Abstract:

More and more text mining methods are applied to patent data analysis to make technology management easier. Patent keyword extraction is an important research direction in patent mining. State-of-the-art strategies for patent keyword extraction are based on bag-of-words representation or distributed model with patent classification. However, these existing strategies need a lot of related field knowledge or pre-defined category corpus. In this paper, we propose an automatic patent keyword extraction method (APKEM) based on patent clustering. Firstly, we apply Skip-gram model on 250,000 patent abstracts to train word embeddings. Secondly, we cluster related patents by Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), and then find each cluster’s center vector. Next, the distance between each candidate keyword vector to its cluster center is calculated by using cosine similarity. Finally, the top n keywords for each test patent document are obtained by sorting the distance calculated in the previous step. The experimental results on our Chinese patent dataset demonstrate the effectiveness of our method.