Latent Dirichlet Allocation Approach for Analyzing Text Documents

D

Parvathi Chundi
University of Nebraska at Omaha, USA

Sussanah Go
University of Nebraska, USA

INTRODUCTION

The amount of digital information accumulated in the form of document data in various fields is ever increasing. Most of the digital information available today is typically unstructured and unlabeled. For the data to be of any use, it must be organized and managed. Due to the enormity of the data, tools must be employed to automatically organize the information hidden in document sets. One popular method for organizing a document collection is to label documents with topics. For example, given a set of news articles, one may want to identify all documents, if any, which are about foreign policy, or domestic issues, etc. Topic labels help users identify relevant documents quickly. Automatically labeling documents with topics is a classic problem in information retrieval. Document classification approaches have traditionally been used for topic labeling. These approaches learn rules or knowledge from a small subset of documents which are labeled with topics. The knowledge is then employed to assign labels to the rest (or new) documents in the collection.

Having access to documents labeled with topics is becoming increasingly difficult in many domains including blogs and social media where the content and topics may change frequently. For these domains, it is crucial to discover suitable topics for a given document set. Topic modeling algorithms are statistical methods used to discover the topics or themes appearing in a document collection. The topic modeling algorithms do not require previously labeled documents, but analyze the words appearing in documents, to discover topics from a document collection. Thus, topic modeling algorithms are well suited for organizing large document collections that would be impossible to label otherwise.

Latent Dirichlet Allocation is a simple topic model proposed by Blei et al (Blei, Ng, & Jordan, 2003) which operates under the assumption that a document may include different topics where a topic is a probability distribution over words. Documents in a collection are treated as observed, whereas the topic structure, how topics are distributed over documents, and how words are distributed over topics are thought to l*atent* or *hidden*. This hidden topic structure is revealed by the LDA approach.

The rest of this article is organized as follows. The section on *Background* gives an overview of the topic model problem and the LDA approach. This section is followed by a *Main Focus* section where details of the LDA approach are discussed. The *Sample Applications* section discusses a few applications of the LDA approach. Then, the *Future Trends* section presents some of the improvements and extensions to the LDA approach and the *Conclusion* section concludes the article. Several important terms and their definitions are also included at the end of the article.

BACKGROUND

The problem that latent Dirichlet allocation (LDA) seeks to solve is as follows: Given a corpus, find short descriptions of the documents that facilitate efficient processing of the corpus while keeping intact the statistical relationships between the documents and words in the corpus that may be needed for other types of processing. LDA solves the problem by assuming that documents may be represented as a mixture of latent (unknown) topics, and then uses statistical inference to create groups of co-occurring words, which are used

DOI: 10.4018/978-1-4666-5888-2.ch175

to form topics. LDA is a generative model. It specifies how to generate each document in a document collection from a specific set of topics. The *words* in a document can be generated in two steps:

Step 1: Randomly choose a distribution over topics.
Step 2: Each word in the document is generated as follows.
a. Select a random topic from the distribution of topics chosen in Step 1.
b. Select a word randomly from the word distribution corresponding to that topic.

The order of words is ignored by LDA, making it a bag-of-words model. LDA resolves some of the issues present in previous topic models; the Unigrams Model did not allow for multiple topics in one document, and the probabilistic latent semantic indexing (pLSI) model was prone to overfitting.

MAIN FOCUS

The main goal of LDA is to infer the topic structure hidden in a document collection by analyzing the distribution of words over documents. A corpus is a collection of M documents, and a document is a sequence of words. A word is the basic unit of discrete data to be analyzed. If there are N words in a corpus, then each word is an item from vocabulary V with indices $1 \ldots N$. Each word in V is represented as a vector over dimensions N, where the component corresponding to the word is one and all other components are zero. That is, the v^{th} word in the vocabulary is represented by a N-vector p where $p^v = 1$ and $p^u = 0$ for all $u \neq v$, where indices represent the vector component. A *topic* is a probability distribution over a collection of words V. A *topic model* is a statistical model used to represent the semantic structure underlying the corpus.

LDA can be described as follows. Let β_i ($1 \leq i \leq K$) is a topic which represents a distribution over the vocabulary V where K is the total number of topics for the given corpus. Let θ_d denote a distribution over topics for a document d where $\theta_{d,i}$ specifies the proportion of topic β_i for the document d. Let $z_{d,n}$ denote the topic assignment of n^{th} vocabulary word in document d. Finally, n^{th} word in a document d is denoted by $w_{d,n}$. Then, the joint probability of observing a corpus is given by Equation (1).

$$p(\,,\,,z,w) = \prod_{i=1}^{K} p(\,_i) \prod_{d=1}^{M} p(\,_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\,_d)\, p(w_{\mathrm{d,n}}|_{1\ldots K}\ ,\ z_{d,n}) \right) \quad (1)$$

Note here that the topic assignment to a word observed, $z_{d,n}$, depends on the distribution of topics assigned to d, $\theta_{d.}$. The observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ as well as *all* topics. Figure 1 shows a graphical representation of the above computation.

In the graphical model, each node is a random variable and is labeled according to its role in the generative process. The unshaded nodes, topics, topic assignments, and proportions, denote the hidden variables whereas the shaded node represents the observed variables – words in a document. Each rectangle stands for "plate" notation, which denotes replication. The inner N plate denotes the repeated sampling of topics and words until N words are generated for a document d. The outer plate denotes the sampling of a distribution over topics for each document d in the corpus containing M documents. The plate surrounding β_i illustrates the

Figure 1. Graphical representation