MIT2000-08
Adaptive Information Filtering with Minimal Instruction

Tommi Jaakkola, Tomaso Poggio and David Gifford

Abstract: 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) has become one of the pervasive challenges of information technology. In this project we exploit and further develop automated information filtering methods arising from a specific synthesis of modern machine learning techniques. 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. The results of this project can be readily translated into various applied and commercial uses. We plan to build proof of concept tools specifically aimed to allow flexible document retrieval and filtering algorithms for various databases in molecular biology.

Introduction: Effective separation of objects on the basis of their properties is essential in many application areas. This includes text classification such finding relevant documents in response to queries expressed through keywords or via a few examples of relevant documents. Many problems in image categorization, retrieval, as well as aspects of speech recognition can be addressed with such separation methods. Mining scientific databases also falls within this class of problems. For example, we may wish to distinguish protein sequences in large databases on the basis of their inferred functional roles or search for protein binding sites in genomic DNA. There are also various commercial applications such as fraud detection and targeted advertising or marketing. While there are considerable differences in the details of these application areas, many of the primary difficulties associated with techniques aspiring towards efficient separation are mostly shared across these problem areas. These difficulties give rise to several challenges that modern information filtering algorithms need to be able to solve. We identify here four main challenges: 1) finding object representations that facilitate efficient separation, 2) dealing with various forms of incomplete information, 3) be able to fuse information from multiple sources, and 4) be able to exploit knowledge about one task to better solve another, related task (concept known as transfer).

Representation: The representation used for the objects such as documents in the filtering process should be conducive to efficient separation on the basis of the type of distinctions that a specific user is interested in making; these representations must therefore be adaptive and modified through experience with the user. Moreover, the objects to be processed in the areas discussed above have special structure such as variable length sequences which refuse to lie in a vector space (this holds in many cases even if we could, in principle, interpret the examples as points in a vector space as in the context of pixel images). Any comparison of such structured examples typically also presumes a pairwise alignment. For example, speech signals arising from different utterances of the same word should be compared by matching phonetically identical components. Moreover, the relevance label that we wish to assign to the examples in the database often encode rather abstract properties, e.g., research area of an article or the functional role of a protein sequence. This abstract categorization can give rise to rather diverse class-conditional populations, making it more challenging to represent and find the appropriate decision rule.

Incomplete information: We can expect the user to provide only a mininal number of example documents (or other information) about what is relevant and what is not. For example, when searching for research articles, the user may be willing to provide only a single prototype of what is meant by "relevant". To apply any standard discrimination method, however, we would need to amass a large representative sample of already labeled examples (documents known to be relevant or not). The construction of such a labeled training set would invariably require considerable human intervention. The prohibitive cost of this approach implies that any reasonable discrimination method must obtain reliable decision rules on the basis of only few labeled examples. However, considerable benefit can be derived from any available unannotated examples. Another source of missing information comes from partially specified objects (e.g., research articles without abstracts or references). Our ability to effectively deal with various types of missing information is necessary to adequeately utilize the increasingly large and overwhelmingly incomplete databases such as the web, various databases in molecular biology, finance, and other fields.

Information fusion: Information fusion is required at multiple levels. Even the simplest filtering method must fuse information from different sources. For example, the presense or absense of specific words in a document can be viewed as indicators contributing to the decision about whether the document is relevant or not. More generally, the fusion of different types of sources can be achieved by treating the sources as individual expert predictions pertaining to the same decision. The expert predictions can be weighted according to how reliable they are, where the reliability is either modeled explicitly or estimated through experience.

Transfer: Solving many related filtering tasks independently is not an efficient approach. We would expect some benefit from already solved but related filtering tasks. This is the problem of transfer and it is closely related to the challenges described above, particularly the problem of information fusion. Other forms of transfer are also possible and arise frequently in the context of information filtering. Consider, for example, the situation where the user indentifies a few example documents and insists that they are related without actually explicating what the relation is. The filtering task here is to find all the documents that share this unspecified relation with the given example documents. Note that the task is not merely to find the intersection of documents that are separately related to the individual documents given by the user (this is the way the problem is currently addressed). Filtering methods that can exploit the notion of transfer (even in a restricted form) can significantly and fundamentally change the utility of information filtering techniques.

Project goals: The different challenges described above can be already addressed to a degree through a specific synthesis of existing techniques in machine learning and statistical inference. Our plan in this project is to exploit and further develop this synthesis while, in parallel, building proof of concept tools in two specific application areas for the filtering methods

1. Inferential tools for document retrieval. We will initially start from small collections of research articles (electronic journals) and move towards more general filtering problems on the web.

2. Databases in molecular biology including protein and DNA sequence databases as well as high density DNA array datasets. These databases contain diverse types of objects/information and therefore provide a reasonable test bed for general purpose filtering methods. In addition, the resulting tools can be readily applied to mine scientific knowledge from a collection of such databases.

Technical approach: Recent work on information retrieval and extraction combined with our work on statistical graph models (Jaakkola and Jordan, 1999), kernel machines for regression, classification and density estimation (Evgeniou et al., 1999; Mukherjee and Vapnik, 1999) and their application to various categorization tasks (Papageorgiou et al.,1999; Mukherjee et al. 1999), efficient representations for discrimination tasks (Jaakkola and Haussler 1998, Jaakkola et al. 2000), and various generalizations of discrimination tasks and associated techniques (Jaakkola et al 1999) provide the essential components for the synthesis that we exploit and further develop as part this project.