Text Classification in USENET Newsgroups: A Progress Report

Scott A. Weiss and **Simon Kasif** and **Eric Brill**
{weiss, kasif, brill}@cs.jhu.edu
Dept. of Computer Science
The Johns Hopkins University

Abstract

We report on our investigations into topic classification with USENET newsgroups. Our framework is to determine the newsgroup that a new document should be posted to. We train our system by forming "metadocuments" that represent each topic. We discuss our experiments with this method, and provide evidence that choosing particular documents or words to use in these models degrades classification accuracy. We also describe a technique called classification-based retrieval for finding documents similar to a query document.

A Domain For Text Classification

Most work in classification has involved articles taken off of a newswire, or from a medical database(Lewis 1992). In these cases, correct topic labels are chosen by human experts. The domain of USENET newsgroup postings is another interesting testbed for classification. The "labels" here are just the newsgroups to which the documents were originally posted. Since users of the Internet must make this classification decision every time they post an article, this is a nice "real life" application of text categorization. Our approach is to create a model for each group, and then compare future documents to each model to find the best match.

Newsgroups have been studied in other projects. Mostly there has been work done in filtering. In this problem, users attempt to select documents from a large pool that match their interests or are similar to other documents. The SIFT project at Stanford(Yan & Garcia-Molina 1995) has users submit profiles of their interests. It then uses new documents as queries on the profiles. At CMU, the Newsweeder project(Lang 1995) has users rate new documents in terms of relevance, and attempts to learn these ratings to distinguish if future postings are relevant. We will address this problem using the techniques we tried for classification.

Classification Experiments

Our underlying system is SMART, developed by Salton(Salton 1971). SMART indexes a collection of text documents by first removing stopwords, and then stemming words to join those with similar meanings. It then creates a vector for each document whose features represent the remaining terms. A query document is converted to a vector in the same manner, and then compared to each document in the collection by computing the cosine of the angle between the two vectors. Our job then is to create a set of metadocuments that represent each of the newsgroups we wish to classify. We then match a posting with the metadocument that gives the highest cosine similarity.

There were three major approaches that we took towards forming the topic models. The first method is *naive pooling*. For each group, we form a single document that is the concatenation of all of that topic's representatives in the training set. This method involves no analysis of documents. It creates a very large feature space using ostensibly many irrelevant terms. It can incorporate "fluke" documents that are off-topic or without content.

Rather than take all of the documents in the training set, we can try to identify those that are most useful in discriminating topics. We call this method *document selection*. Using SMART, we use each training set posting as a query on the remaining training set elements. Consider a particular document Q, and let N be the number of representatives of Q's group in the training set. (In our experiments, $N = 100$ for all groups.) The documents most similar to Q should be exactly those from its own newsgroup. Hence, optimally, Q should appear amongst the top N retrieved documents exactly for those postings from its own group. While it is foolish to expect this behavior for all documents (and possibly for any), those documents that come close to it could be good discriminators.

We thus use each document in our training set as a query on the entire set. A document is considered to be retrieved by a query if it appears in the top N documents. (This is the breakeven point where recall equals precision.) For each posting, we count how many times

it is retrieved per group. We consider a document to be *discriminating* if it is retrieved more than T times by posts from its own group, and less than T times by the documents in each other group. In other words, such documents are retrieved often by members of their own group, and rarely by members of other groups. Instead of forming a metadocument from all of those available for a particular topic, we just concatenate the group's discriminating documents.

The third method is intuitively appealing. When a person performs this categorization task, it is unlikely that he or she will need to read the entire document. Instead, a person may look for key words that exemplify the topic. If the computer can identify such key terms, then it could also look for those terms in new postings. So we attempt to build metadocuments by *term selection*.

Our approach here is similar in spirit to document selection. We again wish to choose terms that are associated highly with one group, and rarely with the remainder. Hence we count the occurrences of each term with respect to each of the topics. For a given term, we computed the term's frequency in a topic's documents divided by the total number of terms in those documents. This value can be interpreted as the probability that the term is used when describing the topic. We are really interested in the probability that we are describing a particular topic given a term. We compute this using a minor variant on cue validity as used by Jeffrey Goldberg(Goldberg 1995). Here we estimate the probability of a newsgroup G given a term W as:

$$P(G \mid W) = \frac{P(W \mid G)}{\sum_g P(W \mid g)} \qquad (1)$$

If this probability is higher than some chosen threshold, then the term could be a good discriminator. We again choose a term only for the topic to which its cue validity is highest (i.e. its probability is maximum). Cue validity does reward terms that are unique to a particular topic. A term that occurs infrequently in that topic may be a randomly occurring term, and hence will appear rarely in new documents. So along with the cue validity measure, we also require that a term occur with some minimum probability in the appropriate topic. A nice feature of this method is that it automatically removes stopwords such as "the" or "and", since they occur with high probability in all groups. It can also identify a term that is essentially a stopword for a given collection (for example, if the word "computer" appears across topics).

The metadocuments provided to SMART are then lists containing those terms that pass the criteria for a given topic. Each term appears in at most one list, and exactly once in that list. Hence all terms in a metadocument are indexed with the same weight in a given vector.

We experimented with postings taken from different areas of the USENET hierarchy. There were two areas where we extracted postings. One was in the groups with a rec prefix. These groups are concerned with entertainment and hobbies. Another was in newsgroups from the comp hierarchy, i.e. those concerned specifically with computers. We wanted a sufficient amount of data to train and test on, so we only considered groups that had at least 350 messages available at the time of collection. We also only extracted from groups with no subgroups. In the rec collection, we did not take two groups from the same subtree of rec, or rec.arts. This was to give some distance between the topics in our experiment, minimizing overlap between concepts. When dealing with the comp groups, we did not follow this restriction. We extracted 300 documents from each group we selected; 100 of these would be selected randomly for forming the metadocuments, and the remaining 200 would be used for classification testing.

Table 1 gives the results of our experiments. We give results for a collection of 10 groups from the rec hierarchy, and for a collection of 10 groups from the comp hierarchy. The accuracies reported are an average of the classifications reported over ten runs. Notice that, of all of these methods, the naive pooling does the best. This is somewhat counterintuitive. One would believe that our selection routines provide a distilled version of the newsgroup. However, we have apparently eliminated information that SMART can find useful.

We do not currently know exactly what information that is. We have made a few hypotheses. First, people often put signatures on their postings. If a person is a frequent poster to one group, or has a message copied in a particular thread, his name would appear in multiple documents. If we use just one of his posts in training under naive pooling, his name or email address will appear in a metadocument. It is possible that his postings or his address will not occur often enough to warrant inclusion under selection. Secondly, in naive pooling, terms are weighted according to their frequencies in the metadocument. In the term selection process, terms are essentially weighted by the size of the term list. It may be necessary to weight a term based on its cue validity or some other function. Another hypothesis is that term selection does very poorly on a group such as rec.arts.poems where many of the postings are poems, and thus do not have common content. Surprisingly, naive pooling gets 85% accuracy on this group, while the cue validity method of selection gets 47% correct. Possibly long phrases (and hence large sets of words) are picked up in the naive method.

Classification Based Retrieval

While our metadocuments were created for the purpose of classification, we can use them to perform the related task of retrieving related documents. Suppose we have created metadocuments, and now have a large pool of new documents. We describe three ways of finding documents in this collection similar to a given query

How to form Models	Rec groups Avg. Accuracy	Rec groups Std. Dev.	Comp groups Avg. Accuracy	Comp groups Std. Dev.
Naive pooling	0.89	0.01	0.79	0.01
Document Selection Threshold: 50	0.76	0.01	0.72	0.01
Term selection cue validity > 0.6 Prob. in topic > 0.001	0.73	0.02	0.67	0.03

Table 1: Classification accuracies with the rec and comp groups

document:

Ignore the metadocuments

1. Index every document in the pool.
2. Compare every document to the vector for the query.
3. Return the best documents by the cosine similarity measure.

Use the metadocuments as queries

1. Classify the query document as described in the previous section.
2. Compare the topic's metadocument to each document in the pool via cosine similarity.
3. Return the best documents by this measure.

Classification-based retrieval

1. Classify the query document.
2. Classify each other document in the same manner.
3. Return those documents with the same topic as the query.

We investigated the usefulness of this technique in our newsgroup domain. Let us assume that, for a given document, the relevant documents are exactly those from the same group. We can attempt to find documents relevant to a query by comparing each one with the query. We consider as "retrieved" those documents ranked in the top 100. This gives us results at the breakeven point where recall is the same as precision. With respect to the ten groups from the rec domain, this method retrieves an average of only 28% of the relevant documents per query. When we use the metadocuments as queries, we retrieve an average of 57% at the breakeven point. If we use the classification based retrieval described above, we then observe a recall of 79% and a precision of 78%.

One problem in retrieval is that there could be a large number of pairs of documents to compare. Our method avoids this by only requiring that a document be compared once with a small set of topic models. It then never needs to be compared with any other documents (or with just a small set after successive classifications). Also an old document never needs to be compared with a new one as long as both have been classified by topic.

Conclusions

We described different ways of creating metadocuments to exemplify the topics of newsgroups. Each involved using a training set of postings from the groups. In naive pooling, we use all of the available training documents. In document selection, we take posts that we expect will be good discriminators. In term selection, we find words that could represent the ideas in the topic. We discovered that attempts to filter the training set actually hurt classification. We presented some hypotheses as to why.

We next described a technique called classification-based retrieval which can be used to find documents relevant to a query. This does very well in finding documents from the same newgroup as a given posting. It can also be used to reduce the search space for further querying. This method not only outperforms a direct document-document comparison, but also does better than the use of the metadocuments themselves as queries on the new documents.

References

Goldberg, J. 1995. Cdm: An approach to learning in text categorization. In *Proc. of TAI 95, 7th IEEE International Conference on Tools with Artificial Intelligence.*

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, 331–339.

Lewis, D. D. 1992. *Representation and Learning in Information Retrieval.* Ph.D. Dissertation, Department of Computer and Information Science, University of Massachusetts.

Salton, G. 1971. *The SMART Retrieval System - Experiments in Automatic Document Processing.* Prentice Hall.

Yan, T., and Garcia-Molina, H. 1995. Sift - a tool for wide-area information dissemination. *Proceedings of the 1995 USENIX Technical Conference* 177–186.