

**An Efficient Method for High Quality and Cohesive Topical Phrase Mining**

**Abstract:**

A phrase is a natural, meaningful, and essential semantic unit. In topic modeling, visualizing phrases for individual topics is an effective way to explore and understand unstructured text corpora. However, from phrase quality and topical cohesion perspectives, the outcomes of existing approaches remain to be improved. Usually, the process of topical phrase mining is twofold: phrase mining and topic modeling. For phrase mining, existing approaches often suffer from order sensitive and inappropriate segmentation problems, which make them often extract inferior quality phrases. For topic modeling, traditional topic models do not fully consider the constraints induced by phrases, which may weaken the cohesion. Moreover, existing approaches often suffer from losing domain terminologies since they neglect the impact of domain-level topical distribution. In this paper, we propose an efficient method for high quality and cohesive topical phrase mining. In our framework, we integrate quality guaranteed phrase mining method, a novel topic model incorporating the constraint of phrases, and a novel document clustering method into an iterative framework to improve both phrase quality and topical cohesion. We also describe efficient algorithmic designs to execute these methods efficiently. The empirical verification demonstrates that our method outperforms the state-of-the-art methods from the aspects of both interpretability and efficiency.