Increasing Representational Power and Scaling Reasoning in Probabilistic Databases

Amol Deshpande (amol@cs.umd.edu)
Department of Computer Science and UMIACS
University of Maryland, College Park, MD, USA

ABSTRACT

Increasing numbers of real-world application domains are generating data that is inherently noisy, incomplete, and probabilistic in nature. Statistical analysis and probabilistic inference, widely used in those domains, often introduce additional layers of uncertainty. Examples include sensor data analysis, data integration and information extraction on the Web, social network analysis, and scientific and biomedical data management. Managing and querying such data requires us to combine the tools and the techniques from a variety of disciplines including databases, first-order logic, and probabilistic reasoning. There has been much work at the intersection of these research areas in recent years. The work on probabilistic databases has made great advances in efficiently executing SQL and inference queries over large-scale uncertain datasets [2, 1]. The research in first-order probabilistic models like probabilistic relational models [5], Markov logic networks [10] etc. (see Getoor and Taskar [6] for a comprehensive overview), and the work on lifted inference [9, 3, 8, 11] has resulted in several techniques for efficiently integrating first-order logic and probabilistic reasoning.

In this talk, I will present some of the foundations of large-scale probabilistic data management, and the challenges in scaling the representational power and the reasoning capabilities of probabilistic databases. I will use the PrDB probabilistic data management system being developed at the University of Maryland as a case study for this purpose [4, 7, 12]. Unlike the other recent work on probabilistic databases, PrDB is designed to represent uncertain data with rich correlation structures, and it uses probabilistic graphical models as the basic representation model. I will discuss how PrDB supports compact specification of uncertainties at different abstraction levels, from "schema-level" uncertainties that apply to entire relations to "tuple-specific" uncertainties that apply to a specific tuple or a specific set of tuples; I will also discuss how this relates to the work on first-order probabilistic models. Query evaluation in PrDB can be formulated as inference in appropriately constructed graphical models, and I will briefly present some of the key novel techniques that we have developed for efficient query evaluation, and their relationship to recent work on efficient lifted inference. I will conclude with a discussion of some of the open research challenges moving forward.

ICDT 2010, March 22–25, 2010, Lausanne, Switzerland.

References

[1] Charu C. Aggarwal, ed. *Managing and Mining Uncertain Data*. Springer, 2009.

[2] Nilesh N. Dalvi, Christopher Ré, and Dan Suciu. Probabilistic databases: diamonds in the dirt. *Communications of the ACM*, 52(7):86–94, 2009.

[3] Rodrigo de Salvo Braz, Eyal Amir, and Dan Roth. Lifted first-order probabilistic inference. In *IJCAI*, 2005.

[4] Amol Deshpande, Lise Getoor, and Prithviraj Sen. *Graphical Models for Uncertain Data*. Managing and Mining Uncertain Data. Charu Aggarwal ed., Springer, 2009.

[5] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *International Joint Conference on Artificial Inteligence*, Stockholm, Sweden, 1999.

[6] Lise Getoor and Ben Taskar, ed. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[7] Bhargav Kanagal and Amol Deshpande. Indexing correlated probabilistic databases. In *SIGMOD*, pages 455–468, 2009.

[8] Brian Milch, Luke Zettlemoyer, Kristian Kersting, Michael Haimes, and Leslie Kaelbling. Lifted probabilistic inference with counting formulas. In *AAAI Conference on Artificial Intelligence*, 2008.

[9] David Poole. First-order probabilistic inference. In *International Joint Conference on Artificial Inteligence*, 2003.

[10] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

[11] Prithviraj Sen, Amol Deshpande, and Lise Getoor. Bisimulation-based approximate lifted inference. In *UAI*, 2009.

[12] Prithviraj Sen, Amol Deshpande, and Lise Getoor. PrDB: managing and exploiting rich correlations in probabilistic databases. *VLDB Journal*, 18(5):1065–1090, 2009.