LIMTopic: A Framework of Incorporating Link based Importance into Topic Modeling

Abstract:

Topic modeling has become a widely used tool for document management. However, there are few topic models distinguishing the importance of documents on different topics. In this paper, we propose a framework LIMTopic to incorporate link based importance into topic modeling. To instantiate the framework, RankTopic and HITSTopic are proposed by incorporating topical pagerank and topical HITS into topic modeling respectively. Specifically, ranking methods are first used to compute the topical importance of documents. Then, a generalized relation is built between link importance and topic modeling. We empirically show that LIMTopic converges after a small number of iterations in most experimental settings. The necessity of incorporating link importance into topic modeling is justified based on KL-Divergences between topic distributions converted from topical link importance and those computed by basic topic models. To investigate the document network summarization performance of topic models, we propose a novel measure called log-likelihood of ranking-integrated document-word matrix. Extensive experimental results show that LIMTopic performs better than baseline models in generalization performance, document clustering and classification, topic interpretability and document network summarization performance. Moreover, RankTopic has comparable performance with relational topic model (RTM) and HITSTopic performs much better than baseline models in document clustering and classification.