499 Project: Classification of Academic Papers Using Document Embedding Techniques

Jianguo Lu, <http://cs.uwindsor.ca/~jlu>

Academic papers need to be classified for a variety of reasons. Traditional methods such as bag of words model can not handle word orders and semantic relations effectively. Besides, the dimensionality of the data is extremely high, because the number of distinct words is often in the order of millions. This high dimensional data restrict us with basic classification algorithms such as Naïve Bayes method only. We will explore the document embedding techniques that represent documents with lower dimension, and at the same it can capture the semantic and word order information. Different from the existing word and document embedding techniques, we will add citation network to enrich the representation. It will be a combination of document embedding and network embedding. Our approach will be tested on millions of research papers from CiteSeer, DBLP, and arXiv. The classification results will be evaluated using papers labeled by conferences and journals. There are several challenges of the research. One is the scalability of the existing algorithms. We will utilize the Smart Computing platform via our SOSCIP/OCE/NSERC CRD project, which provides us with high performance computers including dedicated clusters. We will utilize the Spark framework to run our distributed algorithms on the clusters.

The project is to run embedding algorithms on the distributed environment on massive data. We will start by running word embedding algorithms such as Word2vec and Glove. Then, we will expand the experiments to doc2vec algorithm. Once this is done, the next step is to extend doc2vec to include citation information. We will also evaluate the results by classifying the papers, and by visualizing the research papers in a two-dimensional plot using t-SNE.