Document Classification on Enron Email Dataset

Enron Email Dataset is distributed by William Cohen. The dataset consists of 517,431 messages that belong to 150 users, mostly senior management of the Enron Corp. Although the dataset is huge, topical folders of particular users are often quite sparse. We use email directories of seven users which are especially large. The users are: Sally Beck (Chief Operating Officer), Darren Farmer (Logistics Manager), Vincent Kaminski (Head of Quantitative Modeling Group), Louise Kitchen (President of EnronOnline), Michelle Lokay (Administrative Assistant), Richard Sanders (Assistant General Counsel) and William Williams III (Senior Analyst).

Preprocessing. We remove non-topical folders: *all_documents, calendar, contacts, deleted_items, discussion_threads, inbox, notes_inbox, sent, sent_items and _sent_mail*. We then flatten all the folder hierarchies. After that we remove all folders that contain less than three messages. We also remove *X-folder* field from message headers that actually contains the class label. We do not entirely remove message headers. See the table below for statistics on the seven preprocessed datasets:

Messages in largest folder	Messages in smallest folder	Number of messages	Number of folders	User
166	3	1971	101	*beck-s*
1192	5	3672	25	*farmer-d*
547	3	4477	41	*kaminski-v*
715	5	4015	47	*kitchen-l*
1159	6	2489	11	*lokay-m*
420	4	1188	30	*sanders-r*
1398	3	2769	18	*williams-w3*

Download seven preprocessed datasets (14.7 Mb tarred, gzipped).

Experimental setup. We apply four classifiers (MaxEnt, Naive Bayes, SVM and Winnow). We use Mallet implementations of MaxEnt, Naive Bayes and Winnow (Avrim Blum's version), and SVM*light* implementation of Support Vector Machines. Since email is heavily time-dependent, we cannot use standard random splits for training and test sets. We sort all messages according to the *Date* field and apply incremental timeline splits: we initially train on the first 100 messages and test on the following 100 messages, then we train on the first 200 messages and test on the following 100 messages etc.

Download dataset timelines (130 Kb tarred, gzipped).

Classification results. We report on accuracy over the timeline train/test splits. Click here to see accuracy/timeline graphs of the seven datasets. As it can be seen on the graphs, MaxEnt, SVM and Winnow show similar results, while the results of Naive Bayes are significantly worse. Overall, the results are surprisingly low, probably due to the fact that we apply no feature selection. On one dataset (*williams-w3*) the results are extremely high, while it can be seen from the table above that one half of the dataset belongs to one category, so it is probably not an interesting dataset.

We also plot accuracy over the percentage of test set coverage. For each split, after performing the actual classification, we sort all the test messages according to a classification score and threshold the sorted list so that first 10%, 20%, ..., 100% of messages are chosen. Then we calculate accuracy at each of the 10 thresholds. After that, we average accuracies at each threshold over all the train/test splits. We report on mean accuracy and standard error of the mean at each threshold. Click here to see accuracy/coverage graphs of the seven datasets.

Download all the graphs in EPS, FIG, JPG and PDF formats (1.4 Mb tarred, gzipped).

Publication. We present an extensive case study of email foldering (including the proposal of the evaluation method, discussion on various design choices, application of the four classifiers and comparative analysis of their results) in the following paper:

Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora. Joint work with A. McCallum and G. Huang. *CIIR Technical Report IR-418 2004* pdf