

TPMTM: Topic Modeling over Papers' Abstract

Than Than Wai
thanwai85@gmail.com

University of Computer Studies, Mandalay, UCSM

Sint Sint Aung
ssaung@gmail.com

University of Computer Studies, Mandalay, UCSM

Probabilities topic models are active research area in text mining, machine learning, information retrieval, etc. Most of the current statistical topic modeling methods, such as Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). They are used to build models from unstructured text and produce a term-based representation to describe a topic by choosing single words from multinomial word distribution. There are two main weaknesses. First, popular or common words are different topics, often causing ambiguity for understanding the topics; Second, lack of consistent semantics for single words to be represented correctly. To address these problems, this paper proposes a model (A Two-Phase Method for Constructing Topic Model, TPMTM) that combines statistical modeling (LDA) with frequent pattern mining and produces better presentations of rich topics and semantics. Empirical evaluation shows that the results of the proposed model are better than LDA.