Enhancing Digital Book Clustering by LDAC Model

Lidong WANG†,††a†, Student Member and Yuan JIE†, Nonmember

SUMMARY In Digital Library (DL) applications, digital book clustering is an important and urgent research task. However, it is difficult to conduct effectively because of the great length of digital books. To do the correct clustering for digital books, a novel method based on probabilistic topic model is proposed. Firstly, we build a topic model named LDAC. The main goal of LDAC topic modeling is to effectively extract topics from digital books. Subsequently, Gibbs sampling is applied for parameter inference. Once the model parameters are learned, each book is assigned to the cluster which maximizes the posterior probability. Experimental results demonstrate that our approach based on LDAC is able to achieve significant improvement as compared to the related methods.

key words: digital book clustering, topic model, Gibbs sampling, LDAC model

1. Introduction

Nowadays, the development goal for digital library is to access digital collections anytime, anywhere. The availability of long document collections, such as digital books, presents new opportunities to mine deep knowledge based on correlation analysis and clustering. More and more research on document clustering aims to find a robust and practical algorithm for document semantic mining, which attempts to discover new, previously unknown knowledge by applying techniques from Natural Language Processing (NLP) and data mining. As one of the common data mining techniques and an unsupervised learning method, clustering methods attempt to segregate the documents into groups where each group represents some topic that is different with those topics represents by the other groups [1].

Traditional clustering techniques reply on three concepts: representation model [2]–[4], similarity measure [5] and clustering model [6]. However, these traditional models have a number of limitations. First, these methods virtually neglect the most important criteria of clustering, which is topic discrepancy. Clustering the documents just based on terms co-occurrence would not reflect whether two documents share common topics or not. Second, most of the improved algorithms are based on the BOW representation or a variation of it. The BOW model is limited if applied on digital books because of the problem of data sparseness and high dimension, and also neglects the implicit relationships between terms and documents, that is topic. Therefore, traditional document clustering methods may fail to achieve satisfactory results when they are directly applied to digital book clustering task.

Furthermore, high dimensionality is a common problem for digital books (long documents). Lots of recent work concentrates on projecting data into lower dimensional spaces to remove irrelevant dimensions, including feature transformation [7] and feature selection [8], then conduct clustering. Dimension reduction is successful in uncovering latent structure and improving clustering performance. However, since they preserve the relative distances between objects, they are less effective when there are large numbers of irrelevant attributes that hide the clusters in noise data. Also, the process of dimension reduction makes the clustering algorithm more complicated. A different clustering method for high dimensional datasets is subspace clustering [9]. This method selects a small number of original dimensions in some unsupervised way so that clusters become more obvious in this subspace. Focusing on the original dimensions has the advantage of easy implementation, but the rigidity of original feature does not have enough flexibility to handle clusters which extends along a mixture of directions [10].

To achieve a more accurate document clustering for digital books (long documents), we present a unique clustering approach through the way of topic analysis. Our approach is based on the intuition that digital books clustered in a group share one or more common topics. As for digital books, we utilize the multiple-granularity information of books, including Contents and Texts. Texts mentioned here represent the normal main body of a book that does not contain Contents information. In this paper, we extend the Latent Dirichlet Allocation (LDA) [11] model as LDAC, which effectively extracts the topics by integrated topic modeling for Contents and Texts.

The rest of the paper is organized as follows. In Sect. 3, we present the graph model of LDAC, which explicitly models Contents and Texts for each book, significantly outperforming other related methods on long document tasks. In Sect. 4, the experiments performed on digital books corpus and the results are presented. Our final conclusion and suggestions for future work are discussed in Sect. 5.
2. Problem Statement

In this paper, a common word “document” is used to represent digital book. A document clustering algorithm partitions a set of documents into groups of similar documents. We call the groups of similar documents clusters. In this paper, we look at a series of clustering algorithm, each of which has the following input and output:

**Input** A target number of clusters $c$, and a set of documents/books numbered $\{1 \ldots D\}$. However, $c$ is not required for LDAC based algorithm.

**Output** An assignment of documents/books to clusters. The assignments is represented as a mapping from each document/book to a particular cluster $c$.

This description is similar to a standard document clustering task. Furthermore, many clustering algorithms make soft rather than hard assignments. With hard assignments, every document is a member of one and only one cluster. Soft assignments allow for degrees of membership and membership in multiple assignments. In our paper, we conduct hard assignment by selecting the single most likely cluster for each document. Meanwhile, our output is a flat set of clusters. We focus on flat (nonhierarchical) clustering algorithms rather than hierarchical clustering algorithms, since the time complexity of former tend to be $O(n^2)$ or $O(n^3)$ [15].

3. LDAC Based Clustering

Using topic models for document representation has recently been an area of considerable interest in document-related tasks. One of the well-known topic models is LDA. Recently, document clustering based on topic models have been done in several ways. Kim [12] proposed a new PLSA-based algorithm to solve the PLSA limitation of too restrictive in cases where there are more topics than document clusters in a collection. Cassar [14] proposed using LDA to learn latent factor from the corpus of service descriptions and group them in clusters according to the result of topic modeling. However, if we directly apply LDA model for books clustering, the books that contain Contents and Texts should be considered as a whole ordinary document for topic modeling, which has a limitation that the importance of Contents would be neglected, in that Contents intend to contain key topic words. Moreover, it is not proper to conduct clustering on Contents directly, since it will cause the distortion of representing a whole book, and would get bad performance. Although Texts contain more information than Contents, the Contents may represented by key words that do not appear in Texts. Therefore, Texts and the Contents can be considered as independent sets of observations for topic modeling. We assume that Contents and Texts share common topic.

3.1 Introduction of LDA Model

LDA is a hierarchical Bayesian network that constructs relationship between terms and documents through latent topics. It assumes that one document contains several latent topics, then the relationship is constructed by additional latent variables $z_w$, which represents the certain topic in a document. Dirichlet priors $\alpha$ and $\beta$ are set over the document and topic distributions, respectively. We denote a document corpus as $D = \{d_1, d_2, \ldots , d_M\}$ where $M$ is the number of documents. Each document contains a sequence of words $w = (w_1, w_2 \ldots w_N)$ where $N$ is the number of terms in certain document. These words are derived from a dictionary in size of $|V|$. The Bayesian network of LDA is depicted in Fig.1.

As shown in Fig.1, the generative process of a document under LDA is given as follows: First, draw a multinomial distribution $\theta$ from a Dirichlet distribution with parameter $\alpha$. $\theta \in \mathbb{R}^{M \times K}$, represents the document-topic distribution of the corpus. Second, for all the words in a document, draw a topic $z_w$ from the multinomial distribution $\theta$. Third, draw distribution $\delta$ from a Dirichlet distribution with parameter $\beta$. $\delta \in \mathbb{R}^{K \times V}$, represents the topic-term distribution. Finally, draw word $w_i$ on the basis of $p(w_i | z_w; \beta)$.

3.2 LDAC Model

We denote a digital book corpus as $\mathbb{R} = \{d_1, d_2 \ldots d_D\}$, $D$ is the number of digital books. The best way to describe LDAC is to outline the process of generating a book. The Bayesian network of LDAC model is described as follows: As shown in Fig.2, LDAC generates a collection of books by the process below:

(i) For each book $d_i$, $i \in [1 \ldots D]$ in the collection, draw $\theta_i$ of size $K$ from a Dirichlet distribution with parameter $\alpha$. Each $\theta_i$ represents the probability of certain topic in book $d_i$.

(ii) For Texts in each book, draw $\delta_k$ of size $|T|$ from a Dirichlet distribution with parameter $\beta$. $\delta_k$ represents the probability of seeing all words given topic $k$, $k \in [1 \ldots K]$. 

![Fig. 1 Bayesian network of LDA model. Shaded circles represent observed variables, unshaded circles represent hidden variables, diamonds represent model parameters, and plates represent replications. $K$ represents the topic number.](image-url)
(iii) For Contents in each book, draw $\phi_i$ of size $|C|$ from a Dirichlet distribution with parameter $\beta$. $\phi_i$ represents the probability of seeing all words given topic $l$, $l \in [1 \ldots L]$.

(iv) For each term index $t \in [1 \ldots N]$ from Texts in book $d_i$: (a) Draw a topic $z_t$ from $\theta_i$, $z_t \in [1 \ldots K]$. (b) Draw a term $w_t \in T$ from $\delta_z$.

(v) For each term index $c \in [1 \ldots M]$ from Contents in book $d_i$: (a) Draw a topic $z_c$ from $\theta_i$, $z_c \in [1 \ldots L]$; (b) Draw a term $w_c \in C$ from $\varphi_c$.

The relationship between $K$ and $L$ should be $K \gg L$ because of the discrepancy in the amount of information. It is assumed that Contents and Texts keep common topic $z$, $z \in [1 \ldots L]$, since the Contents is the summarization of Texts. However, there really exists rare condition that Contents and Texts contain different topics, which we will discuss in future work.

3.3 Learning LDAC Parameters

Variational inference [17] and Gibbs sampling [18] are two general techniques that have been used in parameter inference. Both the variational and Gibbs sampling approaches have their advantages: the variational approach is faster computationally, and the Gibbs sampling approach is in principal more accurate since it asymptotically approaches the correct distribution. Our paper employs Gibbs sampling for learning the parameters $\theta, \delta, \varphi$. The algorithm is given as follows:

1. Iterate repeatedly through the books in random order.
   For all terms in each book, randomly sample topic $z_t, z_c$ from $[1 \ldots K]$ and $[1 \ldots L]$, respectively. This step is considered as the initialization of assigning certain topic for each term.
2. For all terms in Texts of each book, use the assignment of $z_t$ for term $w_t$ based on $p(z_t \mid w_t)$. Similarly, assign the topic $z_c$ for term $w_c$ based on $p(z_c \mid w_c)$. The approximately computation of $p(z_t \mid w_t)$ is described in Eq. (1). We use $p(z_t = j \mid z_{-t}, w_t, w_{-t})$ to simulate $p(z_t = j \mid w_t)$, which estimates the probability of assigning the current term to each topic ($p(z_t = j)$), conditioned on the topic assignment to all other terms ($z_{-t}$).

$$p(z_t = j \mid z_{-t}, w_t, w_{-t})$$

$$\propto p(w_t \mid z_t = j, z_{-t}, w_{-t})p(z_t = j \mid z_{-t})$$

$$\propto \int p(w_t \mid z_t = j, \delta_j)p(\delta_j \mid z_{-t}, w_{-t})d\delta_j$$

$$\propto \int p(z_t = j \mid \theta_i)p(\theta_i \mid z_{-t})d\theta_i$$

$$\propto \frac{n_{w_{-t}}^{z_t=j} + \beta}{n_{z_t=j} + \beta} \frac{n_{z_{-t} \neq j}^{w_{-t}} + \alpha}{n_{z_{-t} \neq j} + \alpha} + \frac{N\beta}{n_{z_{-t} \neq j} + K\alpha}$$

(1)

Where $z_t = j$ means assigning current term $w_t$ to topic $j$, $z_{-t}$ denotes the topic assignments for all word tokens except word $w_t$. The meaning of $n_{z_t=j}^{w_{-t}}$, $n_{z_{-t} \neq j}^{w_{-t}}$, $n_{z_{-t} \neq j}^{w_{-t}}$, and $n_{z_{-t} \neq j}^{w_{-t}}$ is the same as corresponding components in Table 1, but not including the current assignment instance $t$ (represented by the token $-t$). For $p(z_c \mid w_c)$, it is estimated in analogous way (see Eq. (2)).

$$p(z_c = j \mid z_{-c}, w_c, w_{-c})$$

$$\propto \frac{n_{w_{-c}}^{z_c=j} + \beta}{n_{z_c=j} + \beta} \frac{n_{z_{-c} \neq j}^{w_{-c}} + \alpha}{n_{z_{-c} \neq j} + \alpha} + \frac{M\beta}{n_{z_{-c} \neq j} + M\alpha}$$

(2)

(3) Repeats step 2 until a maximum iteration number is reached, then read out parameter $\theta, \delta, \varphi$. The estimations for parameters $\theta, \delta, \varphi$ are listed below, and the meanings of each component in Eq. (3)–Eq. (5) are showed in Table 1.
4. Experiments

In our experiments, we look at two broad families of clustering algorithms. The first family is based on vector space model (VSM), including K-means algorithm, and Affinity Propagation (AP) algorithm[19]. K-means has the advantage of being simple to understand, efficient, and standard. AP algorithm is a new algorithm and is overestimated because of its simplicity, general applicability and good performance. Moreover, ISOMAP [20] is employed for dimension reduction for the first family to alleviate the problem of data sparseness and high dimension. The second family is based on probabilistic models, including PLSA, LDA, and LDA-derived model named LDAC. PLSA model introduced by [13] is a model which characterizes each word in a document as a sample from a mixture model, where mixture components are conditionally-independent multinomial distributions. The detail process for LDA based clustering can be found in [14]. PLSA based method is conducted similarly in [12]. The common characteristic in these two methods and traditional methods is that Contents and Texts in each book are considered as a whole ordinary document.

In this section, we ask two questions about our method:

(1) Does topic model based clustering algorithms perform better than traditional clustering methods? (Discussed in Sect. 4.3)

(2) Can we do better than PLSA based method and LDA based clustering by creating a model that explicitly accounts for Texts and Contents as separate observations in each book? (Discussed in Sect. 4.3)

4.1 Data Sets

To effectively evaluate our method, we construct the Books corpus: the corpus contains 500 books, which are downloaded from Internet and our digital library\(^1\). The corpus is manually divided into 10 categories that are Management, Sports, Economy, Military, Tourism, Diet, Philosophy, Medicine, Computer and Environment. The relative document number of each category is listed as follows: 48, 39, 61, 62, 50, 49, 42, 44, 48, 57. The page number of each book falls in the region of 200–300.

The corpus should be pre-processed before experiment. All of the data are under the manipulation of word segmentation and part-of-speech tagging, then stopwords are removed, and the nouns as well as gerunds are retained, in that topic words are mostly determined by the nouns and gerunds in Chinese expression. Furthermore, we assume that the words appear less than five times in each book are not important terms in most of cases, so the words appearing less than five times are filtered out to simplify computation. For evaluation, F-measure evaluation measure [16] and entropy [1] are applied to comparing the generated clusters with the original classes. The higher the overall F-measure, the better the clustering, due to the higher accuracy of the clusters mapping to the original classes. Entropy denotes the degree of homogeneity, the higher the homogeneity of a cluster, the lower the entropy is, and vice versa.

4.2 Parameters Selection

Several parameters are essential to be determined for LDAC

\(^1\)http://www.cadal.zju.edu.cn
model, which are $\alpha$, $\beta$, iteration number and topic number $K$, $L$. In this section, we randomly select 100 documents from the corpus for experiments of parameter selection.

(1) Topic number

The method for the selection of topic number can be directly found in our published paper [18], which we have directly confirmed that the method can find the optimum topic number. According to the experimental result, the optimum number of $K$ and $L$ should be set to 110 and 20, respectively.

(2) Parameters $\alpha$, $\beta$

Parameters $\alpha$, $\beta$ are set as follows: $\alpha = 50/K$, $\beta = 0.01$, which are common settings in the literature. Our experience shows that clustering results are not very sensitive to the values of these parameters.

(3) Iteration number

Iteration number in LDAC estimation plays an important role in its complexity. Generally, after a certain number of iterations means the invariant distribution of the Markov chain is equivalent to the true distribution. So it would be ideal if we take samples right after the convergence. However, in clustering task, it is almost impossible to run a very large number of iterations due to the size of data set. In order to get a good iteration number that is effective for clustering task, we randomly select 100 documents from Books corpus for training and maximizing the F-measure as the optimization criterion since it is our final evaluation metric. Different iteration numbers are tried in our experiment. The results are presented in Fig. 3. After 60 iterations, performance is quite stable. Therefore, in consideration of a good trade-off between accuracy and running time, the iteration number is set to 60 in our final experiment.

4.3 Overall Performance

We conducted several clustering methods based on: K-means, AP, LDA, PLSA, LDAC, K-means+ISOMAP, AP+ISOMAP and K-means+Content. Among which, “*+ISOMAP” means that algorithms run under the dimensionality reduction by ISOMAP before directly running of original methods. Most of research concentrate on the combination of traditional methods with dimensionality reduction methods for long documents, we hope to present whether it can achieve ideal improvement for digital books clustering. “K-means+Contents” denotes that we apply K-means algorithm on Contents of each digital book, nor the full text.

Firstly, we present the complexity analysis for our method. Complexity is often a big concern for topic model. Representative algorithms from two families that are LDA, LDAC and K-means are selected for comparison of computation time. Because of little time fluctuation in each running, we set the average time of 10 runs for each method as the final record.

The above comparison shows that LDA based method is the most time-consuming. However, the difference in running time among these methods is trivial. Next we take detail analysis for these three algorithms. The time complexity of the Gibbs sampling in the LDA and LDAC is $O(I_K * K * M * \overline{d_l})$, where $I_K$ is the iteration number, $K$ is the topic number, $M$ is the number of documents for experiment, and $\overline{d_l}$ is the average number of words in one document. Comparing the running time between LDA and LDAC, the former is larger, in that the iteration number for LDA model is set to 100, which is the common settings in other literatures, and is bigger than the iteration number 60 in LDAC model. As for K-means, the time complexity is $O(I_K * M * \psi * \overline{\psi_w})$, where $I_K$ is the number of iterations, $\psi$ is the number of clusters, $\overline{\psi_w}$ is the average number of unique words in each cluster. The detail comparison between LDAC and K-means is given as follows: As to $I_K$ vs $I_K$, they will probably have a similar value. $I_K$ changes a lot for different datasets. In our experiment, given the condition of convergence, the iteration number for K-means falls in the region of [50, 65]. In Sect. 4.2, we have shown that 60 is reasonable for LDAC algorithm. As to $K$ vs $\psi$, $K > \psi$, since $\psi$ is always set to the real number of cluster. As to $\overline{d_l}$ vs $\overline{\psi_w}$, $\overline{\psi_w}$ is mostly determined by $\psi$, $\overline{\psi_w}$ is often much larger than $\overline{d_l}$ in our experiments, since if the cluster number is smaller, one cluster always contains much more documents and words.

Therefore, based on the discussion above, no evidence shows that the one is obviously more efficient, which also conforms to the experimental results shown in Table 2.

Secondly, we discuss on the situation that whether dimension reduction can improve the performance for long document clustering. The result from “K-means+Contents” gets the worst performance, which can demonstrate that the clustering conducted on small abstract such as Contents is not proper for digital book clustering. Each algorithm is implemented through dimensionality reduction by Isomap algorithm, as shown in Table 3. All methods implemented via dimension reduction appear to have better performance. However, the improvement is not obvious. For

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Running time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>2.85</td>
</tr>
<tr>
<td>LDA</td>
<td>3.06</td>
</tr>
<tr>
<td>LDAC</td>
<td>2.91</td>
</tr>
</tbody>
</table>

Fig. 3 Clustering results with different number of iterations. $K = 110$, $L = 20$, $\alpha = 50/K$, $\beta = 0.01$. |
Table 3 F-measure for different traditional methods.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means + Contents</td>
<td>0.105</td>
</tr>
<tr>
<td>K-means</td>
<td>0.187</td>
</tr>
<tr>
<td>K-means + ISOMAP</td>
<td>0.202</td>
</tr>
<tr>
<td>AP</td>
<td>0.213</td>
</tr>
<tr>
<td>AP + ISOMAP</td>
<td>0.235</td>
</tr>
</tbody>
</table>

Table 4 Comparison of four methods for dataset.

<table>
<thead>
<tr>
<th></th>
<th>BASE1</th>
<th>PLSA</th>
<th>LDA based</th>
<th>LDAC based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>0.235</td>
<td>0.562</td>
<td>0.643</td>
<td>0.753</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.583</td>
<td>0.492</td>
<td>0.481</td>
<td>0.316</td>
</tr>
</tbody>
</table>

Kmeans + ISOMAP, there only exists an increment of 0.015 on F-measure compared with Kmeans. Another method (AP + ISOMAP) has an analogous effect.

To compare topic model based clustering algorithm with traditional clustering methods, AP + ISOMAP is selected as baseline and denote it as BASE1, which has been demonstrated its superiority compared with other conventional methods in Table 3. As shown in Table 4, three methods based on topic model all perform better than BASE1, and the increase of F-measure between LDAC based clustering and BASE1(0.511) outclasses the promotion rate between LDA and LDAC(0.11). This result shows that the general clustering method based on topic modeling significantly outperforms other conventional methods. The method based on LDAC performs best for dataset. Its advantage reflects in the following aspect: (1) Compared with BASE1, there is significant increase of 0.518 in F-measure, and -0.267 drop in Entropy; (2) Based on the results of two quality measures, it can be demonstrated that taking the Contents and Texts as separate observations for topic modeling can effectively improve the performance of clustering over a standard LDA model and PLSA model, both of which just combine the Contents and Texts as a whole ordinary document.

The stability of the algorithm is also an essential factor. In Fig. 4, we depict the clustering performance with different category numbers under the run of LDA and LDAC, respectively. Compared to LDA, the LDAC method consistently performs better with different category numbers. Clearly, our goal about these comparisons is to demonstrate that our proposed method based on LDAC is suitable for books clustering, and the performance will not be influenced by the size of corpus.

Based on the above discussion on different situations, we believe that the clustering based on topic analysis is a more effective way than traditional methods for digital book clustering. It is demonstrated that the effect of the relationship between topic analysis and clustering is obvious, which can be explained by the fact that “the documents in one group share one or more common topics”. Moreover, LDAC based algorithm proposed in our paper outperforms other related algorithms in book clustering task.

5. Conclusion and Discussion

This work addresses the problem of effectively clustering the documents (books) through the way of topic analysis. We assume that Contents and Texts share common topic, then exploit Contents and Texts as separate observations for integrated topic modeling, and propose a new probabilistic topic model named LDAC. The main goal of LDAC model is to improve the books clustering performance in general. Empirical evaluations demonstrate that the clustering method based on topic model can obviously achieve better performance than traditional methods.

In future work, we will develop a query mechanism based on latent topics rather than matching key words in a query to the texts of each book. This kind of books clustering makes it easier to recommend books with similar semantic information since documents are grouped in the same cluster according to topics. Furthermore, the model in our paper assumes that the topics from Contents can also be derived from Texts, there exists another possibility that different two topic are drawn from two parts, which will be conducted in our future work.

References


Lidong Wang was born in December, 4, 1982. She received the M.S. degree in computer Science from Ningbo University. She is currently pursuing the Ph.D. degree in College of Computer Science and Technology, Zhejiang University. In 2007, she joined Hangzhou Normal University, Hangzhou, Zhejiang Province, China. Her current research interests include image processing, machine learning and text mining.

Yuan Jie received the M.S. degree in Zhejiang Normal University, Zhejiang province, China. He is currently a Ph.D. student in College of Computer Science at Zhejiang University (ZJU), Hangzhou, Zhejiang Province, China. He worked as an assistant in Hangzhou Normal University. His current research interests are image processing, pattern recognition and information retrieval.