別紙様式５(Attached Form 5)

学位論文要旨 Abstract of Thesis

所属専攻 Field: Computer Science and Electrical Engineering 専攻(Field)

氏名 Name: RAHUTOMO, Faisal

Title of Thesis

Studies of econo-ESA Method in Semantic Text Similarity
意味的テキスト類似度における econo-ESA 法に関する研究

Abstract

Explicit semantic analysis (ESA) utilizes an immense Wikipedia index matrix in its interpreter part. This part multiplies the huge matrix by a term vector to produce a high-dimensional concept vector. Similarity measurement between two texts is performed between two concept vectors with plenty of dimensions. The cost is expensive in both interpretation and similarity measurement steps.

This thesis proposes an economic scheme of ESA, named econo-ESA. Econo-ESA reduces the ESA index matrix dimensions by "safe dimensional reduction" steps. The aim of this reduction is to reduce the procedure runtime with similar results. We investigate several aspects to econo-ESA proposal: dimensional reduction, experiments with various data, index matrix reduction schemes, and impact of the index matrix density. We run experiments with several index matrix reduction schemes: random selection, k-means clustering, norm-based clustering, densest, and sparsest schemes. We also run experiments with several percentage reductions: 40%, 50%, 60%, and 70%.

This thesis also proposes eight recycling test collections which are used for examining the econo-ESA proposal in semantic text similarity. Semantic text similarity uses specific test collections as its performance evaluation measurement. The test collections consist of text pairs with the same meaning even though in different text

form. The existence is scarce compared with information retrieval test collections. Therefore, this thesis investigates the possibility to reuse information retrieval test collections for semantic text similarity tasks. Text pairs are derived from the relevant pairs of information retrieval test collections. We recycle Glasgow test collections, which consist of eight test collections with various characteristics. This thesis examines the econo-ESA proposal with the recycling test collections.

Evaluation of the recycling test collections yields a promising outcome; the evaluated test collections have low Jaccard index, and their recall values lie between the two baselines. Experimental results for our econo-ESA proposal show both concept reduction and test collection characteristics influence the results; an appropriate concept reduction of econo-ESA can decrease the cost with minor differences from the original ESA results. The results also show both 50% and 60% index matrix reductions, which are called econo50 and econo60, respectively, can be considered as a good candidate for an econo-ESA proposal. We recommend the use of econo50 for long texts and econo60 for short texts. Experimental results for index matrix reduction scheme show that the random selection scheme, which has the nearest density to the original index matrix, gives the best results. We thus conclude that the index matrix density is an econo-ESA feature which has to be considered during the reduction of the index matrix dimensions.