Adaptive Information Filtering with Minimal Instruction
MIT2000-08

Progress Report: July 1, 2000—December 31, 2000

Tommi Jaakkola, Tomaso Poggio and David Gifford

Project Overview

This project concerns with automated methods for finding a few pieces of relevant information (such as research articles) within a large dataset of predominantly incomplete and superficially similar information (such as technical report archives). While many such information filtering tasks vary considerably depending on the context, the primary difficulties associated with automated techniques are mostly shared across different tasks. These difficulties give rise to several challenges that modern information filtering algorithms need to be able to solve. In this project we exploit and further develop a specific synthesis of modern machine learning techniques to address these challenges. The tools we develop have the ability to function accurately with mininal instruction of what is relevant, learn from related filtering problems, and make use of any optional feed-back provided by or automatically queried from the user.

Progress Through December 2000

We have made significant progress despite the fact that we were able to start the project at full speed only recently.

First, we have developed efficient ways of exploiting incomplete information. Our method expands documents (or other examples) into a representation that captures relations among all the available documents. This expanded representation permits fast convergence to the optimal classification decisions as the number of labeled examples increases. Preliminary experimental results are encouraging. Part of this work was published in the NIPS 2000 proceedings.

Second, we have developed a new approach to eliciting user feed-back for retrieval purposes. This approach is firmly founded on information theory and works as follows. Based on limited information initially provided by the user (for example in the form of keywords), the system presents to the user a brief list of documents (forced choice set) along with the usual list of top ranked documents. The brief list of documents/pages/links are designed to reveal as much information about the unknown document of interest as possible. Our framework is built on a similarity measure between the documents from which we can derive "substitution probabilities" among the documents. On the basis of these probabilities as well as our initial or current beliefs (ranking) of the documents, we can identify the most revealing set of documents to present to the user. The elicited user response (choice of one of the highlighted documents) subsequently yields most information about the document of interest. The user

feed-back in this formulation is unambiguous and can be readily incorporated into the existing beliefs so that the procedure may continue. We are currently preparing a publication on this material.

Third, we have developed new sequential methods of generating error correcting codes for multi-way classification tasks. As most information filtering tasks involve a large number of potential categories, such methods are necessary. We are in the process of implementing these approaches and will test them in the context of document retrieval problems.

Fourth, as most information filtering problems involve large amounts of potentially useful but largely incomplete information, it is important to determine when we can and cannot expect to be able to exploit such information. We have made considerable progress in understanding the fundamental limitations of various retrieval methods that involve the use of generative probability models and the standard EM algorithm. We are currently writing a publication about these results.

We are also carrying out related work in computational biology where fusion of various sources of information is necessary for making accurate predictions of the function of specific genes.

Research Plan for the Next Six Months

The development of the underlying methodology is starting to reach the point where we can begin implementing proof of concept tools. We will do this in parallel with further generalization and refinement of the methodology. Among other things, we will implement and test our method of optimally eliciting user feed-back test multi-class filtering methods relying on sequential construction of error correcting codes. This part of the project is likely to be finished as a Master's thesis project. refine existing methods for exploiting incomplete information based on the insights we have already derived from the characterization of fundamental limitations. We will test and demonstrate the effectiveness of such modifications develop and test a formulation of transfer of knowledge across multiple filtering tasks. We believe that incorporating such inferential power in the retrieval system will have revolutionary consequences.