Adaptive Information Filtering with Minimal Instruction
MIT2000-08
Tommi Jaakkola and Tomaso Poggio

Collaboration with NTT

The project got started quite late and thus we have not yet established an active collaboration.

The most significant progress:

We have developed a new approach to eliciting user feed-back for the purpose of retrieving documents from large relatively unstructured databases. The approach is firmly founded on information theory and optimally selects the type and level of information to be queried in order to minize the effort on the part of the user.

Brief 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 filtering 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. 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 appropriate use of feed-back automatically queried from the user.

Proposed research:

The research in this project will focus on two related key problems: active information retrieval and dealing with incomplete information in adaptive information filtering tasks. Both of these problems have remained largely unsolved. The user is often perceived as an unwilling or ineffective participant and no serious attempt has been made about dealing with the fragmented and incomplete nature of the available information. Both problems, however, admit effective solutions with substantial benefits.

In our active formulation of the retrieval process, the user is successively queried for distinctions at varying levels of abstraction (links, summaries, topics) and is permitted to respond with multiple selections or may choose not to respond. In each case, the information is unambiguously interpreted and incorporated by the system. The next query is chosen optimally to minimize the need for any further exchange. Our information theoretic formulation also permits us to determine whether the document of interest is in the (portion of the) database being consulted. We will focus on various extensions of this basic paradigm as well as on implementing and testing proof of concept tools.

The second part will concentrate on a related challenge of dealing with incomplete information in adaptive filtering systems. The setup is complementary to active retrieval and the user is regarded here as a passive source, willing to provide only mininal information about what is relevant and what is not. When searching for technical articles, for example, the user may be willing to provide only a single prototype of what is meant by "relevant". Accurate filtering in this context necessarily involves systems with the ability to extract some structure from the large set of unlabeled documents in the database and combine this information with the few annotated examples. The primary research objective here is to develop and test algorithms that attain near optimal filtering performance in the presence of incomplete information.

The two complementary approaches will be subsequently combined into a more effective system and tested along with the invidual components.