Nonparametric Bayesian Topic Modelling with Auxiliary Data

Kar Wai Lim

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

August 2016
Abstract

The intent of this dissertation in computer science is to study topic models for text analytics. The first objective of this dissertation is to incorporate auxiliary information present in text corpora to improve topic modelling for natural language processing (NLP) applications. The second objective of this dissertation is to extend existing topic models to employ state-of-the-art nonparametric Bayesian techniques for better modelling of text data. In particular, this dissertation focuses on:

- incorporating hashtags, mentions, emoticons, and target-opinion dependency present in tweets, together with an external sentiment lexicon, to perform opinion mining or sentiment analysis on products and services;

- leveraging abstracts, titles, authors, keywords, categorical labels, and the citation network to perform bibliographic analysis on research publications, using a supervised or semi-supervised topic model; and

- employing the hierarchical Pitman-Yor process (HPYP) and the Gaussian process (GP) to jointly model text, hashtags, authors, and the follower network in tweets for corpora exploration and summarisation.

In addition, we provide a framework for implementing arbitrary HPYP topic models to ease the development of our proposed topic models, made possible by modularising the Pitman-Yor processes. Through extensive experiments and qualitative assessment, we find that topic models fit better to the data as we utilise more auxiliary information and by employing the Bayesian nonparametric method.


BIBLIOGRAPHY


