AOEC: A Tool for Automated Online Email Categorization

Zhe Yu, Shijie Li
Amritanshu Agrawal, Di Chen
Com Sci, NC State, USA
{ zyu9,sli41, aagrawa8,dchen12}@ncsu.edu

ABSTRACT

KEYWORDS: Email Categorization, Software Engineering.

Briefly describe general goal (automatically categorize new incoming emails into different folders predefined by users) and motivation here.

1. GENERAL IDEA

Structured as PUTG. The whole project would be the solution.

2. SPECIFIC PROBLEMS

Literature and several user experiences are studied to identify the challenges we may face to achieve our goal. Start with one seed paper [1], each member of the team reviewed three papers, either highly cited or most recent. In addition to literature review, several users are interviewed to express their difficulties using existing tools. The following problems are summarized in the structure of P. U. T. G. S. to provide a guideline for the next step of our project.

2.1 Not Every Fold is Content-aware

2.1.1 Pattern

There are three types of email folders– content-aware, time-aware, and participants-aware– while most of the existing methods can only correctly classify emails with content-aware folders [3].

Content-aware: Folders contain emails of the same topic, e.g. "sports", "music".

Time-aware: Emails in this type of folders are categorized regarding the time they are received, e.g. "2012 Summer".

Participants-aware: This type of folders contain emails with the sender or recipients from a particular group of users, e.g. "Supervisor", "PhD Council".

2.1.2 User Group

CSC 510 North Carolina State University

DOI: 10.1145/1235

The user groups are developers of email client or plug-ins for email client. The common method is entirely based on text categorization and aims on content-aware folders.

2.1.3 Tool

Text categorization is most commonly used for email categorization. Features extracted from the header and body are always specific to the contents.

2.1.4 Goal

The goal is to achieve an email categorization system that can automatically identify the best suited folder for the incoming emails.

2.1.5 Solution

Suggested by [3], three separate models can be built. Each one of the models focuses on one type of folders by constructing the feature set and classifiers specifically according to the characteristic of the target folder type. In the prediction step, the predicted probability of the three models will be used to make a decision fusion and thus decide which folder the incoming email belongs to.

In addition, users will be asked to select the type when they are creating new folders. This simple action of users can greatly facilitate the system in training.

2.2 Less Training Examples

2.2.1 Pattern

Most of the text mining classifiers requires thousands of training examples to build a reliable model. However, more than half the users we interviewed expressed their unwillingness to provide such a great amount of training examples. The number of training examples acceptable is dozens for each folder and hundreds in total according to our interview.

2.2.2 User Group

The user groups are developers of email client or plug-ins for email client. The common method is to collect enough training examples first and then train the classifier [1].

2.2.3 Tool

The common email categorization tools only train once. Therefore it relies heavily on the initial training set and this is why the population of the training set is required to be at least thousands.

2.2.4 Goal

The goal is to achieve an email categorization system that can automatically identify the best suited folder for the incoming emails.

2.2.5 Solution

The incremental learning ability can be the key solution of this problem. Start with hundreds of training examples and a simple model, feedback from user can be collected for the further training of the model.

Three types of user activity can be collected as feedback to the system:

a) user read, replied, or forwarded a certain email without moving it into another folder. This email will be treated as correctly classified.

b) user manually moved an email from the predicted folder to another folder. This email will be treated as misclassified and the latter folder becomes the true label of it in the next round of training process.

c) inspired by active learning, the most confusing emails can be put into a specific folder called "Confusing" that requires the user to manually choose the correct folder.

With the help of user feedback, the system can grow stronger and more adaptive to this certain user and produce less errors as time passes by.

2.3 Importance of Folders may Vary

2.3.1 Pattern

Users definitely do not want to miss a single important email. Some of the folders can be of high importance that False Negative costs much more than False Positive. This phenomenon is commonly described in the binary classification case of email categorization– spam filtering [2]. Failure to identify a spam is always less important to failure to identify non-spam.

2.3.2 User Group

The user groups are developers of email client or plug-ins for email client. The common method is to treat each folder equally in the training process [1].

2.3.3 Tool

The common email categorization tools will predict the folder an incoming email belongs to by comparing the probability of each folder. One email will only be put into one folder with the highest probability.

2.3.4 Goal

The goal is to achieve an email categorization system that can automatically identify the best suited folder for the incoming emails.

2.3.5 Solution

One possible solution for this problem would be to allow each incoming email belongs to several folders. This will change the problem to multi-label, multi-classification problem.

In addition, different weight can be addressed on the labels. User can be asked to put an importance rank when creating folders. For the folders with high rank, the threshold can be reduced to allow emails to go into that folder more easily.

3. REFERENCES

[1] R. Bekkerman. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. 2004.

[2] G. V. Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, 2007.

[3] M. Dehghani, A. Shakery, and M. S. Mirian. Alecsa: Attentive learning for email categorization using structural aspects. *Knowledge-Based Systems*, 2016.