Ontology-based Topic Labeling

Abstract:

Probabilistic topic models such as Latent Dirichlet Allocation have been shown to be powerful techniques to analyze the content of documents and extract the underlying topics represented in the collection. Topic models typically assume that documents are mixtures of topics, while topics are probability distributions over the vocabulary. These models have been extensively used in a variety of text processing tasks, such as word sense disambiguation, relation extraction, text classification and information retrieval. However, due to the fact that topic models are entirely unsupervised, purely statistical and data driven, they may produce topics that are not meaningful and understandable to humans. In this talk, we address some of the research problems in ontology based topic modeling. We will focus more closely on ontology-based topic labeling. Topic labeling aims to assign meaningful labels for discovered topics. I will discuss how we can significantly improve the quality of topic labeling by considering ontology concepts rather than words alone. I will explain an ontology-based topic model that integrates ontological concepts with topic models in a single framework, where each topic is a multinomial distribution over the concepts while each concept is represented as a multinomial distribution over words. Furthermore, I will explain a topic labeling method based on the ontological meaning of the concepts included in the discovered topics, and present experimental results that demonstrate the effectiveness of the method for generating meaningful labels for the discovered topics.