Introducing the Enron Corpus

Bryan Klimt, Yiming Yang

Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

A large set of email messages, the Enron corpus, was made public during the legal investigation concerning the Enron corporation. This dataset, along with a thorough explanation of its origin, is available at http://www-2.cs.cmu.edu/~enron/. This paper provides a brief introduction and analysis of the dataset. The raw Enron corpus contains 619,446 messages belonging to 158 users. We cleaned the corpus before this analysis by removing certain folders from each user, such as "discussion_threads". These folders were present for most users, and did not appear to be used directly by the users, but rather were computer generated. Many, such as "all_documents", also contained large numbers of duplicate emails, which were already present in the users' other folders. Our goal in this paper is to analyze the suitability of this corpus for exploring how to classify messages as organized by a human, so these folders would have likely been misleading.

In our cleaned Enron corpus, there are a total of 200,399 messages belonging to 158 users with an average of 757 messages per user. Figure 1 shows the distribution of emails per user. The users in the corpus are sorted by ascending number of messages along the x-axis. The number of messages is represented in log scale on the y-axis. The horizontal line represents the average number of messages per user (757).

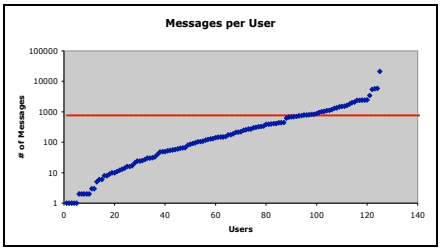

Fig. 1.

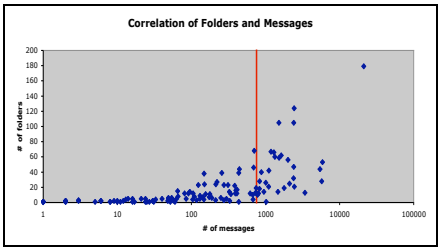

Fig. 2.

As can be seen from the graph, the messages are distributed basically exponentially, with a small number of users having a large number of messages. However, there are users distributed along the entire graph from one message to 100,000 messages, which shows that the Enron dataset provides data for users with all amounts of email. More important in folder classification, though, is the number of folders each user has. The distribution of folders for each user is shown in figure 2. Each point is a user, and shows the number of folders and messages the user has.

Figure 2 illustrates that the Enron dataset is consistent with many of the assumptions made about email folder classification. Most importantly, it shows that most users do use folders to organize their email. Secondly, it shows that the number of messages a user has does not necessarily provide a lower bound for the number of folders that person uses. Some users with many messages have a relatively small number of folders. The number of messages does, however, obviously provide an upper bound for the number of folders the user has. The upper bound for the number of folders of a user appears to be a log of the number of messages of that user. In other words, users with more total messages tend to have more messages in each individual folder.

We also analyzed the characteristics of the email threads in this corpus. A thread is a sequence of messages that form a conversation about a particular topic. For this analysis, emails were considered to be in the same thread if they contained the same words in their subjects and were among the same users (addresses). Out of the total 200,399 messages, we detected 30,091 threads in the corpus, consisting of 123,501 emails. In other words, 61.63% of emails in the corpus are in threads. This makes the average thread size 4.10 messages. The median thread size, however, is only 2. So, there are a few large threads in the corpus, and many small ones. In fact, the distribution of thread sizes is as follows:

thread size:	2	3	4	5	6	7	8	9	10	(10-20]	20+
# of threads	16736	4782	3049	1282	879	903	378	214	178	1260	430

More important than the size of the thread, though, is the amount of information it can provide. The average number of folders containing the messages of a thread is 1.37. This means, given an average thread, the messages in that thread are distributed among only 1.37 folders. This information could be very useful in email folder classification.

The major drawback of this thread information is that it may be redundant when used with the other kinds of evidence. One example of a redundancy problem is with the largest thread, messages with the subject "Demand Ken Lay Donate Proceeds from Enron Stock Sales" belonging to user "lay-k". There are 1124 messages in this thread and they are all in the same folder (Deleted Items)! This would be incredible evidence for folder classification by itself. However, all of the messages in the thread are virtually identical; they appear to be SPAM. Since the messages are identical, there is already incredibly strong evidence from other features, without even detecting the thread.

With the large number of users, messages, and folders in the Enron corpus, we believe it will be a suitable corpus for evaluation of email classification methods. The corpus also has a large number of threads, and would be useful as a test set for email analysis methods that use thread information. For more information, please see "The Enron Corpus: A New Dataset for Email Classification Research", to be published in the proceedings of the 2004 European Conference on Machine Learning (ECML).