The choice of prior to latent dirichlet allocation for fitting topic models

Yi-Hsiung Lin[1]* (林億雄) and Huey-Hong Hsieh[2]
[1]Department of Health Business Administration, Taiwan Shoufu University
[2]Department of Leisure Management, Taiwan Shoufu University

Abstract

As our collective knowledge continues to be digitized and stored in the form of blogs, news, web documents in HTML, XHTML and other social network textual data. We need new tools to help us organize and understand these huge data sets. Topic modeling provides a flexible latent variable framework for modeling high-dimension spare count data and allows the probabilistic modeling of term frequency occurrences in documents in a given corpus. From a statistical machine learning perspective, topic models such as Latent Dirichlet Allocation (LDA) have been recognized as useful tools for analyzing large, unstructured collections of documents. LDA is a three level hierarchical Bayesian model to group data like documents, images or biological sequences. It is convenient to use a symmetric Dirichlet prior on the Document-topics and Topic-terms distribution. We discuss the problems of several classes of structured priors for topic models. Using an asymmetric Dirichlet prior over the Document-topics distribution and a symmetric Dirichlet prior over the Topic-terms distribution has a significant modeling benefit. Here, we will present an overview of topic models first. The basic idea for applying topic models will be proposed, using examples for illustration. Research results of choosing priors on topic models and future research issues will be discussed and proposed.

Keywords: hierarchical Bayesian model, latent Dirichlet allocation, prior choosing, statistical machine learning, topic models