Relational Graphical Models of Computational Workflows for Data Mining

William H. Hsu

Department of Computing and Information Sciences, Kansas State University
Manhattan, KS 66506-2302
http://www.kddresearch.org, bhsu@cis.ksu.edu

Collaborative recommendation 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. Recent systems such as *CiteSeer* [1] have succeeded in providing some specialized but comprehensive indices of full documents, but the kind of probabilistic models used in such indexing do not extend easily to information Grid databases and computational Grid workflows. The collection of user data from Grid portals [4] provides a test bed for the underlying IR technology, including learning and inference systems. To model workflows created using the *TAVERNA* editor [3] and *SCUFL* description language, the *DESCRIBER* system, shown in Figure 1, applies score-based structure learning algorithms, including Bayesian model selection and greedy search (cf. the K2 algorithm) adapted to relational graphical models. Figure 2 illustrates how the decision support front-end of *DESCRIBER* interacts with modules that learn and reason using probabilistic relational models,. The purpose is to discover interrelationships among, and thereby recommend, components used in workflows developed by other Grid users.

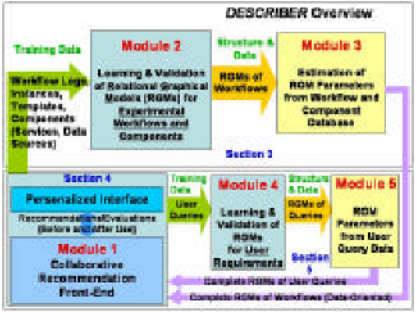

Fig. 1. *DESCRIBER* overview

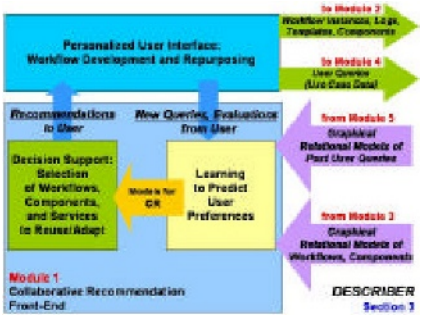

Fig. 2. Recommender component

In this work, we adapt score-based structure learning algorithms for discovering *relational graphical models* for domain-specific collaborative recommendation from scientific data and web service repositories, as well as use cases for software and data sets retrieved from them. Statistical evaluation is performed using validation user data over a larger set of *TAVERNA* workflows built using the same services. The object-relational representation of these transactional workflow models include probabilistic

M. Bouzeghoub et al. (Eds.): ICSNW 2004, LNCS 3226, pp. 309–310, 2004.

graphical models, such as *Bayesian networks* and *decision networks*, that have recently been applied to a wide variety of problems in intelligent information retrieval, information extraction, and link analysis. [2] Our approach combines score-based algorithms for learning the structure of relational probabilistic models with existing techniques for constructing relational models of metadata about computational Grid services (including data sources, functions, and nested workflows).

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. A major challenge in bioinformatics is to discover gene/protein interactions and key features of a cellular system by analyzing microarray scans. Figure 3 depicts an example of a relational graphical model where each transactional workflow is mapped to one instance of an objectrelational schema as shown. Figure 4 shows another example, adapting the toy decision support network *DEC-Asia* to a relational extension of a decision network.

Our recent projects in computational genomics focus on the problem of automatically extracting gene regulatory dependencies from gene expression data sources such as cDNA microarrays. Most data resources we plan to use in developing *DESCRIBER* are in the public domain, while some are *TAVERNA* workflows [3] developed by *myGrid* project team [4].

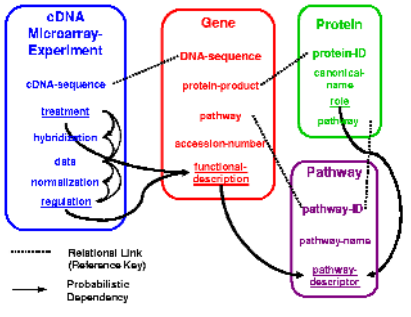

Fig. 3. Example relational model for bioinformatics domain

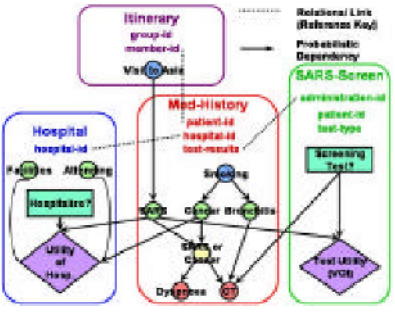

Fig. 4. Relational decision network for diagnostic utility (SARS-Screen)

References

1. Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6):67-71.
2. Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2002). Learning probabilistic models of link structure. *J. of Artificial Intelligence Research*, **3**(12), 679-707.
3. Oinn, T., Marvin, D., & Rice, P. (2004). *TAVERNA* workflow editor for myGrid. Available from: http://taverna.sourceforge.net .
4. Stevens, R. D., Robinson, A. J., & Goble, C. A. (2003). myGrid: personalized bioinformatics on the information Grid (ISMB-2003). *Bioinformatics* **19**:302-305.