Automated E-mail Filtering

Jason Rennie

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The rise of the World Wide Web and the ever-increasing amounts of machine-readable text has caused text classification to become a important aspect of machine learning. One specific application that has the potential to affect almost every user of the Internet is automated e-mail filtering. The WorldTalk Corporation estimates that over 60 million business people use e-mail [4]. Many more use e-mail purely on a personal basis and the pool of e-mail users is growing daily.

As people use e-mail more frequently, it becomes more difficult to organize e-mail in a reasonable fashion. Many e-mail clients include tools that allow the user to define rules for filtering. However, any user that keeps a large corpus of e-mail will find it difficult to maintain such a set of rules. Also, any organizational changes may require the rewriting of most of the constructed rules.

A user with a high volume of e-mail can also benefit from a prioritization system which alerts the user to important e-mail and allows the user to read less important e-mail in batches. This is a principle that Helfman and Isbell built in to Ishmail [1]. However, as with the handling of filtering in most mail clients, Ishmail also requires the user to maintain a large collection of rules and could greatly benefit from an automated system.

Motivation: With the rise of the Internet, computer-readable text has become available in mass quantities. As a result, many researchers have applied traditional machine learning techniques. This application has been a great success and has encouraged entire new fields to develop. The problem of mail filtering is itself a classification problem with close ties to many of the text domains that have been more closely examined by the research community.

Previous Work: While mail filtering has not been given the attention of Web page classification, the research community has already given it some attention. Helfman and Isbell constructed an e-mail prioritization system where the user specifies rules for how frequenly certain types of e-mail are to be viewed [1]. Segal and Kephart have recently applied text classification to the problem of suggesting folders for e-mail filtering [2]. Sahami, et. al. attacked the specific problem of junk e-mail classification and met with great success [3]. William Cohen attacked exactly the problem of e-mail filtering, however, he used a decision rule model and found found that his system may not be efficient enough for real-world use [5].

Difficulty: The two major difficulties associated with mail filtering are the availablity of experimental data and the efficiency required for its real world application. We tackle the first of these by distributing a filter that we created, ifile, for free. Some of the users of our system have been kind enough to allow us to run experiments on their e-mail collections. We deal with the second problem by only using techniques which are efficient and which have been shown to be effective in text classification problems such as mail filtering.

Approach: While there has been previous work in the area of e-mail filtering, not enough attention has been paid to making it practical. In our work, we have focused on this aspect. In that vein we chose Naive Bayes, an efficient yet effective text classifier, for the classification model. Unlike the usual posing of classification problems, the mail filtering problem requires that the model be updated very frequently, optimally with every e-mail that is filtered or moved between mail boxes. The simplicity of Naive Bayes makes it easy to conduct such frequent updating. With efficiency in mind, we also chose a feature selection algorithm that is amenable to efficiency. Yang has shown that selection by word frequency is an effective feature selection technique [6]; it is also efficient for incremental selection.

Impact: The efficiency of our filtering system, ifile, combined with its classification performance have resulted in many users finding it to be a usable and welcome addition to their attempts to prioritize and organize their e-mail collections. We tested the classification performance of ifile on e-mail corpuses of four different users and found that e-mails were, on average, filtered correctly 89% of the time. This translates to the user having to correct an action

file=figures/jrennie-users.ps,width=2.4in,angle=270

Figure 1: Classification accuracies for seven different experiments on four different users. Experiments are ordered from left to right within each user block. Note that no settings produce the best results across all users.

of the filter one in every 9 times. This compares well against similar experiments that Segal and Kephart ran where accuracy scores are between 52% and 76%.

Future Work: Although we have already achieved good results, there is still room for improvement. Joachims has shown that support vector machines (SVM) perform significantly better than Naive Bayes on text classification tasks [7]. While no efficient method has been devised for constructing SVM models, such a discovery would promise even better mail filtering accuracy. Lewis and Knowles study methods for recognizing threads and give near-perfect retrieval results by searching for matching quoted/unquoted text [8]. Such techniques could be used to improve our ability to recognize and correctly classify e-mail messages that are part of the same thread.

References:

[1] John Helfman and Charles Isbell. Ishmail: Immediate Identification of Important Information. http://www.research.att.com/∼jon/ishmail, 1995.

[2] Richard B. Segal and Jeffrey O. Kephart. Incremental Learning in SwiftFile. Proceedings of the Seventeenth International Conference on Machine Learning (ICML-00), 2000.

[3] Mehran Sahami, Susan Dumais, David Heckerman and Eric Horvitz. A Bayesian Approach to Filtering Junk E-Mail. Proceedings of AAAI-98 Workshop on Learning for Text Categorization, 1998.

[4] Dawn Harris and Heather Clark. Worldtalk releases first Internet e-mail corproate usage report; concludes e-mail abuse at epidemic levels. http://www.worldtalk.com/Corporate%20Information/ press%20releases/iecur.htm, 1999.

[5] William Cohen. Learning Rules that Classify E-Mail. AAAI Spring Symposium on Machine Learning in Information Access, 1996.

[6] Y. Yang and J.P. Pedersen. Feature selection in statistical learning of text categorization. Fourteenth International Conference on Machine Learning, 1997.

[7] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European Conference on Machine Learning, 1998.

[8] David D. Lewis and Kimberly A. Knowles. Threading Electronic Mail: A Preliminary Study. Information Processing and Management, 1997.