DISSERTATION ABSTRACT

Bayesian Graphical Models for Adaptive Filtering

Yi Zhang (Advisor: Jamie Callan)
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

A personal information filtering system monitors an incoming document stream to find the documents that match information needs specified by user profiles. The most challenging aspect in adaptive filtering is to develop a system to learn user profiles efficiently and effectively from very limited user supervision.

In order to overcome this challenge, the system needs to do the following: use a robust learning algorithm that can work reasonably well when the amount of training data is small and be more effective with more training data; explore what a user likes while satisfying the user's immediate information need and trade off exploration and exploitation; consider many aspects of a document besides relevance, such as novelty, readability and authority; use multiple forms of evidence, such as user context and implicit feedback from the user, while interacting with a user; and handle various scenarios, such as missing data, in an operational environment robustly.

This dissertation uses the Bayesian graphical modelling approach as a unified framework for filtering. We customize the framework to the filtering domain and develop a set of solutions that enable us to build a filtering system with the desired characteristics in a principled way. We evaluate and justify these solutions on a large and diverse set of standard and new adaptive filtering test collections. Firstly, we develop a novel technique to incorporate an IR expert's heuristic algorithm as a Bayesian prior into a machine learning classifier to improve the robustness of a filtering system. Secondly, we derive a novel model quality measure based on the uncertainty of model parameters to trade off exploration and exploitation and do active learning. Thirdly, we carry out a user study with a real web-based personal news filtering system and more than 20 users. With the data collected in the user study, we explore how to use existing graphical modeling algorithms to learn the causal relationships between multiple forms of evidence and improve the filtering system's performance using this evidence.

A copy of the dissertation is available at
http://www.soe.ucsc.edu/~yiz/www/papers/thesis.html