DESCRIBER: Graphical Relational Models for Collaborative Filtering in Microarray Data Mining

William H. Hsu[1], Roby Joehanes[2], Prashanth Boddhireddy
[1]bhsu@cis.ksu.edu, Kansas State University; [2]robbyjo@cis.ksu.edu, Kansas State University

Collaborative filtering is the problem of analyzing the content of an information retrieval system and actions of its users, to predict additional topics or products a new user may find useful. Developing this capability poses several challenges to machine learning and reasoning under uncertainty. The research described in this summary addresses the problem of formulating tractable and efficient problem specifications for probabilistic learning and inference in this framework. It describes an approach that combines learning and inference algorithms for relational models of semi-structured data into a domain-specific collaborative filtering system. Recent systems such as *ResearchIndex* / *CiteSeer* have succeeded in providing some specialized but comprehensive indices of full documents. The collection of user data from such digital libraries provides a test bed for the underlying IR technology, including learning and inference systems. The authors are therefore developing two research indices in the areas of bioinformatics (specifically, functional genomics) and software engineering (digital libraries of source codes for computational biology), to experiment with machine learning and probabilistic reasoning software recently published by the authors and a collaborative filtering system currently under development. The overall goal of this research program is to develop new computational techniques for discovering relational and constraint models for domain-specific collaborative filtering from scientific data and source code repositories, as well as use cases for software and data sets retrieved from them. The focus of this project is on statistical evaluation and automatic tuning of algorithms for learning graphical models of uncertain domains from such data. These include probabilistic representations, such as Bayesian networks and decision networks, that have recently been applied to a wide variety of problems in intelligent information retrieval and filtering. The primary contribution of this research shall be the novel combination of algorithms for learning the structure of relational probabilistic models with existing techniques for constructing relational models of metadata about computational science experiments, data, and programs. The technical objectives center around statistical experiments to evaluate this approach on data from the domains of gene expression modeling and indexing of bioinformatics repositories. Though widely used, systems such as *ResearchIndex* face limitations in their direct application to IR from computational genomics repositories: 1. Over-generality: Citation indices and comprehensive web search engines are designed for the generic purpose of retrieving all individual documents of interest, rather than collections of data sets, program source codes, models, and metadata that meet common thematic or functional specifications. 2. Over-selectivity: Conversely, IR systems based on keyword or key phrase search may return fewer (or no) hits because they check titles, keywords, and tags rather than semi-structured content. 3. Lack of explanatory detail: A typical user of an integrated collaborative filtering system has a specific experimental objective, whose requirements he or she may understand to varying degree depending upon his or her level of expertise. The system needs to be able to explain relationships among data, source codes, and models in the context of a microarray data mining experiment. We present the design of *DESCRIBER*, a prototype research index for consolidated repositories of computational genomics resources, along with preliminary collaborative filtering results using structure learning algorithms for graphical relational models. The unifying goal of this research is to advance the automated extraction of graphical models of use cases for computational science resources, to serve a user base of researchers and developers who work with genome data and models. We present recent results from our own work and related research that suggest how this can be achieved through a novel combination of probabilistic representation, algorithms, and high-performance data mining not previously applied to collaborative filtering in bioinformatics. This continuing work aims at facilitating gene expression modeling and intelligent, search-driven reuse in distributed software libraries.