Using Probabilistic Relational Models for Collaborative Filtering

William H. Hsu **Prashanth Boddhireddy** **Roby Joehanes**

Laboratory for Knowledge Discovery in Databases, Kansas State University
234 Nichols Hall, Manhattan, KS 66506
{bhsu | pbo8844 | robbyjo}@cis.ksu.edu http://www.kddresearch.org

Abstract

This research summary describes some work in progress on using graphical models to represent relational data in computational science portals such as *myGrid*. The objective is to provide a integrative *collaborative filtering* (CF) capability to users of data, metadata, source code, and experimental documentation in some domain of interest. Recent systems such as *ResearchIndex* / *CiteSeer* provide collaborative recommendation through citation indexing, and systems such as *SourceForge* and the Open Bioinformatics project provide similar tools such as content-based indexing of software. Our current research aims at learning *probabilistic relational models* (PRMs) from data in order to support intellignet retrieval of data, source code, and experimental records. We present a system design and a précis of a test bed under development that applies PRM structure learning and inference to CF in repositories of bioinformatics data and software.

Keywords: probabilistic relational models, collaborative filtering, information retrieval, source code repositories, structure learning

1 INTRODUCTION

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*.

1.1 Rationale

Recent systems such as *ResearchIndex* / *CiteSeer* [LGB99] have succeeded in providing cross-indexing and search features for specialized but comprehensive **citation** indices of full documents. The indexing technologies used by such systems, as well as the general-purpose algorithms such as *Google PageRank* [BP98] and *HITS* [KL99], have several advantages: They use a *simple conceptual model* of document webs. They require little specialized knowledge to use, but organize and present hits in a way that allows a knowledgeable user to select relevant hits and build a collection of interrelated documents quickly. They are extremely popular,

encouraging users to submit sites to be archived and corrections to citations, annotations, links, and other content. Finally, some of their content can be automatically maintained.

Despite these benefits, systems such as *ResearchIndex* have limitations that hinder their direct application to IR from bioinformatics repositories:

- **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.
- **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.
- **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 bioinformatics experiment.

1.2 Objectives and Hypothesis

How can we achieve the appropriate balance of generality and selectivity? How can we represent inferred relationships among data entities and programs, and explain them to the user? Our thesis is:

> *Probabilistic representation, learning, and reasoning are appropriate tools for providing domain-specific collaborative filtering capability to users of a scientific computing repository, such as one containing bioinformatics data, metadata, experimental documentation, and source codes.*

Toward this end, we are developing *DESCRIBER*, a research index for consolidated repositories of **computational genomics resources**, along with machine learning and probabilistic reasoning algorithms to refine its data models and implement collaborative filtering. 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. Our project shall also directly advance gene expression modeling and intelligent, search-driven reuse in distributed software libraries.

2 CF IN COMPUTATIONAL SCIENCES

2.1 Collaborative Filtering Objectives

We seek to take existing ontologies and minimum information standards for computational genomics and create a refined and elaborated data model for decision support in retrieving data, metadata, and source codes to serve researchers. A typical collaborative filtering scenario using a domain-specific research index or portal is depicted in **Error! Reference source not found.** 1. We now survey background material briefly to explain this scenario, then discuss the methodological basis of our research: development of learning and inference components that take records of use cases and queries (from web server logs and forms) and produce decision support models for the CF performance element.

As a motivating example of a computational genomics experiments, we use gene expression modeling from microarray data. DNA hybridization *microarrays*, also referred to as *gene chips*, are experimental tools in the life sciences that make it possible to model interrelationships among genes, which encode instructions for production of proteins including the *transcription factors* of other genes. Microarrays simultaneously measure the expression level of thousands of genes to provide a "snapshot" of protein production processes in the cell. Computational biologists use them in order to compare snapshots taken from organisms under a control condition and an alternative (e.g., *pathogenic*) condition. A microarray is typically a glass or plastic slide, upon which DNA molecules are attached at up to tens of thousands of fixed locations, or *spots*. Microarray data (and source code for programs that operate upon them) proliferate rapidly due to recent availability of chip makers and scanners.

A major challenge in bioinformatics is to discover gene/protein interactions and key features of a cellular system by analyzing these snapshots. Our recent projects in computational genomics focus on the problem of automatically extracting gene regulatory dependencies from microarray data, with the ultimate goal of building simulation models of an organism under external conditions such as temperature, cell cycle timing (in the yeast cell), photoperiod (in plants), etc. Genomes of model organisms, such as *S. cerevisiae* (yeast), *A. thaliana* (mouse ear cress or *weed)*, *O. sativa* (rice), *C. elegans* (nematode worm), and *D. melanogaster* (fruit fly), have been fully sequenced. These have also been annotated with the *promoter* regions that contain binding sites of *transcription factors* that regulate gene

expression. Public repositories of microarray data such as the *Saccaromyces* Genome Database (SGD) for yeast have been used to develop a comprehensive catalog of genes that meet analytical criteria for certain characteristics of interest, such as *cell cycle regulation* in yeast. We are using SGD data and a synthesis of existing and new algorithms for learning Bayesian networks from data to build robust models of regulatory relationships among genes from this catalog. Most data resources we plan to use in developing *DESCRIBER* are in the public domain, while some are part of collaborative work with the UK *myGrid* project (Goble).

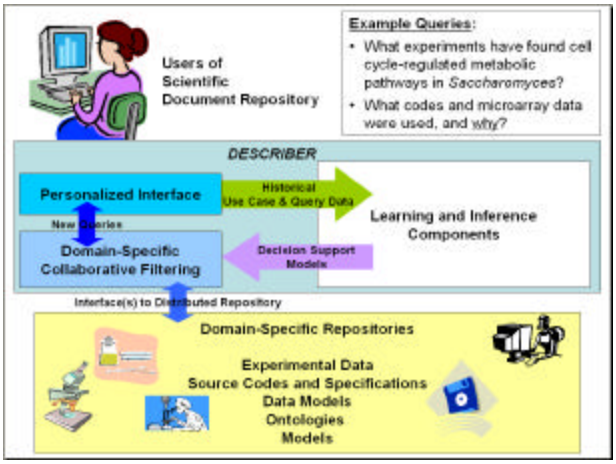

Figure 1. Design overview of DESCRIBER

The next two figures depict our design for *DESCRIBER*. Figure 2 is the block diagram for the overall system, while Figure 3 elaborates Module 1 as shown in the lower left hand corner of Figure 3. Our current and continuing research focuses on algorithms that perform the learning, validation, and change of representation (inductive bias) denoted by Modules 2 and 4. We choose probabilistic relational models as a representation because they can express constraints (cf. Figure 1) and capture uncertainty about relations and entities. We hypothesize that this will provide more flexible generalization over use cases. We have recently developed a system for Bayesian network structure learning that improves upon the *K2* [CH92] and *Sparse Candidate* [FLNP00] algorithms by using combinatorial optimization (by a genetic algorithm) to find good topological orderings of variables. Similar optimization wrappers have been used to adapt problem representation in supervised inductive learning for classification, using decision trees and instance-based learning.

Other relevant work includes *BioIR*, a digital library for bioinformatics and medical informatics whose content is much broader than that of this test bed for genome analysis. *BioIR* emphasizes phrase browsing and cross-indexing of text and data repositories rather than experimental metadata and source codes. Other systems such as *CANIS*, *SPIDER*, and *OBIWAN* also address intelligent search and IR from bioinformatics digital libraries, emphasizing categorization of text documents. We view the technologies in these systems as complementary and orthogonal to our work because of this chief difference.

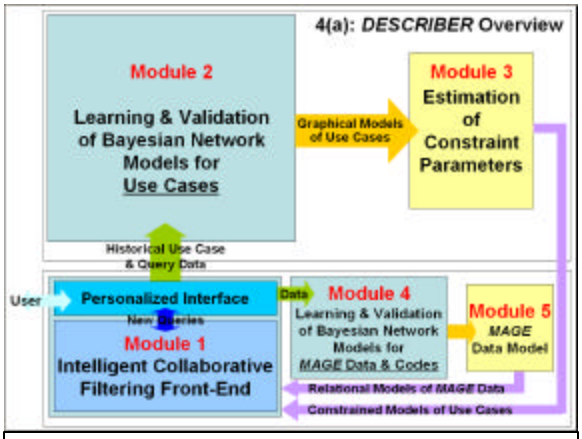

Figure 2. DESCRIBER system

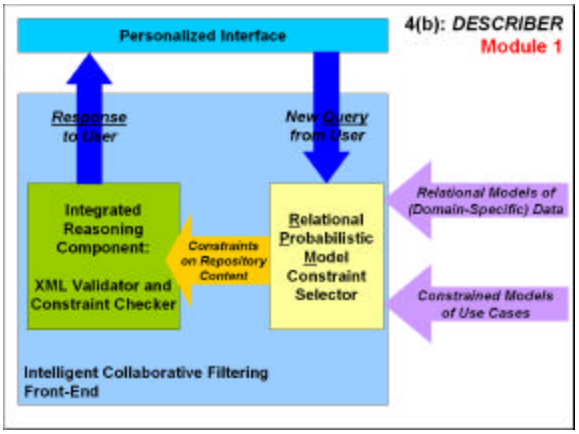

Figure 3. Collaborative filtering component of DESCRIBER

3 LEARNING BN STRUCTURE

3.1 Classifier System for Learning BN Structure

Learning the structure, or causal dependencies, of a graphical model of probability such as a Bayesian network (BN) is often a first step in reasoning under uncertainty. In many machine learning applications, it is therefore referred to as a method of *causal discovery* [PV91]. Finding the optimal structure of a BN from data has been shown to be *NP*-hard [HGC95], even without considering latent (unobserved) or irrelevant (extraneous) variables. Therefore, greedy *score-based* algorithms

[FG98] have been developed to provide more efficient structure learning at an accuracy tradeoff. In this paper we examine a general shortcoming of greedy structure learning – sensitivity to variable ordering – and develop a genetic algorithm to mitigate this problem by searching the permutation space of variables [HH98] using a probabilistic inference criterion as the fitness function.

We make the case in this paper that the probabilistic inference performance element, **in the absence of a known gold standard network** or any explicit constraints, can provide the feedback needed to search for a good ordering. We then derive a heuristic based on validation by inference (exact inference [LS88, Ne90] for small networks, approximate inference by stochastic sampling [CD00] for larger ones). Our primary objective is inferential accuracy *using* the learned structure.

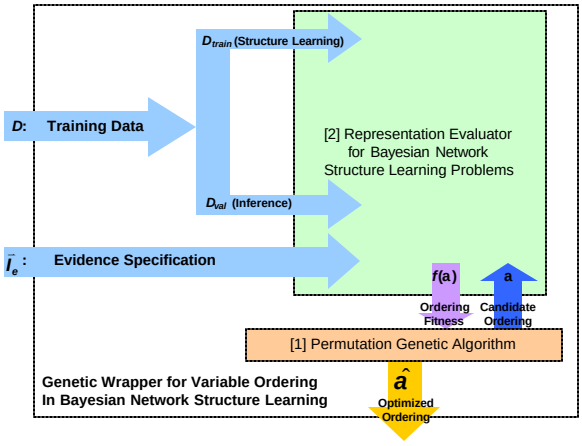

Figure 4. System Design Overview.

Toward this end, we have developed a *genetic wrapper*, similar to a classifier system [BGH89], to search the space of variable orderings in score-based structure learning. This wrapper adapts a composite fitness measure used in other wrappers based upon best-first search [KJ97] and automatically tunes parameters of the learning system [HL99] such as the ordering of input variables. We present the system shown in Figure 1, a genetic algorithm-based wrapper [CS96, RPG+98, Hs03], and show how it provides a parallel stochastic search mechanism for inferential loss-minimizing variable orderings. We demonstrate that, used in tandem with *K2*, it produces structures whose loss under importance sampling is nearly as low as any found by exhaustive enumeration of orderings. Finally, we discuss how this wrapper provides a flexible method for tuning *representation biases* [Mi97] in Bayesian network structure learning using different fitness criteria.

Consider a typical probabilistic reasoning environment, as shown in Figure 2, where structure learning [A] is a first step. The input to this system includes a set D of training data vectors $\mathbf{x} = (x_1, \ldots, x_n)$ each containing n variables. If the structure learning algorithm is greedy, an ordering **a** on the variables may also be given as input. The structure learning component of this system produces a graphical model $B = (V, E, \Theta)$ that describes the dependencies among X_i, including the conditional probability functions. The inferential performance element [B] of this system takes B and a new data set D_{test} of vectors drawn from the desired inference space, where only a subvector **E** of $\mathbf{X} = (X_1, \ldots, X_n)$ is observable, and infers the remaining unobserved values $\mathbf{X} \setminus \mathbf{E}$. We denote the indicator bit vector for membership in **E** by $\mathbf{I}_e$. The performance criterion f is the additive inverse of the (inferential or utility) loss of [B].

4 CONTINUING WORK

Our current research focuses on structure learning of relational models by adapting traditional score-based search algorithms for flat graphical models [Pe03] *and* constrain-based structure search over hierarchical models.

Entity and reference slot uncertainty present new challenges to PRM structure learning. Three of the questions that we are looking into are:

1. *How much relational data is needed?* How can we estimate the sample complexity of PRMs under specified assumptions about entity existence and reference slot distributions?
2. *What constraint-based approaches can be usedf?* Learning reference slot and entity structure in PRMs presents a task beyond flat structure learning.
3. *Can this cut down on the amount of data to learn the low-level model (versus the flat version)?* How can we establish and test sufficient conditions for conditional independenc, and context-specific independence, in PRMs?

5 References

[BGH89] L. B. Booker, D. E. Goldberg, and J. H. Holland. Classifier Systems and Genetic Algorithms. *Artificial Intelligence*, 40:235-282, 1989.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1-7):107-117.

[CD00] J. Cheng and M. J. Druzdzel. AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. Journal of Artificial Intelligence Research (JAIR), 13:155-188, 2000.

[CH92] G. F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, **9**(4):309-347, 1992.

[CS96] K. J. Cherkauer and J. W. Shavlik. Growing Simpler Decision Trees to Facilitiate Knowledge Discovery. In *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, August, 1996.

[FG98] N. Friedman and M. Goldszmidt. *Learning Bayesian Networks From Data*. Tutorial, American

National Conference on Artificial Intelligence (AAAI-98), Madison, WI. AAAI Press, San Mateo, CA, 1998.

[FLNP00] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, Using Bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, ACM-SIGACT, April 2000.

[HGC95] D. Heckerman, D. Geiger, and D. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197-243, Kluwer, 1995.

[HH98] R. L. Haupt and S. E. Haupt. *Practical Genetic Algorithms*. Wiley-Interscience, New York, NY, 1998.

[HL99] G. Harik and F. Lobo. *A parameter-less genetic algorithm.* Illinois Genetic Algorithms Laboratory technical report 99009, 1999.

[Hs03] W. H. Hsu. Control of Inductive Bias in Supervised Learning using Evolutionary Computation: A Wrapper-Based Approach. In J. Wang, *editor*, *Data Mining: Opportunities and Challenges*, p. 27-54. IDEA Group Publishing.

[Kl99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5):604-632.

[KJ97] R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence, Special Issue on Relevance, 97*(1-2):273-324, 1997.

[Mi97] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, 1997.

[LGB99] S. Lawrence, C. L. Giles, and K. Bollacker Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6):67-71.

[LS88] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B* 50, 1988.

[Ne90] R. E. Neapolitan. Probabilistic Reasoning in Expert Systems: Theory and Applications. Wiley-Interscience, New York, NY, 1990.

[Pe03] B. B. Perry. *A Genetic Algorithm for Learning Bayesian Network Adjacency Matrices from Data.* M.S. thesis, Department of Computing and Information Sciences, Kansas State University, 2003.

[PV91] J. Pearl and T. S. Verma, A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. Morgan Kaufmann, San Mateo, CA, 1991.

[RPG+97] M. Raymer, W. Punch, E. Goodman, P. Sanschagrin, and L. Kuhn, Simultaneous Feature Extraction and Selection using a Masking Genetic Algorithm, In *Proceedings of the 7th International Conference on Genetic Algorithms*, pp. 561-567, San Francisco, CA, July, 1997.