Relational Graphical Models for Collaborative Filtering and Recommendation of Computational Workflow Components

William H. Hsu

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

Abstract

This paper describes preliminary work 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: relational graphical models, collaborative information retrieval, explanation of recommendation, user profiling

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*, which 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 to 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 current work in progress and survey results from 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 Figure 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 [ZGSB04].

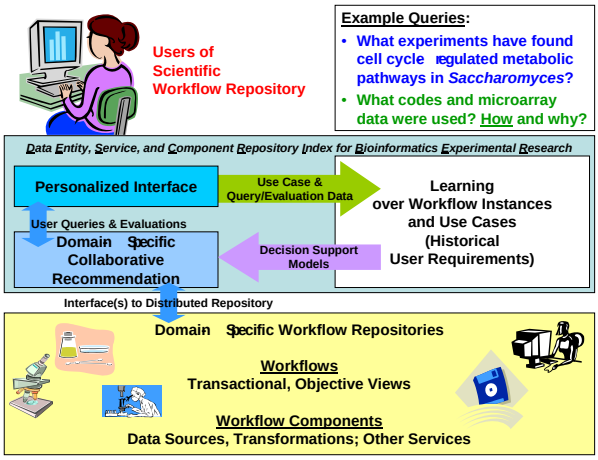

Figure 1. Design overview of *DESCRIBER*.

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 2. 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.

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.

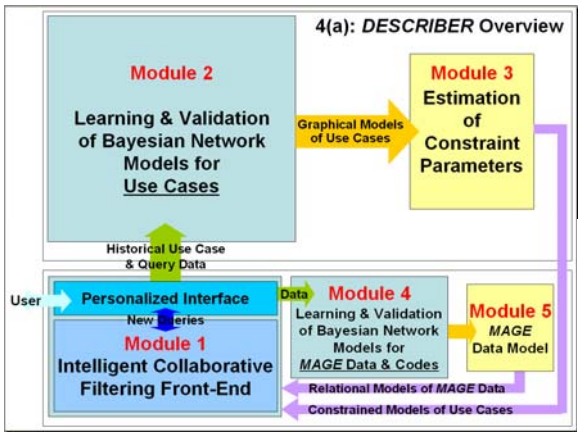

Figure 2. *DESCRIBER* system.

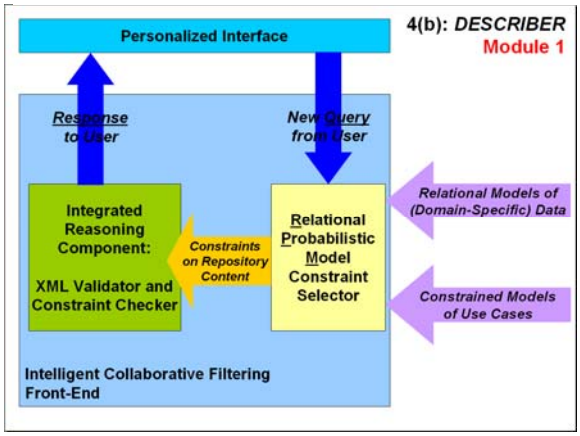

Figure 3. Collaborative filtering component.

3 NEW RGM MODELS

3.1 Managing Decisions under Uncertainty

Decision-theoretic intelligent agents must function under uncertainty and be able to reason and learn about objects and relations in the context of action and utility. This section presents a new relational graphical model (RGM), analogous to the probabilistic relational model (PRM), for representation of decisions under uncertainty. It first analyzes some basic properties of the representation and gives an adaptation of several decision network inference algorithms to these relational decision networks. It then describes some early experimentation with algorithms learning link structure in PRMs, discussing how these can be adapted to learning in decision networks. Finally, it considers the problem of representing dynamic relations in decision networks and sketches an extension of the dynamic PRM representation to include choice and utility.

Uncertainty is a common feature of decision problems for which the decision network or influence diagram is currently one of the most widely-used graphical models. Decision networks represent the state of the world as a set of variables, and model probabilistic dependencies, action, and utility. Though they provide a synthesis of probability and utility theory, decision networks are still unable to compactly represent many real-world domains, a limitation shared by other propositional graphical models such as flat Bayesian networks and dynamic Bayesian networks. Decision domains can contain multiple objects and classes of objects, as well as multiple kinds of relations among them.

Meanwhile, objects, relations, choices, and valuations can change over time. Capturing such a domain in a decision network would require not only an exhaustive representing of all possible objects and relations among them, but also a combinatorially fast-growing space of choices and valuations. This raises two problems. The first one is that the inference using such a dynamic decision network would likely exhibit near-pathological complexity, making the computational cost prohibitive. The second is that reducing the rich structure of domains such as workflow management to very large, "flat" decision networks would make it much more difficult for human beings to comprehend. This paper addresses these two problems by introducing an extension of decision networks that captures the relational structure of some decision domains, and by adapting methods for efficient inference in this representation.

3.2 Extending Decision Networks

The representation we introduce in this paper extends PRMs to decision problems in the same way that the decision networks extend Bayesian networks. We therefore call it the *relational decision network* or RDN. We develop two inference procedures for RDNs: the first based upon the traditional variable elimination algorithm developed by Shenoy and Cowell, the second a more efficient one based upon an adaptive importance sampling-based algorithm.

3.2.1 Probabilistic Relational Models

First-order formalisms that can represent objects and relations, as opposed to just variables have a long history in AI. Recently, significant progress has been made in combining them with a principled treatment of uncertainty. In particular, probablistic relational models, or PRMs, are an extension of Bayesian networks that allows reasoning with classes, objects, and relations.

Probabilistic relational models (PRMs) [GFKT02, SDW03] extend the flat (propositional) representation of the variables and the conditional dependencies among them to an object-relational representation. Before proceeding to discussion the decision network analogues of PRMs, we briefly review the PRM family and the relevant components of a PRM specification. As an example extending the DEC-Asia decision network above, the Patient schema might be used in to represent partial or total patient records, with classes corresponding to information about a patient's pulmonary medical history, smoking history, travel itinerary, and groups of people contacted. The propositional attributes of the medical history include the patient's age, previous contagious pulmonary diseases contracted, and currently extant diseases; the relational attributes might include the patient's membership in a list of quarantine subjects and *links* between patients denoting specific exposure incidents and contexts. Note that some of these are static and some, such as clusters of at-risk destinations and groups of people, may be dynamic relational attributes.

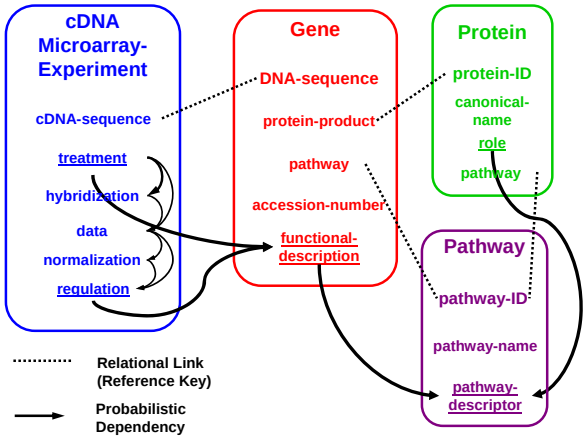

Figure 4. PRM for the *DESCRIBER* domain.

As a further example, Figure 4 depicts a PRM for the domain of computational genomics, particular gene expression modeling from DNA hybridization microarrays. Slot chains can be traced using the reference keys (dotted lines). This PRM contains tables for individual microarrays or gene chips (admitting aggregation of chip objects into classes), putative gene function (where known or hypothesized), putative pathway membership (where known or hypothesized), and protein production (a target aspect of discovered function).

This allows a PRM to be flattened into a large Bayesian network containing ground (propositional) chance nodes, with one variable for every attribute of every object in the relational skeleton of Π and belief functions (usually deterministic) for the aggregation operations. The latter are open-ended in form.

As Getoor *et al.* [GFKT02] and Sanghai *et al.* [SDW03] note, the most general case currently amenable to learning is where an object skeleton is provided and structure and parameter learning problems must be solved in order to specify a distribution over relational attributes. In the epidemiology domain, a PRM might specify a distribution over possible transmission vectors of an infected person (the itinerary, locale of contamination, and set of persons contacted).

3.2.2 Relational Decision Networks

Decision networks are extendible to relational representations using a simple and straightforward synthesis of decision network and PRM specifications.

Thus the relational attributes can include distinguished member *action identifiers* and *outcome identifiers* specifying a representation for equivalence classes of decisions and outcomes. Note that the range of actions may be continuous (e.g., in intelligent control or continuous decision problems) and the range of *utilities* may also be continuous.

When the decision network's object skeleton is not known (i.e., the set of decisions and outcomes is not fully pre-specified), the RDN includes boolean *existence variables* for propositional attributes of the relational tables, and boolean *reference slot variables* for relational attributes.

Inference algorithms that can be used with RDNs include two based on stochastic sampling: Likelihood Weighting and Adaptive Importance Sampling (AIS). For brevity, we refer interested readers to Cheng and Druzdzel [CD00] for detailed descriptions of these algorithms.

A desired joint probability distribution function *P(X)* can be computed using the chain rule for Bayesian networks, given above. The most probable explanation (MPE) is a truth assignment, or more generally, value assignment, to a query $Q = X \setminus E$ with maximal posterior probability given evidence e. Finding the MPE directly using enumeration requires iteration over exponentially many explanations. Instead, a family of exact inference algorithms known as *clique-tree* propagation (also called *join tree* or *junction tree* propagation) is typically used in probabilistic reasoning applications. Although exact inference is important in that it provides the only completely accurate baseline for the fitness function f, the problem for general BNs is ***#P***-complete (thus, deciding whether a particular truth instantiation is the MPE is NP-complete.

Approximate inference refers to approximation of the posterior probabilities given evidence. One stochastic approximation method called importance sampling [CD00] estimates the evidence marginal by sampling query node instantiations.

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

[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*, **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.

[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.

[GFKT02] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning Probabilistic Models of Link Structure. *Journal of Machine Learning Research*, 2002.

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

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

[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.

[SDW03] S. Sanghai, D. Weld, and P. Domingos. Dynamic Probabilistic Relational Models. In *Proceedings of IJCAI-2003*.

[ZGSB04] J. Zhao, C. Goble, R. Stevens, and S. Bechhofer. Semantically Linking and Browsing Provenance Logs for E-science. In *LNCS 3226, Semantics of a Networked World: Semantics for Grid Databases*. Berlin: Springer-Verlag.