別紙４－１（課程博士（英文））

Date of Submission（month day, year）: January 6, 2023

Department of Computer Science and Engineering	Student ID Number	D199304	
Applicant's name	Setio Basuki		

Supervisors	Masatoshi Tsuchiya Norihide Kitaoka

Abstract（Doctor）

Title of Thesis	The Crucial Role of Citation Functions in The Technology-assisted Peer Review

Approx. 800 words

This research aims to develop a citation functions-based prediction method of paper quality to support Technology-assisted Peer Review (TAPR). The prediction method is intended to reduce the review burden which becomes a critical issue in today's paper submission process. Since the review burden problem has gained much attention, many works have developed the TAPR system to handle this issue. However, most of existing works were created by involving reviewers' comments which is considered unapplicable for reducing the review burden. Addressing this issue, this research proposes a prediction method to estimate the paper quality depending only on the paper itself. The estimator of paper quality used in this paper is citation functions which represent the reason why author of research paper cites previous works. Moreover, the citation functions present the position of the proposed research in wide-ranging literature, understand the broad view of the given research topics, indicates the novelty of the proposed research, and estimate the quality of the proposed research.

The challenge for estimating the paper quality using citation functions will depend on whether the labels are representative enough to capture all potential citation roles in the full text of research paper. Handling drawbacks of current available scheme of citation functions that have a small number of citation instances, few types of labels, and suffer from lack of research variety, this research proposes a new labeling scheme of citation functions covering multi-field computer science domains consisting of 5 coarse labels and 21 fine-grained labels. The annotation experiments on the proposed scheme achieved Cohen's Kappa values of 0.85 for coarse labels and 0.71 for fine-grained labels. The scheme is then used to construct a large dataset of citation functions using a semiautomatic approach which follows two classification stages, i.e., filtering and fine-grained. Adopting the Active Learning (AL) techniques using less than half of training dataset, the Bidirectional Encoder Representations from Transformers (BERT)-based AL in the filtering stage and SciBERT-based AL in the fine-grained stage reached the accuracies of 0.90 and 0.81, respectively. Finally, this research released the largest dataset consisting of 1,840,815 instances.

The prediction method for TAPR, which covers two classification tasks and one regression task, is developed based on the proposed scheme and the best models to create the dataset. While the classification tasks focus on predicting the final review decision (accepted-rejected) and estimating the paper quality (good-poor), the regression task is used to predict the peer review scores. Both classification and regression are implemented using three features i.e., citing sentence features developed based on labeling scheme of citation functions, regular sentence features created by applying the label of citation functions to non-citation text, reference-based features constructed by identifying the source of citations. The classification experiments on the International Conference on Learning Representations (ICLR) 2017-2020 showed that the proposed methods are more effective in the good-poor task compared to the accepted-rejected task by demonstrating the best accuracy of 0.75 and 0.73, respectively. Obtaining as many good papers as possible in the good-poor task, this research reached a satisfying recall of 0.99 by using only the citing sentence features. The regression experiments indicate that the best result in predicting the average review score is higher than in the individual review score by showing RMSE of 1.34 and 1.71, subsequently.

As mentioned above, reaching as high recall as possible is important to get as many good papers as possible, which is more reasonable and applicable for supporting the editor to filter the submitted manuscripts. Interestingly, this highest recall was reached by using only the citing sentences feature. These results prove the hypothesis of this research about the crucial role of citation functions in the manuscript.