Adaptive Information Filtering with Minimal Information MIT2000-08

Progress Report: July 1, 2001—December 31, 2001

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 feed-back automatically queried from the user

Progress Through December 2001

Our work is divided into four complementary areas of the overall problem:

1. active querying of information
2. estimation with predominantly incomplete data
3. exploiting the structure of the elements to be filtered
4. extending multi-way filtering methods

We have made rapid progress in these areas and the work has resulted in several publications and presentations including two Master's theses. We have recently started a collaboration with Dr. Naonori Ueda (NTT Communication Science Labs) on active learning.

1. Active querying of information. The problem of retrieving information from a database (or web) is formulated 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 basics of this active learning formulation were already described in the previous report. We have extended the framework considerably in terms of the allowable user operating modes (contrastive selection and annotation) and the associated user models and inference algorithms needed to properly interpret user responses, maintain beliefs about the documents of interest, and to optimize each successive query. All the algorithms scale linearly with the database size and the capacity of the information channel. The extended framework is explained in

T. Jaakkola and H. Siegelmann. Active information retrieval. In Advances in Neural Information processing systems 14, 2001. http://www.ai.mit.edu/people/tommi/papers/JaaSie-nips01.ps.gz

A presentation of these ideas containing results from our proof-of-concept implementation can be retrieved from

T. Jaakkola and H. Siegelmann. Active information retrieval. NIPS*01 presentation. http://www.ai.mit.edu/people/tommi/projects/ntt/JaaSie-nips01-presentation.pdf

We have also begun developing user interfaces that facilitate proper interpretation and flexible elicitation of information from the user. The interfaces and the user models need to be developed together.

2. Estimation with predominantly incomplete data. 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 a fundamentally new approach to the problem of combining all the available information sources in a stable and accurate manner. Our approach replaces the standard (EM) algorithm and is based on efficient evolution of differential equations that govern how the solutions change at varying levels of source allocation. The paper describing this approach can be obtained from

A. Corduneanu and T. Jaakkola. Stable mixing of complete and incomplete information. MIT AI Memo AIM-2001-030, 2001. http://www.ai.mit.edu/people/tommi/papers/AIM-2001-030.pdf

This work, partially explained in the previous report, has been extended and generalized substantially. These extensions as well as various tests of the ideas can be found in the Master's thesis:

Adrian Corduneanu. Stable mixing of complete and incomplete information. Master's Thesis. Massachusetts Institute of Technology. http://www.ai.mit.edu/people/adrianc/papers/mthesis.pdf

An invited talk about the overall problem and solutions can be retrieved from

T. Jaakkola. Classification with incomplete labels. Invited talk, IJCAI*01 workshop on text learning. http://www.ai.mit.edu/people/tommi/projects/ntt/Jaakkola-IJCAI01.pdf

We are in the process of integrating these results into an active learning framework. Briefly, the idea is to perform annotation queries from the user so as to regain estimation stability when a more aggressive combination of the information sources would otherwise result in unpredictable results.

3. Exploiting element structure. The structure and relations among the elements in the database are important from the point of view of effective filtering. For example, natural clusters among the elements are typically associated with unambiguous labels; this drastically limits the number of ways that the documents in the database can be consistently labeled and serves to minimize the amount of information needed from the user.

We have developed a new approach to exploiting such structure in a comprehensive manner. The approach is based on a neighborhood graph over the database elements as well as a Markov random walk operating on this graph. The random walk captures the manifold structure of the data at multiple levels of resolution. This formulation gives rise to a new representation of the elements which relies on how the elements relate to each other at desired levels of resolution. The paper describing our approach is available from

M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. In Advances in Neural Information processing systems 14, 2001. http://www.ai.mit.edu/people/tommi/papers/SzuJaa-nips01.ps.gz

A conference presentation of these ideas can be found in

M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. NIPS*01 presentation. http://www.ai.mit.edu/people/szummer/papers/SzummerJaakkola-nips01-poster.ps.gz

4. Multi-way classification. Most information filtering tasks involve a large number of potential categories. It is therefore important to develop methods for effective and robust multi-way classification. We have extended and tested approaches based on error correcting codes. This work is discussed in the Master's thesis

Jason D. M. Rennie. Improving Multi-class Text Classification with Naive Bayes. Master's Thesis. Massachusetts Institute of Technology. AI Technical Report AITR-2001-004. 2001. ftp://publications.ai.mit.edu/ai-publications/2001/AITR-2001-004.pdf

and, from the point of view of discriminative classification, in

Jason D. M. Rennie and Ryan Rifkin. Improving Multiclass Text Classification with the Support Vector Machine. Massachusetts Institute of Technolgy. AI Memo AIM-2001-026. 2001. ftp://publications.ai.mit.edu/ai-publications/2001/AIM-2001-026.pdf

Research Plan for the Next Six Months

The emphasis will be on integrating results from the four main areas, implementing proof-of-concept tools, and further developing the underlying retrieval methodology. Our work will concentrate primarily on the following topics:

Integrate stable estimation (area 2) and multi-resolution document representations (area 3) with the overall active retrieval approach (area 1)

Further develop active learning methods for the purpose of accurate classification of examples with minimum number of labels. This is will be done in part in collaboration with Dr. Ueda.

Examine and test problem structures that will facilitate filtering with limited information from the user. The idea is to clarify and formally establish classes of retrieval problems where our active learning approach can be guaranteed to work well

Formalize retrieval across multiple tasks to exploit transfer of knowledge across related retrieval tasks. We have already solved components of this problem, for example, in the contex of optimizing successive active queries. Our stable estimation approach will provide the appropriate inferential foundation.

Design and implement user interfaces supporting the overall interactive approach to information retrieval. We have outlined the necessary semantics for such interfaces and we are in the process of integrating it with automated queries.