Email Feature Extraction

Shijie Li

Blank Page

Blank Page (Continued next page)

DOI: 10.1145/1235

1. INTRODUCTION

Email has become one of the most important communication tools through the past decades. For most people, there will be hundreds of emails flowing into the mailboxes everyday. These emails can differ in various topics and contents. And reading such a large amount of emails manually in a short time is always a challenge and sometimes even impossible. So the categorizations and classifications of email is crucial to improve the efficiency of dealing with the flow of emails. There were lots of research on the automatic email categorizations popping up these years. And several mature machine learning techniques have been investigated to classify the emails according to their text content.

2. PROBLEM

Previously there are lots of research on the categorizations and classifications of emails. Most of these work applied the common text mining algorithms to the text contents of emails, such as Decision Tree(DT), Naive Bayes classifier(NB), Support Vector Machine(SVM), Artificial Neural Network(ANN) and Ensemble classifier that used several of these with voting mechanisms. Most of these algorithms can work only on numerical inputs. Thus the feature extraction that filters and converts the email texts into numerical representations becomes an important preprocessing stage that is closely related to both the efficiency and accuracy of the classification models.

Most of previous research work use Enron email data set as both training and test data, which was published more than 10 years ago. Early research only focus on the text body of these emails, and use simple bag of words methods to extract the verbal features with removal of stop words. These features are expressed in numerical vectors representing the word frequencies in text emails. Later, researcher applied lexical analysis to the text content to extract more information such as time schedule, actions, sender intent and activities. In addition, geometrical analysis is used to parse the email into structures of different importance to reduce the work load. Almost all of these research work discarded all the attachments and only preserved the text content in the email data set.

However, the expressions of email nowadays are not just confined to the text content. Instead, with the convenience of GUI and embedded HTML as well as the support of MIME(multipurpose Internet mail extension), lots of emails can carry graphical attachments, linkages to on-line information and non-text characters. For example, many people use email to share and discuss photos taken together with their friends or families. And they also suggest interesting on-line videos to friends by providing website linkages. Also, people today prefer to use non-text characters such as Emoji to convey emotions and ideas. In these situations, even the names of the attached files and domains of website links might be more useful than the whole text content to determine the email categorizations. And considering these files or websites are probably viewed by thousands of people who share the same interesting, they are more precise to describe the themes of the parent emails. In addition, some emails even contain only the non-text contents. Instead, they embed HTML codes inside email to present the information neatly. To analyze this type of email content, we should change feature extraction methods because these information is organized in different formats with regular text messages. And usually the formats themselves are important source of features. So it is not a wise decision to ignore all the non-text contents for feature extraction process in the current research on email categorizations.

3. USER GROUP

The users of emails can generally be divided into two groups. One is the professional users, such students/faculties in school, staffs in business companies. They usually received a huge amount of emails during weekdays. Many of these emails are long articles and the topics are closely related to their professional fields with the heavy usage of professional terminology and expressions, such as program codes and design graphs. And they are in desperate need of efficient time managements including checking emails. In addition, they often organize their email categorizations in more reasonable ways. Therefore, accurate email categorizations are easier to implement, and thus can drastically save their time and increase their productivity. The other group of users use emails for regular social activities. Such users also receive lots of emails, but the topics can cover a vast variety, including advertisements and subscribed news. Many of such users cannot check the emails instantly but instead save for later reviews. So emails categorizations can assist them to retrieve information of the imperative concern.

4. TOOLS

To extract information from emails, we choose to set up new experimental email accounts on Gmail, and use the Gmail API to receive the email stream for instant categorizations. Also we plan to use the modified Enron email data set and the WEKA, scikit-learn and Spark packages to develop and improve the text mining models for our purposes.

5. GOAL

The objective of this project is to improve the text mining accuracy on email categorization by adding the features of the non-text content. Different methods of non-text feature extractions will be tested to provide the best balance between accuracy and computational speed. Evaluations of the models will be based on the Recall-oriented measurements, since the categorization is a semi-supervised learning process with the increment of new emails, and users may not provide implicit feedback.

6. SOLUTION

In our project, we mainly use the text mining algorithms to categorize the emails. And considering the computational speed requirements, we will not make a thorough analysis of the non-text content in the emails. Instead, we will extract partial information in those non-text parts and convert them into text labels. For example, instead of analyzing the attached graphs using computer visions, we only extract the text properties such as file names, authors, create dates, graph formats and so on. Such information can be easily found in the original emails. And for the embedded codes, we plan to apply simple text feature extraction methods since the syntax of codes are more logical and predictable. In addition, the linkage domains will be checked on-line to categorize the content efficiently. Private content may be analyzed with preprocessing to hide security information. After detections of these information, we convert them into numerical vectors as described in most previous researches.