MACHINE LEARNING TECHNIQUES FOR DOCUMENT FILTERING

Richard M. Tong, Lee A. Appelbaum

Advanced Decision Systems
(a division of Booz•Allen & Hamilton, Inc.)
1500 Plymouth Street, Mountain View, CA 94043

PROJECT GOALS

Booz•Allen & Hamilton's Advanced Decision Systems Division is conducting a program of research to investigate machine learning techniques that can automatically construct probabilistic structures from a training set of documents with respect to a single target filtering concept, or a set of related concepts. These structures can then be applied to individual documents to derive a posterior probability that the document is about a particular target concept.

Our primary goal is to investigate the use of the CART (Classification and Regression Trees) algorithm as the basis of a totally automatic approach to generating document classification structures, working only with information need statements and training data supplied by users. That is, we are interested in testing the hypothesis that effective descriptions of what constitutes a "relevant document" can be constructed using just document exemplars and broad statements about document features. Such a scenario is common in organizations that monitor large volumes of real-time electronic documents.

RECENT RESULTS

Our most recent results are those from the first ARPA sponsored TREC (Text Retrieval Conference) held in November 1992.

The TREC corpus represents a significant challenge for our approach. Our previous results with a small corpus, while encouraging, did not allow us to evaluate how well the technique might do with realistically sized document collections. Our conclusion based on the results we have from TREC is that CART does exhibit some interesting behaviors on a realistic corpus, and that, despite the small size of the training sets and the restricted choice of features, for some topics it produces competitive results. So although the overall performance is moderate (relative to the better performing systems at TREC), we believe that the absolute performance (given that the system is totally automatic) is at least encouraging and definitely acceptable in several instances.

Some specific observations on the performance of the current implementation of the CART algorithm as used for TREC are:

- Relying on the re-substitution estimates for the terminal nodes is a very weak method for producing an output ranking. A scheme that makes use of surrogate split information to generate a *post hoc* ranking shows much promise as a technique for improving our scores in the TREC context.
- While our approach is totally automatic, we restricted ourselves to using as features only those words that appear in the information need statement. This is obviously a severe limitation since the use of even simple query expansion techniques (e.g., stemming and/or a synonym dictionary) is likely to provide a richer and more effective set of initial features.
- Using words as features is possibly too "low-level" to ever allow stable, robust classification trees to be produced. At a minimum, we probably need to consider working with concepts rather than individual words. Not only would this reduce the size of the feature space but would probably result in more intuitive trees.
- We need to work with much bigger and more representative training sets. Our preliminary experiment in this area shows, not surprisingly, that adding more training examples can lead to dramatic changes in the classification trees.

PLANS FOR THE COMING YEAR

The main activity planned for the coming year is to participate in TREC 2. We intend to perform a series of additional experiments designed to explore some obvious extensions suggested by the TREC 1 results. That is we will perform experiments to determine: (1) the effect of training set size on the overall performance of the learning algorithm, (2) the effectiveness of using surrogate splits information to help perform a *post hoc* ranking of the documents classified as relevant, (3) the effectiveness of knowledge-based query expansion techniques, and (4) the value of using concepts, detected using our RUBRIC text retrieval technology, as document features.

To facilitate the experimental procedure we plan to integrate the CART technology into Booz•Allen & Hamilton's distributed information integration system (MINERVA). The primary feature of MINERVA is a distributed operating environment, which is an implementation of an intelligent Ethernet token ring that uses TCP/IP and standard UNIX socket protocols. This provides a unique transport layer that allows multiple databases and information access services to communicate transparently using a common metalanguage.