Adaptive Information Filtering with Minimal Instruction MIT2000-08

Progress Report: January 1, 2002—June 30, 2002

Tommi Jaakkola and Tomaso Poggio

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 challenges associated with automated techniques are often shared across different tasks. In this project, we develop the foundations for adaptive information retrieval tools with the ability to function accurately with minimal instruction of what is relevant, learn from related filtering problems, and make use of any optional feedback automatically queried from the user

Progress Through June 2002

Our progress during the six month period has been in three key areas: 1) active querying of information, 2) estimation with predominantly incomplete information, and 3) collaborative filtering. We provide here a brief description of the results with links to associated publications and presentations.

1. Active querying of information

We formulate the problem of retrieving information from a database (or web) as an active learning problem, where the user is automatically queried for additional information in response to a user-initiated query. The information is elicited from the user at multiple levels of abstraction to quickly determine the set of elements that the user is after. The interaction with the user defines a restricted information channel, which we exploit to minimize the overall time of the exchange. The overall framework has been described in previous reports along with the resulting publications.

For this framework to be applicable we must understand the structure and relations among the elements in the database. When the structure is unknown we seek to recover the structure with minimal effort. We approach this sub-problem also as an active learning problem. This sub-problem can be cast more generally as the problem of recovering the structure of a graphical model (Bayesian network) with minimal number of queries. We have developed a computationally feasible approach to this problem. The key idea is to ensure that we can maintain and update only a few candidate structures at each stage and minimize an appropriate measure of disagreement among the candidates. The details can be found in

Steck and Jaakkola, "Unsupervised Active Learning in Large Domains", Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence, 2002.

We are also in the process of extending the overall active retrieval framework to cases where the documents (or elements in the database) naturally belong to overlapping sets of subtopics. This changes the type of information the user may be after, the set of queries that can be performed, as well as how we can expect the user to respond to the new queries. We are pursuing this work in collaboration with Dr. Naonori Ueda (NTT) whose recent work on multi-category labeling complements our progress on active learning.

2. Estimation with predominantly incomplete information

The available data for text filtering involves predominantly incomplete or fragmented information. A typical realization of this problem involves a few annotated documents exemplifying the filtering task and a large database of unannotated documents. In general, incorporating large amounts of incomplete data can either dramatically increase or decrease the filtering performance.

We have developed two fundamentally new approaches to this problem. First, we address the inherent instability of exploiting heterogeneous sources of information in estimation. Our approach (partially described in previous reports) efficiently recovers a continuum of estimation solutions at varying levels of mixing of the sources. This permits us to identify and avoid any instabilities due to mixing. The paper is available from

Corduneanu and Jaakkola, "Continuation Methods for Mixing Heterogeneous Sources", Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence, 2002.

The associated UAI 2002 conference presentation is also available

We have given several seminar talks about this approach including those at MIT, University of Massachusetts at Amherst, and Yale University.

Our second approach characterizes how unlabeled documents can be tied to their hypothetical labels. The key intuition is that any tight clusters of documents should possess unambiguous labels. We can express this idea mathematically as a regularization approach, where the marginal distribution of documents constrains how the conditional distribution of labels given documents is chosen. The regularization limits the amount of information that related documents can contain about the labels. The details can be found in

Szummer and Jaakkola, "Information Regularization with Partially Labeled Data", submitted, 2002.

3. Collaborative filtering

It is often beneficial to solve multiple filtering tasks concurrently so as to exploit any similarities and regularities across the tasks. Collaborative filtering exploits this idea in contexts that involve multiple users or a single user across multiple filtering criteria. We have developed a new approach to this problem by casting the joint filtering task as a latent matrix factorization problem. The resulting low rank matrix captures the regularities across the related filtering tasks. The underlying mathematical problem and our solution is described in

Srebro and Jaakkola, "Generalized Low-Rank Approximations", submitted, 2002.

Research Plan for the Next Six Months

There are several key goals that we hope to achieve over the next six months:

In collaboration with Dr. Naonori Ueda, we seek to extend the overall active learning framework to settings where the documents possess multiple simultaneous category labels (topics). More precisely, we extend the framework to efficiently annotate documents in large databases with multiple topic labels. A related problem is to efficiently identify documents (in unannotated databases) that conform to user specified constraints on the sub-topics.

We hope to release software for combining heterogeneous data sources. Our method provides a stable solution to a wide variety of estimation tasks involving multiple sources of information of varying type or quality.

We extend and test our information regularization approach in filtering tasks involving large numbers of unlabeled documents. The computational complexity of this approach is inherently tied to the resolution at which we wish to solve the problem and does not scale directly with the size of the dataset.

We employ the new matrix factorization approach to concurrently solve multiple related filtering tasks. The goal is to extend and test the approach so that it can be appropriately integrated with the active learning framework.

A more longer term goal involves the design and implemention of user interfaces supporting the overall interactive approach to information retrieval. The basic elements for this are already in place; our on-going work progressively refines these elements.