Conceptualization topic modeling

Yi-Kun Tang¹,² · Xian-Ling Mao¹ · Heyan Huang¹ · Xuewen Shi¹ · Guihua Wen³

Received: 3 August 2017 / Revised: 20 August 2017 / Accepted: 22 August 2017
© Springer Science+Business Media, LLC 2017

Abstract Recently, topic modeling has been widely used to discover the abstract topics in the multimedia field. Most of the existing topic models are based on the assumption of three-layer hierarchical Bayesian structure, i.e. each document is modeled as a probability distribution over topics, and each topic is a probability distribution over words. However, the assumption is not optimal. Intuitively, it’s more reasonable to assume that each topic is a probability distribution over concepts, and then each concept is a probability distribution over words, i.e. adding a latent concept layer between topic layer and word layer in traditional three-layer assumption. In this paper, we verify the proposed assumption by incorporating the new assumption in two representative topic models, and obtain two novel topic models. Extensive experiments were conducted among the proposed models and corresponding baselines, and the results show that the proposed models significantly outperform
the baselines in terms of case study and perplexity, which means the new assumption is more reasonable than traditional one.

**Keywords** Conceptualization topic modeling · Hierarchical Bayesian structure · Conceptualization latent Dirichlet allocation · Conceptualization labeled latent Dirichlet allocation

## 1 Introduction

In recent years, topic modeling is becoming more and more popular in identifying latent semantic components in the multimedia field. Lots of topic models have been proposed. The existing topic models can be divided into four categories: *Unsupervised non-hierarchical topic models* [1, 5, 7, 31], *Unsupervised hierarchical topic models* [2, 10, 25], and their corresponding supervised counterparts [12, 13, 21, 24].

The basic assumption of most existing topic models is that each document is modeled as a probability distribution over topics, and each topic is directly a probability distribution over words, i.e. three-layer hierarchical Bayesian structure, shown in Fig. 1a. However, this assumption is not optimal, because it does not consider the importance of the concepts in topics. Concepts are very important in natural language and textual semantic understanding. Concepts can also help people better understand knowledge, as psychologist Gregory Murphy wrote: “Concepts are the glue that holds our mental world together” [16].

Intuitively, it’s more reasonable that if we add a latent concept layer between topic layer and word layer in traditional three-layer assumption, i.e. a four-layer hierarchical Bayesian structure, shown in Fig. 1b. In this novel assumption, each document is considered as a probability distribution over topics, each topic is a probability distribution over concepts, and each concept is a probability distribution over words. The assumption is similar to the writing process. For example, if we want to write a article about the topic “military”, we then focus on the concepts related to the topic, such as army, navy and air force. Finally, we select related words from these concepts, maybe the word tank from the concept army, the word torpedo from the concept navy, and the word fighter from the concept air force.

As we known, Latent Dirichlet Allocation (LDA) [1] is the beginning of topic modeling, and is the most important component in all kinds of topic models. If the novel assumption

![Fig. 1](image-url) a Three-layer hierarchical Bayesian structure of existing topic models; b Four-layer hierarchical Bayesian structure of conceptualization topic modeling

---

![Springer](logo-url)
performances better than the traditional one in LDA, it’s reasonable to infer that the novel assumption is more suitable for topic modeling than the traditional one. Thus, in this paper, we first propose a novel topic model, called Conceptualization Latent Dirichlet Allocation (CLDA), which applies the novel four-layer assumption in LDA, to verify our assumption. Furthermore, we also apply the novel assumption in a supervised topic model, Labeled LDA (LLDA) [21], to proof the novel assumption is more effective. We call this novel supervised topic model Conceptualization Labeled Latent Dirichlet Allocation (CLLDA). The distribution of each concept over words in our models can be obtained from Probase knowledge base [27], which is a universal probabilistic taxonomy concept knowledge base.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, two novel topic models, CLDA and CLLDA, are proposed by using new four-layer assumption. Extensive experiments on two real datasets are introduced in Section 4. Finally, we conclude the paper in Section 5.

2 Related work

2.1 Topic modeling

Topic modeling has been widely used to discover the abstract topics in the multimedia applications. Multi-modal event topic model (mmETM) [19] can capture multi-modal topics of social events and can obtain the evolutionary trends of social events and generate effective event summary details over time. Twitter-Network (TN) topic model [11] can jointly model the text and the social network in a full Bayesian nonparametric way. Dynamical causal topic model (DCTM) [6] mines activities from complex video surveillance scenes. Visual Sentiment Topic Model (VSTM) [3] can obtain the sentiments implied in Microblog images.

The existing topic models can be divided into four categories: Unsupervised non-hierarchical topic models, Unsupervised hierarchical topic models, and their corresponding supervised counterparts.

Unsupervised non-hierarchical topic models are widely studied, such as LDA [1], Probase-LDA [30], TCC [9] and COT [31] etc. The most famous one is Latent Dirichlet Allocation (LDA). LDA is similar to pLSA [7], except that in LDA the topic distribution is assumed to have a Dirichlet prior. Probase-LDA [30] is a state-of-the-art knowledge-based topic model by incorporating knowledge graph embeddings into topic modeling, but it has to set the concept number artificially.

However, the above models cannot capture the relation between super and sub topics. To address this problem, many models have been proposed to model the relations, such as Hierarchical LDA (HLDA) [2], Hierarchical Dirichlet processes (HDP) [25], Hierarchical PAM (HPAM) [15], PIE [10] and Guided HTM [23] etc. The relations are usually in the form of a hierarchy, such as the tree or Directed Acyclic Graph (DAG).

Although unsupervised topic models are sufficiently expressive to model multiple topics per document, they are inappropriate for labeled corpora because they are unable to incorporate the observed labels into their learning procedure. Several modifications of LDA to incorporate supervision have been proposed in the literature, such as Labeled LDA [21], Prior-LDA [22], Partially LDA (PLDA) [20], NTM [4], DOLDA [12] and Labeled Phrase LDA [24] etc.

None of these non-hierarchical supervised models, however, leverage on dependency structure, such as parent-child relation, in the label space. Lots of models, such as hLLDA [18], HLSLDA [17], SSHLLDA [13], SHDP [32] and EHL LDA [14], have been proposed to solve the problem.
All of these topic models are mainly based on the assumption of three-layer hierarchical Bayesian structure. However, the assumption is not optimal. Intuitively, it’s more reasonable to add a latent concept layer between topic layer and word layer in traditional three-layer assumption. Different from latent concept topic model (LCTM) [8] which adds another layer of latent variables to indicate the conceptual similarity of words, our proposed assumption can combine not only the characteristics of the datasets, but also the concept knowledge base. Therefore, it is considered to be better than LCTM. In this paper, we will verify the proposed assumption by incorporating the new assumption in two representative topic models.

2.2 Concept knowledge base

It is easy for mankind to acquire the meaning of an article and extract the topics of the article, because there is a certain background conceptualized knowledge base in a brain. For example, when seeing a sentence: “Microsoft announced a project named, Microsoft Azure Information Protection”, a man will never mistake *Microsoft* as a person or other things, because we have known that *Microsoft* is a concept about software company.

However, machines cannot conceptualize what they read, which is a great challenge for machines to understand natural language. Concept knowledge base is a kind of knowledge base that uses taxonomies and ontologies to obtain concepts and extract the relationships between instances and concepts. Therefore, Concept knowledge base is a kind of tool to make machines understand nature language [28, 29].

There are many existing concept knowledge bases, such as Probase [26, 27], Freebase and WordNet etc. Among them, Probase is a state-of-the-art one, which contains above 5.4 million concepts that is greater than other concept knowledge bases. The main advantage of Probase is that it is the first to measure the correlation between instances and concepts with probabilities, while other concept knowledge bases use a boolean variable to represent relationships between instances and concepts.

Therefore, in this paper, we use Probase API [26] to get the probability distribution of each concept over words.

3 Conceptualization topic modeling

In this section, we will demonstrate that how to incorporate the four-layer assumption in unsupervised and supervised topic models, to verify the effectiveness of the novel assumption. For unsupervised topic modeling, we choose LDA as the manipulating object because it is the basic component of most existing topic models. For supervised topic modeling, we choose Labeled LDA [21] because it is one of the most representative supervised models.

3.1 Conceptualization LDA

To incorporate the four-layer assumption in LDA, we propose a novel topic model, called Conceptualization LDA (CLDA). It models each document as a mixture of underlying topics. Different from existing topic models, CLDA assumes that each topic is a distribution over concepts rather than directly over words, and regards concepts as distributions over words.
In addition, as for neologisms, which are not in the dictionary of the concept knowledge base, they will be regarded as new concepts. In other words, we define these neologisms as atomic concepts. In CLDA, the distribution of a concept over words is acquired from the concept knowledge base, Probase. The graphical model of CLDA is shown in Fig. 2.

In CLDA, each document consists of a group of words represented as $w^{(d)} = (w_1, ..., w_{N_d})$. $\alpha$ is the parameter of the Dirichlet distribution of the topic prior, and $\theta^{(d)}$ is the parameter of the multinomial distribution of the $d^{th}$ document. $\beta$ is the parameter of the Dirichlet distribution of the concept prior, and $\phi_k$ is the parameter of the multinomial distribution of the $k^{th}$ topic. $\lambda$ is the concept distribution over words gets from Probase. $m$ is the number of words that do not belong to any concept in the concept knowledge base, and $R$ is the size of concept set. $z_{d,i}$ is the latent topic for the concept or atom concept of the $i^{th}$ word in the $d^{th}$ document. $c_{d,i}$ is the concept of the $i^{th}$ word in the $d^{th}$ document.

The generative process of our CLDA is summarized in Algorithm 1. The generative process can be divided into three parts. Firstly, draw the concept and atom concept distribution from Dirichlet distribution for each topic in the datasets (line $1 \sim 2$). Secondly, draw the topic distribution for each document from Dirichlet distribution (line $3 \sim 4$). Finally, to generate the word $w_{d,i}$, we first select a latent topic $z_i$ (line $5 \sim 6$), and then generate a variable $\xi$ from Bernoulli distribution, where 0 indicates the word does not belong to any concept, and 1 indicates the word $w_{d,i}$ belongs to some concepts in the given concept knowledge base. If $\xi$ equals to 0, then generate a word from $Mult(\cdot|\phi_{z_i})$; otherwise, generate a concept from $Mult(\cdot|\phi_{z_i})$, and then select a word from the concept, which conditionally is related to the concept distribution from the knowledge base (line $7 \sim 12$).

3.1.1 Learning and inference

In this section, we use collapsed Gibbs sampling to estimate parameters.

Specifically, if the word $w_{d,i}$ belongs to some concepts in the given concept knowledge base, the sampling probability for a topic and a concept in position $i$ in document $d$ can be expressed as follows:

$$P(z_{d,i} = k, c_{d,i} = j | w, e_{-(d,i)}, z_{-(d,i)}; \alpha, \beta) \propto \frac{\beta_{k,c_{d,i}} + n_{-(d,i),k}}{\sum_{x=1}^{R} \beta_{x} + n_{-(d,i),x}} \cdot \frac{\alpha_k + n^{(d)}_{-(d,i),k}}{\sum_{t=1}^{K} \alpha_t + n^{(d)}_{-(d,i),t}} \cdot P(w_{d,i}|c_{d,i})$$  \hspace{1cm} (1)

![Fig. 2 Graphical model for CLDA](image-url)
Algorithm 1 Generative process for CLDA

1: For each topic $k \in \{1, \ldots, K\}$:
2: \hspace{1em} Generate $\phi_k = (\phi_{k,1}, \ldots, \phi_{k,C}, \phi_{k,C+1}, \ldots, \phi_{k,C+m})^T$
3: \hspace{1em} $\sim \text{Dir}(\cdot|\beta)$
4: For each document $d \in \{1, \ldots, D\}$:
5: \hspace{1em} Generate $\theta^{(d)} = (\theta_1, \ldots, \theta_K)^T \sim \text{Dir}(\cdot|\alpha)$
6: \hspace{1em} For each $i$ in $\{1, \ldots, N_d\}$:
7: \hspace{2em} Generate $z_{d,i} \in \{1, \ldots, K\} \sim \text{Mult}(\cdot|\theta^{(d)})$
8: \hspace{1em} Generate $\xi \sim \text{Bernoulli}$, where $0$ indicates the word $w_{d,i}$ is an atom concept, and $1$ indicates the word $w_{d,i}$ belongs to some concepts in the given concept knowledge base.
9: \hspace{1em} If $\xi = 0$:
10: \hspace{2em} Generate $w_{d,i} \in \{1, \ldots, V\} \sim \text{Mult}(\cdot|\phi_{z_i})$
11: \hspace{1em} Else:
12: \hspace{2em} Generate $c_{d,i} \in \{1, \ldots, R\} \sim \text{Mult}(\cdot|\phi_{z_i})$
13: \hspace{1em} Select a word $w_{d,i}$ from $\lambda$, a probability distribution gets from Probase.

And if the word $w_{d,i}$ does not belong to any concept in the given concept knowledge base, the sampling probability for a topic in position $i$ in document $d$ can be expressed as follows:

$$P(z_{(d,i)} = k|w, e_{-(d,i)}, z_{-(d,i)}; \alpha, \beta) \propto \frac{\beta_{k,e_{d,i}} + n_{(e_{d,i}),k}}{\sum_{x=1}^E \beta_{k,x} + n_{(x),k}} \cdot \frac{\alpha_k + n_{(d,i),k}}{\sum_{t=1}^K \alpha_t + n_{(d,i),t}}. \quad (2)$$

where $e$ is the vector of concepts and atomic concepts related to the words. $e$ denotes a concept or an atomic concept. $E$ is the number of concepts and atomic concepts. $n_{(x),k}$ is the count of concepts and atomic concepts in $e$ without $z_{d,i}$. $n_{-(d,i),k}$ is the number of tokens in $e$ assigned to topic $k$ in document $d$ without $z_{d,i}$, and $n_{-(d,i),t}$ indicates a summation over that dimension. In (1), $n_{(e_{d,i}),k}$ is the count of concept $c_{d,i}$ in topic $k$, that does not include the current assignment $z_{d,i}$. And the conditional probability $P(w_{d,i}|c_{d,i})$ describes the probability of word $w_{d,i}$ in concept $c_{d,i}$, which can be obtained from Probase. In (2), $n_{(e_{d,i}),k}$ is the count of atomic concept $e_{d,i}$, non-concept word, in topic $k$, that does not include the current assignment $z_{d,i}$.

Finally, the parameters can be estimated as follows:

$$\hat{\phi}_{e,k} = \frac{\beta_{e,k} + n_{(e),k}}{\sum_{x=1}^E \beta_{e,x} + n_{(x),k}} \quad (3)$$

$$\hat{\theta}_{k}^{(d)} = \frac{\alpha_k + n_{(d),k}}{\sum_{t=1}^K \alpha_t + n_{(d),t}} \quad (4)$$

where $e$ denotes a concept or an atomic concept. The two equations for parameter estimation are important. We can use the topic-specific distribution $\phi$ to obtain topical abstracts for topics; meanwhile the topic distribution for each document $\theta$ can be used to discover the most relevant topics for a document and find documents with similar topics.
3.2 Conceptualization labeled LDA

The proposed four-layer Bayesian assumption can be used in most of existing topic models, and we have demonstrated that the assumption can be used in unsupervised topic model, i.e. LDA. In this section, we will further demonstrate the use of the assumption in supervised topic modeling. Labeled Latent Dirichlet Allocation (Labeled LDA) [21] which is a classical supervised topic model, will be extended by incorporating conceptualization assumption. The novel model is called Conceptualization Labeled Latent Dirichlet Allocation (CLLDA).

Labeled LDA is very similar to LDA. Different with LDA, Labeled LDA assumes that the topics of each document are restricted to its labels. The topic distribution of each document in Labeled LDA is generated from a Dirichlet distribution, whose dimensionality of the prior parameter is the same as the number of labels of each document, rather than the number of the total topics of the datasets in LDA. Thus, CLLDA is also similar to CLDA.

Specifically, in order to restrict the latent topics to the label set of each document in CLLDA, we define an indicator function $I^{(d)}(k)$ as follows:

$$I^{(d)}(k) = \begin{cases} 1 & \text{if the } k^{th} \text{ topic is in the label set of the } d^{th} \text{ document.} \\ 0 & \text{otherwise.} \end{cases}$$  (5)

If the word $w_{d,i}$ belongs to some concepts in the given concept knowledge base, the sampling probability for a topic and a concept in position $i$ in document $d$ can be expressed as follows:

$$P(z_{d,i} = k, c_{d,i} = j | w, e_{-(d,i)}, z_{-(d,i)}; \alpha, \beta) \propto I^{(d)}(k) \cdot \frac{\beta_{k,c_{d,i}} + n^{(e_{d,i})}_{-,(d,i),k}}{\sum_{x=1}^{E} \beta_{k,x} + n^{(c)}_{-,(d,i),k}} \cdot \frac{\alpha_k + n^{(d)}_{-,(d,i),k}}{\sum_{t=1}^{K} \alpha_t + n^{(d)}_{-,(d,i),t}} \cdot P(w_{d,i} | c_{d,i})$$  (6)

And if the word $w_{d,i}$ does not belong to any concept in the given concept knowledge base, the sampling probability for a topic in position $i$ in document $d$ can be expressed as follows:

$$P(z_{(d,i)} = k | w, e_{-(d,i)}, z_{-(d,i)}; \alpha, \beta) \propto I^{(d)}(k) \cdot \frac{\beta_{k,e_{d,i}} + n^{(e_{d,i})}_{-,(d,i),k}}{\sum_{x=1}^{E} \beta_{k,x} + n^{(c)}_{-,(d,i),k}} \cdot \frac{\alpha_k + n^{(d)}_{-,(d,i),k}}{\sum_{t=1}^{K} \alpha_t + n^{(d)}_{-,(d,i),t}}$$  (7)

where $I^{(d)}(k)$ is the indicator function, and other notations have the same meaning as that in CLDA stated above.

Finally, the parameter can be estimated as follows:

$$\hat{\phi}_{e,k} = \frac{\beta_{k,e} + n^{(e)}_{-,(d,i),k}}{\sum_{x=1}^{E} \beta_{k,x} + n^{(c)}_{-,(d,i),k}}$$  (8)

$$\hat{\theta}^{(d)}_k = \frac{\alpha_k + n^{(d)}_{-,(d,i),k}}{\sum_{t=1}^{K} \alpha_t + n^{(d)}_{-,(d,i),t}}$$  (9)

the notations have the same meaning as that in CLDA stated above.
Table 1  The statistics of the datasets

| Datasets              | Conf   | AP    |
|-----------------------|--------|-------|
| Size of Documents     | 2317   | 106222|
| Size of Concepts      | 4740   | 4773  |
| Size of Vocabulary    | 18487  | 38419 |

4 Experiment

4.1 Experiment setting

We conducted the experiments on two real datasets. One of them, called Conf, contains 2,317 full papers of four conferences (CIKM, SIGIR, SIGKDD and WWW) of three years (2011 ∼ 2013). And the other dataset named AP is a public dataset, which contains more than 106K full Associated Press news articles published in 1989. Both of the raw datasets contain more than 2 million concepts according to Probase, which is much larger than the size of vocabulary. It leads to the imbalance between concepts size and vocabulary size. Moreover, lots of concepts in Probase Concept Graph are similar to each other associated with the same word. For example, the word microsoft associates with concept company, concept software company and concept technology company, which are semantically similar. In order to address this issue, we use the concept clustering results provided by Probase, to reduce the number of concepts. Totally, it contains 4,819 concept clusters. In the above example, all concepts about company can be represented by a concept cluster company.

The statistics of the two datasets are summarized in Table 1. And we conduct all the experiments on a server with an Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz and 125GB memory. In the rest, we will compare the proposed models with corresponding baselines in terms of case study and perplexity. For all models, we set the number of iterations in each collapsed Gibbs sampler as 1000, and set the same initial hyperparameters, where $\alpha$ and $\beta$ both equal to 0.01.

4.2 Experiments for CLDA

In the experiments, we removed the standard stop words for both datasets, and then we further removed words that occurred less than ten times. We use three state-of-the-art topic models as baselines, i.e. Latent Dirichlet Allocation (LDA) [1], Probase-LDA [30] and latent concept topic model (LCTM) [8]. We trained the four topic models using 80% of each dataset, and tested the models using the remaining 20% of each dataset. For each dataset, we ran each topic model with different number of topics from 10 to 100.

4.2.1 Case studies

Tables 2 and 3 show top ten words and concepts associated with 5 topics learned on Conf and AP respectively, when the topic number equals to 100. The topics learned by CLDA\(^1\) were matched to a topic in the baselines with smallest Kullback-Leibler divergence. It is

---

\(^1\)Source code can be found at https://github.com/anonymity01/CLDA
| Table 2 | Top ten words and concepts associated with five topics learned on Conf |
|---------|------------------------------------------------------------------|
| **Topic 1** | **Term** | **Words in LDA** | **Concepts in CLDA** | **Words in CLDA** | **Words in Probase-LDA** | **Words in LCTM** |
| **Topic 1** | user, server, web, request, content, site, page, browser, client, policy | heading, system -internal, antihypertensives, buzzword, predicate, user authentication information;security information, non-personal information, metric, organizer;graphic organizer, functionality | password, system, security, www, time, education, http, law, web, file | servers, provide, file, policy, website, system, sites, browser, security, http | bergeron, unplug, afe, netflow, icj, reising, meola, snowman, sanjoy, awa |
| **Topic 2** | user, data, children, model, inform, distribute, crowd, result, network, number | appearance, subheading, classifier, safety aspect, stamp, permanent insurance policy;life insurance policy, maois, analytical performance parameter;analytical parameter, engaging topic, price-related information;travel related information | train, svm, cluster, position, linear, unlabel, policy, feature:weight, precision, chemical | kernel, function, learning, svm, auc, dataset, share, space, compute, linear | unplug, rwi, afe, envi, gal, numeri, eif, paget, montana, cerri |
| **Topic 3** | xml, filter, data, set, user, function, query, bloom, slice, number | indies, campsite, engaging topic, governance model;advanced model, dbpedia, object-oriented feature, polyenes;condo, construct, company, financial statement | rdf, rdfa, dbpedia, ontology, http, rdfs, www, inherit, aztec, xml | related, extraction, semantic, ontology, knowledge, fact, type, instance, information, based | vsr, skyview, blanc, ofo, geni, luxembourg, brisson, saint-jean, lorenzo, schill |
| **Topic 4** | page, web, extract, site, data, text, node, content, form, inform | html element;page element, photojournalist, organizer;graphic organizer, supercomputer, relaxation method;technique, cracker, requirement, nicety, oral surgical procedure;gynecological procedure, skill | image, table, photograph, page, produce, binder, devo, snb: text, head, list | tags, role, semantic, cloud, navigation, image, proposed, annotation, audio, resources | taluk, jahr, klm, dame, afield, abhijit, opportunist, yul, cted, regis |
| **Topic 5** | ticket, set, video, number, user, data, system, model, result, time | applied engineering concept, statistical procedure, probabilistic model;statistical model, logistical problem, price-related information;travel related information, free blogging service;blog service, entertainment feature, permanent insurance policy;life insurance policy, boundary condition, folder | load, model, slim, regress, anova, correlation, transport, frequency, observed, sample | sample, estimate, random, distribution, method, compute, size, value, error, correlation | ofo, ctrl, stang, icj, meola, zarra, rawashdeh, bergeron, hambleton, dyk |
| Topic   | Words in LDA                        | Concepts in CLDA                                                                 | Words in CLDA                                                                 | Words in Probase- LDA                                                                 | Words in LCTM                                                                 |
|---------|------------------------------------|---------------------------------------------------------------------------------|--------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Topic 1 | store, sale, retail, year, shop, sell, busy, company, chain, sear | cloud property, business structure, disclosure, fixed income investment; debt security, real estate service, real estate transaction, tax-free investment; bond, broadband service, asset, company | bond, partnership, corporate, financial, sale, trust, merger, purchase, density, stock | year, percent, million, billion, sales, increase, tax, reported, cost, month      | astm, schertz, keron, moorea, stockport, mehren, arecibo, speea, kneed, hlava sa |
| Topic 2 | year, state, people, govern, million, report, week, day, time, nation | heading, food assistance program; welfare program, destination category; application category, button, angle, predicate, dental problem, organ, trial, high-risk group; patient actor, political idea, dissident, political group, slogan, online discussion, gif, contentious area, policy initiative, subobjects | medicaid, wic, infant, home, children, family, health, child, education, food       | women, children, abortion, state, court, child, law, parents, woman, case         | apeldoorn, sneed, deveau, feltham, nto, animaux, scale-back, mcshane, stockport, stahel |
| Topic 3 | year, state, dutch, people, report, van, million, nation, official, because | conflict, actor, compass direction; precinct, muslims, foreigner, vocabulary, militia, faction, munition, page factor | democracy, nation, wed, government, country, leader, democratic, socialist, political, equity | country, state, u.s, bush, president, united, foreign, government, national, world | cast-off, kickstart, deveau, dehnert, endicott, rosen, krizan, oarsmen, corbi, anti-serb |
| Topic 4 | year, state, sky, report, before, official, people, government, nation, because | mayor, governor, president, democratic, roller, royalist, ceo, populist, nsk, housw | party, elections, leaders, democratic, political, mayor, government, vote, national, city | party, elections, leaders, democratic, political, mayor, government, vote, national, city | kitsap, escobedo, piri, illini, segregationist, bur, rosen, wankel, goldfarb, corbi |
noted that the bold phrase in the third column of Tables 2 and 3 is the clustering concepts where pattern “A, ..., B; C” means “A, ..., B” is similar to “A, ..., C”, and the un-bold word is the atomic concepts.

From the two tables, we can see that in CLDA, a topic can be reflected by concepts, and a topic can be represented by a distribution of concepts. Both the concepts and the words can explain a specific topic. The top-n words learned by CLDA are obtained not only from the characteristics of the datasets, but also from the concept knowledge base, while neither of LDA’s nor LCTM’s are learned from the external database. Compared with Probase-LDA, the number of concepts in CLDA is automatically determined by the dataset without manual determination, while Probase-LDA has to set the concept number artificially. Top-n words of the topics in CLDA, LDA and Probase-LDA are generally understandable, while top-n words in LCTM are hard to understand.

Therefore, our proposed CLDA performs better than the baselines, and thus our conceptualization method for topic modeling sounds good.

4.2.2 Perplexity

In this experiment, we trained both CLDA and the baselines for ten times with different number of topics varying from 10 to 100 in turn. We computed the perplexity of the proposed CLDA and the baselines, which can quantificationally measure the quality of different models. A lower perplexity score indicates better generalization performance. The perplexity can be computed as follows:

\[
\text{Perplexity}(D_{\text{test}}) = \exp \left\{ -\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\}
\]  

Figure 3 shows the perplexity of the four models on the two datasets for different number of topics. As we can see in Fig. 3, the perplexity curves of CLDA, LDA and Probase-LDA decrease as the number of topics increases on both of the two datasets, while LCTM’s ups and downs, does not have too much law. In most cases, the perplexity value of CLDA is smaller than that of the baselines, which indicates that CLDA generally performs better than the baselines.

![Fig. 3 A comparison of the perplexity of CLDA and LDA with different number of topics on two datasets](image-url)
4.3 Experiments for CLLDA

In this experiment, we train CLLDA and LLDA over \textbf{Conf} dataset. The keywords of each paper will be used as labels of the corresponding paper, and the number of labels for the whole dataset is 4760.

4.3.1 Case study

Table 4 shows top ten words and concepts from five topics learned on \textbf{Conf}, where we can easily acquire the concepts under topics. From Table 4, we can learn that concepts is very important in understanding a document. For example, in topic \textit{social media}, concepts such as \textit{limitation, activity} and \textit{computer function; complex function}, are different aspects of the topic \textit{social media}, and the words like \textit{social, user, media} are related to these concepts. Therefore, from the results we know that the proposed model performances better than the baseline, which means our conceptualization method for topic modeling sounds good, and our assumption is more reasonable.

4.3.2 Perplexity

To compute perplexity for the two models with different number of topics, we segment documents in \textbf{Conf} into ten groups, and the first nine groups all contain 200 documents, while the last group contains the rest of the documents.

We train CLLDA and LLDA for ten times. The first time we use the first group, the second we use the first two groups, and so on. The comparison of the two models’ perplexity is

| Topic                  | Words in LLDA          | Concepts in CLLDA                  | Words in CLLDA          |
|------------------------|------------------------|------------------------------------|-------------------------|
| Social networks        | social, network, number, model, graph, set, inform, result, show, problem | socioeconomic variable, limitation, exercise program; class, operation, asset, activity, perturbation, product, basic contact information; contact information, anomaly | network, social, user, number, result, set, inform, node, topic, figure |
| Wikipedia              | wikipedia, article, inform, feature, test, set, word, page, entity, data | activity, poem, article, exercise program; class, limitation, answer, metric, construct, requirement, disadvantage | wikipedia, article, feature, inform, entity, evalu, page, category, figure, table |
| Query recommender systems | query, term, suggest, recommend, node, compute, model, set, result, graph | query, limitation, artifact, writing system; script, suggest, activity, reinforcers; essential, requirement, famous name, high wear area | query, suggest, recommend, model, user, generate, set, list, node, qfg |
| Social media           | social, user, media, inform, data, number, work, time, figure, twitter | limitation, exercise program; class, activity, asset, tax implication, operation, construct, company, perturbation, computer function; complex function | social, user, media, inform, topic, number, result, network, set, work |
| Data mining            | data, mine, set, inform, result, system, work, number, provide, perform | activity, limitation, requirement, skill tab, asset, metric, mega-projects, operation, construct, reinforcers; essential | data, mine, result, set, user, inform, learn, model, provide, number |
5 Conclusion and future work

In this paper, we propose a novel assumption of four-layer hierarchical Bayesian structure for topic modeling, which adds a latent concept layer between topic layer and word layer in the traditional assumption. To verify the effectiveness of the novel assumption, we apply the assumption in two representative topic models (LDA and LLDA). Extensive experiments have been conducted on two real datasets. The experimental results show that the proposed assumption performances better than the traditional assumption.

In the future, we will further verify the novel four-layer assumption in more topic models over more datasets. Also, we will explore the use of the new assumption in multimedia data, such as images and videos.

Acknowledgements This work was supported by 863 Program (2015AA015404), China National Science Foundation (61402036, 60973083, 61273363), Beijing Technology Project (Z151100001615029), Science and Technology Planning Project of Guangdong Province (2014A010103009, 2015A020217002), Guangzhou Science and Technology Planning Project(201604020179). Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201738).

References

1. Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
2. Blei D, Griffiths T, Jordan M, Tenenbaum J (2004) Hierarchical topic models and the nested chinese restaurant process. Adv Neural Inf Proces Syst 16:106
3. Cao D, Ji R, Lin D, Li S (2016) Visual sentiment topic model based microblog image sentiment analysis. Multimedia Tools and Applications 75(15):8955–8968
4. Cao Z, Li S, Liu Y, Li W, Ji H (2015) A novel neural topic model and its supervised extension. In: AAAI, pp 2210–2216
5. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
6. Fan Y, Zhou Q, Yue W, Zhu W (2017) A dynamic causal topic model for mining activities from complex videos. Multimedia Tools and Applications:1–16. https://link.springer.com/article/10.1007/s11042-017-4760-4
7. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 50–57
8. Hu W, Tsujii J (2016) A latent concept topic model for robust topic inference using word embeddings. In: The 54th annual meeting of the association for computational linguistics, p 380
9. Jayabharathy J, Kannmani S, Sivaranjani N (2014) Correlated concept based topic updation model for dynamic corpora. Int J Comput Appl 89(10):1–7
10. Joshi A, Bhattacharyya P, Carman M (2016) Political issue extraction model: a novel hierarchical topic model that uses tweets by political and non-political authors. In: Proceedings of NAACL-HLT, pp 82–90
11. Lim KW, Chen C, Buntine W (2016) Twitter-network topic model: A full bayesian treatment for social network and text modeling. arXiv:1609.06791
12. Magnusson M, Jonsson L, Villani M (2016) Dolda-a regularized supervised topic model for high-dimension classifikation. arXiv:1602.00260
13. Mao XL, Ming ZY, Chua TS, Li S, Yan H, Li X (2012) SSHLDA: a semi-supervised hierarchical topic model. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, pp 800–809
14. Mao XL, Xiao Y, Zhou Q, Wang J, Huang H (2015) Ehllda: a supervised hierarchical topic model. In: Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer, Cham, pp 215–226
15. Mimno D, Li W, McCallum A (2007) Mixtures of hierarchical topics with pachinko allocation pp 633–640
16. Murphy GL (2004) The big book of concepts. J Child Lang 31(1):247–253
17. Perotte A, Bartlett N, Elhadad N, Wood F (2011) Hierarchically supervised latent Dirichlet allocation. Neural Information Processing Systems (to appear)
18. Petinot Y, McKeown K, Thadani K (2011) A hierarchical model of web summaries. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol 2. Association for Computational Linguistics, pp 670–675
19. Qian S, Zhang T, Xu C, Shao J (2016) Multi-modal event topic model for social event analysis. IEEE Trans Multimedia 18(2):233–246
20. Ramage D, Manning C, Dumais S (2011) Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 457–465
21. Ramage D, Hall D, Nallapati R, Manning C (2009) Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol 1. Association for Computational Linguistics, pp 248–256
22. Rubin T, Chambers A, Smyth P, Steyvers M (2011) Statistical topic models for multi-label document classification. arXiv:1107.2462
23. Shin SJ, Moon IC (2017) Guided htm: Hierarchical topic model with Dirichlet forest priors. IEEE Trans Knowl Data Eng 29(2):330–343
24. Tang YK, Mao XL, Huang H (2016) Labeled phrase latent Dirichlet allocation. In: International conference on web information systems engineering. Springer International Publishing, pp 525–536
25. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. J Am Stat Assoc 101(476):1566–1581
26. Wang Z, Wang H, Wen JR, Xiao Y (2015) An inference approach to basic level of categorization. In: Proceedings of the 24th ACM international on conference on information and knowledge management. ACM, pp 653–662
27. Wu W, Li H, Wang H, Zhu KQ (2012) Probbase: a probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data. ACM, pp 481–492
28. Yan Y, Yang Y, Meng D, Liu G, Tong W, Hauptmann AG, Sebe N (2015) Event oriented dictionary learning for complex event detection. IEEE Trans Image Process 24(6):1867–1878
29. Yan Y, Ricci E, Subramanian R, Liu G, Lanz O, Sebe N (2016) A multi-task learning framework for head pose estimation under target motion. IEEE Trans Pattern Anal Mach Intell 38(6):1070–1083
30. Yao L, Zhang Y, Wei B, Qian H, Wang Y (2015) Incorporating probabilistic knowledge into topic models. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, Cham, pp 586–597
31. Yao L, Zhang Y, Wei B, Li L, Wu F, Zhang P, Bian Y (2016) Concept over time: the combination of probabilistic topic model with wikipedia knowledge. Expert Systems With Applications 60:27–38
32. Zhang C, Ek C, Gratel X, Pokorny F, Kjellstrom H (2013) Supervised hierarchical Dirichlet processes with variational inference. In: Proceedings of the IEEE international conference on computer vision workshops, pp 254–261

Yi-Kun Tang received the B.S. degree from Beijing Institute of Technology in 2016, and now she is studying towards the M.S. degree in computer science and technology from Beijing Institute of Technology. She once published a paper in the 17th International Conference on Web Information Systems Engineering (WISE2016).

Xian-Ling Mao received the PhD degree from Peking University in 2013. Now, he is an assistant professor of Computer Science at Beijing Institute of Technology. His current research interests include Topic Modeling, Learning to Hashing, and Question Answering. Various parts of his work have been published in top forums including AAAI, TKDE, IJCAI, EMNLP, COLING and CIKM.
Heyan Huang received her B.E. degree in 1983 from Wuhan University in computer science, a M.E. degree in 1986 from the National University of Defense Technology in computer science and technology and a Ph.D. degree from China Academy of Sciences, Institute of computer technology in 1989. And now she is a Professor, doctoral tutor, the president of Beijing Institute of Technology in School of computer, and the director of Research Center of High Volume Language Information Processing and Cloud Computing. Her current research interests mainly focus on machine translation. Her work have been published in top forums including TKDE, AAAI, IJCAI, ACL and COLING.

Xuewen Shi received the B.S. degree from Hunan University, and now he is Ph.D. candidate of computer science at Beijing Institute of Technology.
Guihua Wen born in 1968, Ph.D., was a professor and doctor supervisor. In 2005–2006, he did visiting research on machine learning and semantic web in School of Electronics and Computer, University of Southampton, UK. His main research interests are computational creativity, data mining, machine learning, and cognitive computing. He has published many papers in journals and conferences, including Pattern Recognition, Neurocomputing, Journal of Software, Journal of computer Research and Development, and IJCAI. He also directed the projects from China National Natural Science Foundation, State Key Laboratory of Brain and Cognitive Science, etc.