Banff, Canada, July 7–11, 2004
10. Griffiths T L, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences. USA, 2004, 101: 5228–5235
11. Teh Y W, Jordan M I, Beal M J, Blei D M. Hierarhical Dirichlet Processes. Technical Report 653, Department of Statistics, UC Berkeley, 2004
12. Blei D M, McAuliffe J. Supervised topic models. In: Advances in Neural Information Processing Systems (NIPS) 21. Cambridge, MA, MIT Press, 2007, 121–128
13. Buntine W L. Operations for learning with graphical models.

Journal of Artificial Intelligence Research, 1994, 2: 159–225

14. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, August 22–25, 2004
15. Wang X, Li W, McCallum A. A continuous-time model of topic co-occurrence trends. In: AAAI Workshop on Event Detection. Boston, Massachusetts, USA, July 16–20, 2006
16. Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Journal of Machine Learning, 2000, 39(2–3): 103–134
17. Griffiths T L, Steyvers M. A probabilistic approach to semantic representation. In: Proceedings of the 24th Conference of the Cognitive Science Society. USA, 2002
18. Griffiths T L, Steyvers M. Prediction and semantic association. In: Advances in Neural Information Processing Systems (NIPS) 15. Cambridge, MA, MIT Press, 2003
19. Wray L, Buntine, Jakulin A. Applying discrete PCA in data analysis. In: Proceedings of 20th Conference on Uncertainty in Artificial Intelligence (UAI), Banff, Canada, July 7–11, 2004, 59–66
20. Minka T, Lafferty J. Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI), Alberta, Canada, August 1–4, 2002, 352–359
21. Hofmann T, Puzicha J, Jordan M I. Learning from dyadic data. In: Advances in Neural Information Processing Systems (NIPS) 11. Cambridge, MA, MIT Press, 1999
22. Cohn D, Hofmann T. The missing link- a probabilistic model of document content and hypertext connectivity. In: Advances in Neural Information Processing Systems (NIPS) 13. Cambridge, MA, MIT Press, 2001
23. Blei D M, Moreno P J. Topic segmentation with an aspect hidden Markov model. In: Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans. LA USA, September 9-13, 2001, 343–348
24. Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. In: Proceedings of the National Academy of Sciences, USA, 2004, 101: 5220–5227
25. Nallapati R, Cohen W. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In: Proceedings of International Conference for Weblogs and Social Media, Seattle, Washington, USA, March 30-April 2, 2008
26. McCallum A, Corrada-Emmanuel A, Wang X. The Author-recipient-topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Technical Report UM-CS-2004-096, 2004
27. Blei D M, Lafferty J. Correlated topic models. In: Advances in Neural Information Processing Systems (NIPS) 18. Cambridge, MA, MIT Press, 2006, 147–154
28. Li W, McCallum A. Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, June 25-29, 2006, 577–584
29. Newman D, Chemudugunta C, Smyth P, Steyvers M. Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20–23, 2006, 680–686
30. Zhang H, Giles C L, Foley H C, Yen J. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proceedings of 22nd AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, July 22–26, 2007, 663–668
31. Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences. In: Proceedings of 24th International Conference on Machine Learning (ICML), Corvallis, Oregon, USA, June 20–24, 2007
32. Gruber A, Rosen-Zvi M, Weiss Y. Latent topic models for hypertext. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 9–12, 2008
33. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z. ArnetMiner: extraction and mining of academic social networks. In: Proceedings of ACM SIGKDD, 2008
34. Daud A, Li J, Zhu L, Muhammad F. A generalized topic modeling approach for maven search. In: Proceedings of International Asia-Pacific Web Conference and Web-Age Information Management (APWEB-WAIM), Suzhou, China, 2009
35. Daud A, Li J, Zhu L, Muhammad F. Conference mining via generalized topic modeling. In: Proceedings of European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML PKDD), Bled, Slovenia, 2009
36. Griffiths T L, Steyvers M, Blei D M, Tenenbaum J B. Integrating topics and syntax. In: Advances in Neural Information Processing Systems (NIPS) 17. Cambridge, MA, MIT Press, 2005, 537–544
37. Gruber A, Rosen-Zvi M, Weiss Y. Hidden topic Markov models. In: Proceedings of Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico, USA, March 21–24, 2007
38. Wallach J M. Topic modeling: Beyond bag-of-words. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA, June 25–29, 2006
39. Mei Q, Zhai C X. A mixture model for contextual text mining. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20–23, 2006, 649–655
40. Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman

R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391–407

41. Wang X, McCallum A, Wei X. Topical N-grams: phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), Omaha NE, USA, October 28–31, 2007
42. Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE, 1989, 77(2): 257–286
43. Blei D M, Lafferty J. Dynamic topic models. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA, June 25–29, 2006
44. Nallapati R, Cohen W, Ditmore S, Lafferty J, Ung K. Multiscale topic tomography. In: Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12–15, 2007
45. Wang C, Blei M D, Heckerman D. Continuous time dynamic topic models. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 9–12, 2008
46. Uhlenbeck G E, Ornstein L S. On the theory of Brownian motion. Physics Reviews, 1930, 36: 823–841
47. Wang X, McCallum A. Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20–23, 2006
48. Daud A, Li J, Zhu L, Muhammad F. Exploiting temporal authors interests via temporal-author-topic modeling. In: Proceedings of 5th International Conference on Advance Data Mining and Applications (ADMA), Beijing, China, 2009
49. Blei D M, Jordan M. Modeling annotated data. In: Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28-August 1, 2003, 127–134
50. Flaherty P, Giaever G, Kumm J, Jordan M, Arkin A. A latent variable model for chemogenomic profiling. Bioinformatics, 2005, 21(15): 3286–3293
51. Murphy K. An Introduction to Graphical Models. Technical report, University of California, Berkeley, May 2001
52. Bilmes J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov modals. Berkeley, ICSI TR-97-021, 1997
53. Jordan M I, Ghahramani Z, Jaakkola T S, Saul L K. An introduction to variational methods for graphical models. In: Jordan M (Eds), Learning in Graphical Models. MIT Press, 1998
54. Buntine W. Variational Extensions to EM and Multinomial PCA. In: Elomaa T et al. (Eds.): ECML, LNAI 2430, Springer-Verlag, Berlin, 2002, 23–34
55. Gilks W R, Richardson S, Spiegelhalter D J. Markov Chain Monte Carlo in Practice. London: Chapman & Hall, 1996
56. Andrieu C, Freitas N D, Doucet A, Jordan M. An introduction to MCMC for machine learning. Journal of Machine Learning, 2003, 50: 5–43
57. Erosheva E A. Grade of membership and latent structure models with applications to disability survey data. Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University, 2002
58. Teh Y W, Newman D, Wellingm M. A collapsed variational Bayesian inference algorithm for latent dirichlet allocation. In: Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2006
59. Azzopardi L, Girolami M, Risjbergen K V. Investigating the relationship between language model perplexity and IR precision-recall measures. In: Proceedings of the 26th ACM SIGIR, Toronto, Canada, 2003
60. Zhang J, Tang J, Liu L, Li J. A mixture model for expert finding. In: Proceedings of the PAKDD, Washio T et al. (Eds). LNAI, 2008, 5012: 466–478
61. Chang Y L, Chien J T. Latent dirichlet learning for document summarization. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009
62. Arora R, Ravindran B. Latent dirichlet allocation based multi-document summarization. In: Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Rext Data, 2008
63. Bíró I, Szabó J, Benczúr A A. Latent dirichlet allocation in web spam filtering. In: Proceedings of the Adversarial Information Retrieval on the Web (AIRWeb'08), 2008
64. Elango P K, Jayaraman K. Clustering images using the latent dirichlet allocation model, 2005
65. Wang Y, Mori G. Human action recognition by semi-latent topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision (T-PAMI), 2009
66. Wang Y, Sabzmeydani P, Mori G. Semi-latent dirichlet allocation: A hierarchical model for fuman action recognition. In: 2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation (ICCV), 2007
67. Rath T M, Lavrenko V, Manmatha R. A Statistical Approach to Retrieving Historical Manuscript Images Without Recognition. Technical Report, 2003