INFORMATION OF THE DOCTORAL THESIS

Thesis title:

A STUDY OF EMAIL PRIORITIZATION METHODS

Speciality: **Information Systems**

Code: **9.48.01.04**

PhD. Candidate: **Nguyen Thanh Ha**

Scientific Supervisors:

1. **Assoc. Prof. Tran Quang Anh, PhD**
2. **PhD. Tran Hung**

Training institution: Posts and Telecommunications Institute of Technology

NEW FINDINGS OF THE THESIS:

The thesis generalizes about email and email prioritization, the remaining issues of the email prioritization problem, from which the thesis proposes to build a Vietnamese email dataset and a few novel methods to determine the priority of Vietnamese emails. The thesis has practical significance because of the essential role of email in daily life and the negative impacts of junk emails and email overload problem. The thesis has scientific significance for providing an analysis of existing problems in current email prioritization methods, thereby proposing new methods to solve those problems. The effectiveness of the proposed methods for dealing with spam and email overload is evaluated based on scientific experiments. The contributions of the thesis are as follows:

(1) Propose a method for generating SpamAssassin spam detection rule sets using neural networks. The current machine learning methods for building SpamAssassin rule sets include two independent steps: feature selection and weight training. With this approach, the effectiveness of the selected feature set is not tested on the training data. The proposed neural network model has the ability to merge the two stages of rule selection and weight training, helping to improve the quality of the selected feature set, thereby improving the quality of the built rule set. The method has superior prediction accuracy compared to previous methods.

(2) Propose a method for building SpamAssassin anti-spam rule sets based on multi-objective optimization. In methods based on single-objective optimization, assigning scores and finding the threshold value of the rule set are performed sequentially. The proposed method allows to perform these two tasks simultaneously, thereby solving the important problem of SpamAssassin rule generation problem which is balancing between

two opposing criteria, recall and false alarm rate (FAR).

(3) Propose a method of suggesting user action for incoming emails on the SpamAssassin framework, based on the classification approach. SpamAssassin is a popular spam filtering framework, but there is no built-in user action suggestion function for email. The thesis has applied several techniques for converting from binary classification model to multi-class classification model in order to add action prediction feature to a SpamAssassin-based email system. This method is highly applicable because of SpamAssassin's popularity of and the fast processing speed of its weighted rule mechanism. The thesis also presents two options to improve the efficiency of the said action prediction model by improving the efficiency of individual binary classifiers. The first option adds the ham rules to the SpamAssassin rule set, which helps to reduce the rate of false positives in suggesting the delete action. The second option applies the SpamAssassin rule generation method based on the neural network from one of the author's published studies, which helps to increase the overall prediction accuracy.

(4) Propose a novel method to prioritize Vietnamese emails based on the classification approach. The proposed classification model used deep learning techniques, including the LSTM, word2vec content representation technique, further combined with social features extracted using custom heuristics. The proposed method has significantly better performance than the previous method based on the SVM classifier and TF-IDF feature set for the email prioritization problem.

APPLICATIONS, PRACTICAL APPLICABILITY AND FURTHER RESEARCH DIRECTIONS:

The research results of the thesis are applicable in email server systems, especially those where Vietnamese users are the majority. The applications of the proposals in the thesis can be listed as follows:

(1) To improve the effectiveness of SpamAssassin spam filter on email server systems. SpamAssassin rule sets generated using the thesis' proposed method can be used directly by SpamAssassin.

(2) To add user action prediction feature to email systems which currently use SpamAssassin spam filter. To implement the thesis's proposal, it is necessary to adjust the source code of the email system to execute multiple SpamAssassin rule sets according to the proposed algorithm instead of executing a single SpamAssassin rule set. At the same time, it is necessary to modify the email client application to display prediction results to

users. The method is suitable and easy to implement on open-source email systems.

(3) To add the email prioritization feature to email systems. To accomplish this, it is necessary to build a standalone application which implements the proposed method, as well as to modify the email client application to execute the standalone application and display email priority labels to users.

Confirmation of representative Scientific supervisor	**PhD. Candidate**
Assic. Prof. Tran Quang Anh, PhD	**Nguyen Thanh Ha**