Adaptive Information Filtering with Minimal Instruction
MIT 2000-08

Progress Report: January 1, 2001—June 30, 2001

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 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 2001

Our work has progressed rapidly along four related but complementary fronts. Several recent publications have resulted from this project.

First, we have formalized an active learning approach to information retrieval, where the user is automatically queried for additional information at multiple levels of abstraction so as to quickly determine the piece of information the user is after. A paper outlining the mathematical foundations and associated algorithms for this approach is available from

T. Jaakkola and H. Siegelmann (2001). Active information retrieval. Submitted.
http://www.ai.mit.edu/people/tommi/projects/ntt/JaakkolaSiegelmann.ps

We have also started developing a user interface that facilitates the proper interpretation of queries and exploits the flexibility permitted within the overall active learning framework.

Second, we have developed a fundamentally new approach to the problem of combining multiple predominantly incomplete sources of information in a stable and accurate manner. The inherent problem, present in most retrieval and information filtering tasks, pertains to the large number of ways that the incomplete sources of information could be completed. Our approach replaces the standard (EM) algorithm in this context and is based on the idea of efficiently evolving differential equations that govern solutions at varying levels of source allocation. The associated paper can be obtained from

A. Corduneanu and T. Jaakkola (2001). Stable mixing of complete and incomplete information. Submitted.
http://www.ai.mit.edu/people/tommi/projects/ntt/CorduneanuJaakkola.ps

We are working on several extensions of this approach including active labeling of documents.

Third, we have developed alternative ways of exploiting the sparse similarity structure of documents for the purpose of accurate labeling. Such structure can be extracted from a given database without prior annotation and can be used to efficiently limit the number of ways that the documents in the database could be labeled consistently. This permits learning with very little input (labels) from the user. A draft paper describing our approach is available from

M. Szummer and T. Jaakkola (2001). Clustering and efficient use of unlabeled examples. Submitted.
http://www.ai.mit.edu/people/tommi/projects/ntt/SzummerJaakkola.ps

Fourth, we are in the process of developed and testing error correcting codes for effective multi-way classification. Since most information filtering tasks involve a large number of potential categories, such extensions are necessary. A *Master of Science* thesis is currently being written on this topic (not available yet).

Research Plan for the Next Six Months

We have started implementing the proof of concepts tools along with further development and refinement of the underlying methodology. Our work will primarily concentrate on the following tasks:

1. Design and implementation of the user interface for active information retrieval. We have already outlined the necessary semantics for such an interface and are in the process of integrating it with the core computational methods.
2. Extension of the basic active learning framework towards more flexible ways of eliciting user feedback. This development complements the user interface design.
3. Further development and implementation of the source allocation method. We also plan to extend this approach to the problem of actively querying further information (e.g., labels) from the user so as to regain stability and predictability in the context where a more aggressive but unguided combination of the sources could fail catastrophically.
4. Begin formulating transfer of knowledge across multiple filtering tasks. Results briefly outlined in the previous section already solve components of this problem. Properly exploiting transfer can substantially enhance the inferential power of the retrieval system.