A Bayesian Semi-supervised Approach to Key Phrase Extraction with Only Positive and Unlabeled Data

Sherry Wang

Southern Methodist University
E-mail: swang@smu.edu

Abstract: A set of keyphrases is often used as a brief summary of a document as it provides a good coverage of the content. Since manual assignment is tedious and time-consuming, learning methods based on the rankings of importance scores have been developed for keyphrase extraction especially when there exists for a large database or collection of documents. Supervised learning requires a labeled training set that needs to be obtained by human effort. Unsupervised learning does not require any labeled set. However, it is often the case that a small portion of keyphrases can be easily obtained from the title or abstract. We propose a model-based semi-supervised Bayesian learning method for keyphrase extraction, which utilizes the information from known positive labels to improve the scoring process. Unlike previous methods that are purely algorithm-driven, our approach is probabilistic and allows for assessment of estimation uncertainty besides its improved performance in controlling the false discovery rate.