Prospects and Challenges for Multi-Relational Data Mining

Pedro Domingos
Dept. of Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195-2350, U.S.A.
pedrod@cs.washington.edu

ABSTRACT

This short paper argues that multi-relational data mining has a key role to play in the growth of KDD, and briefly surveys some of the main drivers, research problems, and opportunities in this emerging field.

1. WHAT DRIVES MRDM

Multi-relational data mining (MRDM) is a field whose time has come. The fact that most data mining algorithms operate on a single relation or table, while most real-world databases store information in multiple tables, is of course not new, and MRDM has a precursor going back over a decade in the field of inductive logic programming (ILP) [12]. However, until recently, three factors were responsible for the slow growth of this area: the limited scalability of multi-relational algorithms, their inability to explicitly handle noise and uncertainty, and a perceived lack of "killer apps." As the other articles in this issue of *SIGKDD Explorations* illustrate, each of these is bottlenecks is beginning to disappear [3; 8; 22]. As a result, MRDM is entering a period of rapid expansion that will increasingly place it at the center of the KDD enterprise and its real-world impact.

We look at scalability, uncertainty and applications in turn, with a particular focus on the latter, since they are the main drivers of the recent growth in MRDM.

Enabling advances of the last few years on the scalability front include: the identification of restrictions of the general ILP problem that allow for computationally efficient solutions (e.g., [21]); the extension of frequent-set algorithms to the multi-relational case (e.g., [25]); and the increasing number and variety of algorithms for extracting patterns and other information from very large graphs (e.g.,[23]). Much remains to be done, including extending to the relational setting scaling-up techniques that have been successful in propositional (single-table) mining, and improving the interface between MRDM and the relational database technology that underlies it (see below).

One of the most significant developments of recent years has been the convergence of ILP and probabilistic reasoning and learning (Bayesian networks, in particular, and their extensions and specializations) [16]. ILP brings the ability to handle multiple relations; probabilistic methods bring the ability to handle uncertainty explicitly and systematically. To a significant extent, MRDM can be viewed as the product of this convergence. The growing stable of approaches that exhibit it includes probabilistic relational models [14], stochastic logic programs [20], Bayesian logic programs [18], relational Markov models [1], first-order Bayesian classifiers [13], and others. One challenge for future research is to systematize this plethora of approaches, elucidating the relations between them and their strengths and limitations.

Perhaps most exciting is the growth of several major KDD applications that are intrinsically relational, and call for algorithms that exploit this. They include:

The World-Wide Web. The WWW is a massive graph, and taking the interconnections between pages into account leads to better results in Web search [4; 19], Web page classification [7], etc. The success of Google's PageRank algorithm [4] and of Kleinberg's "hubs and authorities" algorithm [19] have led to an explosion of research on mining the link structure of the Web [6]. MRDM provides techniques for doing this, and for combining the "links to" relation with the text content of the page, its HTML structure, etc. Further, as the Semantic Web [2] develops as a machine-readable alternative to the WWW, the richness and variety of the relational information available for mining, and the potential benefits of mining it, grow in parallel.

Counter-terrorism. The attacks of September 11 have led to a greatly increased awareness of terrorist threats, and one consequence of this has been a concerted push into research relevant to counter-terrorism, particularly in the United States. Data mining is a central (and sometimes controversial) part of this effort. In particular, the link analysis techniques traditionally used by intelligence and law enforcement agencies can be naturally formalized, automated and extended in the context of MRDM [17], and several major research projects in this direction are now under way.[1]

Viral marketing. Traditional forms of marketing are increasingly being supplemented (or replaced) by *viral marketing*, which explicitly takes network effects into account and seeks to leverage word-of-mouth among consumers [11]. Particularly in the Internet sector, the future of companies increasingly depends on the success of their viral mar-

[1]See http://www.darpa.mil/iao/EELD.htm for an example.

keting plans. Devising them properly involves modeling the interactions among consumers and between consumer attributes and product characteristics, which is fundamentally an MRDM task.

Social networks. More generally, the field of social network analysis [26], which focuses on modeling the relations between social actors, is relevant to all the above applications, and is a prime client for MRDM. The quantity of data available for building social network models is growing rapidly, and much progress can be expected if MRDM provides the necessary tools.

Computational biology. ILP has had some notable successes in molecular biology applications (e.g., predicting carcinogenesis [24]). Most importantly, biology is fast progressing from the study of individual genes and proteins to the study of how they interact in the living cell. Modeling these interactions and their relation to the observable outcomes, and to the properties of the participating molecules, is a vast area for application of MRDM.

Information extraction. The more text-based information becomes available online, the greater the potential impact of better information extraction (IE) techniques. IE is a fundamentally relational problem, not only because its goal is to populate databases from text, but because doing this successfully requires understanding the explicit and implicit relations among the entities the text is about [5].

Ubiquitous computing. The growth in the number and variety of devices that continuously produce data is a potential bonanza for KDD. Mining what goes on in a ubiquitous computing environment requires taking multiple types of objects and relations into account.

This list is by no means exhaustive, of course, but rather a representative sample.

2. RESEARCH CHALLENGES

The wealth of applications described in the previous section means that the demand for better MRDM solutions is high. However, many challenging research problems must be addressed if this demand is to be fully met. Issues that cut across all of the applications described, and where progress will consequently have the broadest impact, include:

Learning from networks of examples. Most ILP research has focused on problems where individual examples have relational structure, but examples are still independent of each other. For instance, an example might be a molecule, with the bonds between atoms as the relational structure, and the task being to predict whether the molecule is a carcinogen. However, arguably the most interesting and challenging problem in MRDM is that of dependences between examples. For example, molecules do not act independently in the cell; rather, they participate in complex chains of reactions whose outcomes we are interested in. Likewise, Web pages are best classified by taking into account the topics of pages in their graph neighborhood, and consumers' buying decisions are often best predicted by taking into account the influence of their friends and acquaintances. The space of models that assume example independence is a minuscule fraction of the space of all possible models. Exploring the latter is likely to prove a very rich area for research. A productive direction may be to focus on limited dependencies, as Bayesian networks do for the propositional case.

Learning from time-changing relational data. Many relational phenomena of interest take place over time (e.g., a shopper cruising through an e-commerce site, a terrorist group preparing an attack, the response of an organism's immune system to an infection). Temporal phenomena pose some of the most intriguing problems for MRDM. How can we model the appearance, disappearance, assembly and disassembly of objects, and the making and breaking of relations among them? The real world is obviously rife with such events, but they have not been addressed in traditional KDD. Further, learning and inference over time (e.g., in dynamic Bayesian networks) are known to be particularly difficult; performing them in relational domains is likely to pose both new challenges and new opportunities.

Integration with traditional KDD. An oft-heard criticism of MRDM is that the problems it addresses could in principle be solved by traditional learners, if suitably formulated. This is akin to saying that high-level programming languages like Java and C++ are not very useful, because the same programs could be written in assembly code. In finite domains, all multi-relational problems can indeed be reduced to equivalent propositional ones, but this does not negate the compact representation, modularity and scaffolding for learning and inference that relational approaches provide. Nevertheless, the emphasis in the MRDM literature has often been on the contrasts between relational and propositional approaches, which is understandable in a new field, but ultimately counterproductive. Emphasizing the continuity between the two will make widespread adoption of MRDM easier. Further, MRDM should build on the extensive research already done on propositional methods, by making it easy to plug them into MRDM algorithms. For example, probabilistic relational models should in principle allow any classifier as a node model, and Bayesian logic programs should in principle allow any classifier as a combining rule.

Integration with databases. Traditionally, a single table for mining is extracted manually from a multi-relational database. Finding the best way to do this is often quite difficult, and can consume a large fraction of a KDD project's time. One of the great potential benefits of MRDM is the ability to automate this process to a significant extent. Fulfiling this potential requires solving the significant efficiency problems that arise when attempting to do data mining directly from a relational database, as opposed to from a single pre-extracted flat file. Although the study of this problem began with traditional KDD (e.g., [15]), many challenging new issues arise in the MRDM context. For example, it would be desirable to support automated *ad hoc* search over join paths, effectively assembling many different tables for learner modules on the fly, but a traditional RDBMS is too slow for this. Extending streaming-data and approximate querying approaches (e.g., [10]) to the MRDM domain is likely to be part of the answer. Another aspect is that, by look-

ing directly at "unpacked" databases, MRDM is closer to the "real world" of programmers formulating SQL queries than traditional KDD. This means it has the potential for wider use than the latter, but only if we address the problem of expressing MRDM algorithms and operations in terms that are intuitive to SQL programmers and OLAP users.

Learning from multiple sources of information. Data for MRDM often comes from multiple sources. They can be of many different types (e.g., relational databases, plain text, HTML, XML, audio, video, sensors, etc.). Even when they are of the same type, different sources often use different representations for the same entities, and can be of widely varying quality. We would like MRDM systems to be able to range autonomously over all sources relevant to their task, perhaps even discovering them on their own. Doing this requires figuring out automatically how the source contents map to the objects and relations of interest to the system. Giving multi-relational miners the ability to use meta-data and self-describing data is one step toward this. Another is using machine learning techniques to generalize from known mappings to unknown ones [9]. A third one is developing methods for automatically gauging the quality of information sources, by evaluating their responses to queries with known answers, by mapping them to sources of known quality, etc.

Using domain knowledge. One of the most attractive features of ILP is its ability to naturally incorporate domain knowledge in the form of Horn rules and other statements in first-order logic. The incorporation of domain knowledge is essential to the success of any KDD project, and the more it can be done automatically, the lower the cost of the project, and the less time it will take. Unfortunately, domain knowledge in purely logical form is very hard to come by. Application domains are messy and complex, and so are experts' minds. If MRDM can take advantage of uncertain reasoning to absorb domain knowledge in less polished forms, this may greatly increase its range of applicability, and its success in the applications it tackles.

Making the results of research accessible. One of the bottlenecks preventing the wider use of MRDM is the fact that relational algorithms tend to be much more involved and difficult to understand than propositional ones. This is doubtless partly the result of relational problems being intrinsically more complex. Nevertheless, MRDM researchers must strive to present their algorithms in the most accessible form possible, to devise frameworks and simplifying ideas that reduce the "delta" required to understand a new algorithm, and to continue and expand the practice of making easily-used software available.[2] Otherwise the uptake of MRDM by students and practitioners will necessarily be slow and uneven.

In the long term, KDD can only be said to have truly succeeded if mining data from multiple relations—as it is stored in databases—and combining it with knowledge expressed in rich and noise-tolerant languages becomes far easier than it is today. This places MRDM at the heart of KDD, and means that the stakes are high for researchers working in this area. The next few years should be an exciting time.

[2]See http://www-ai.ijs.si/∼ilpnet2/systems for a list of publicly-available ILP systems.

3. REFERENCES

[1] C. Anderson, P. Domingos, and D. Weld. Relational Markov models and their application to adaptive Web navigation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 143–152, Edmonton, Canada, 2002. ACM Press.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001.

[3] H. Blockeel and M. Sebag. Scalability and efficiency in multi-relational data mining. *SIGKDD Explorations*, 5, 2003. This volume.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998. Elsevier.

[5] C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18(4):65–79, 1997.

[6] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Francisco, CA, 2003.

[7] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 307–318, Seattle, WA, 1998. ACM Press.

[8] L. De Raedt and K. Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5, 2003. This volume.

[9] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pages 509–520, Santa Barbara, CA, 2001. ACM Press.

[10] P. Domingos and G. Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 71–80, Boston, MA, 2000. ACM Press.

[11] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, San Francisco, CA, 2001. ACM Press.

[12] S. Dzeroski. Inductive logic programming and knowledge discovery in databases. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 117–152. AAAI Press, Menlo Park, CA, 1996.

[13] P. Flach and N. Lachiche. 1BC: A first-order Bayesian classifier. In *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP'99)*, pages 92–103, Bled, Slovenia, 1999. Springer.

[14] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1300–1307, Stockholm, Sweden, 1999. Morgan Kaufmann.

[15] G. Graefe, U. Fayyad, and S. Chaudhuri. On the efficient gathering of sufficient statistics for classification from large SQL databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 204–208, New York, NY, 1998. AAAI Press.

[16] D. Heckerman. Bayesian networks for knowledge discovery. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 273–305. AAAI Press, Menlo Park, CA, 1996.

[17] D. Jensen and H. Goldberg, editors. *Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*. AAAI Press, Orlando, FL, 1998.

[18] K. Kersting and L. De Raedt. Basic principles of learning Bayesian logic programs. Technical Report 174, Institute for Computer Science, University of Freiburg, Freiburg, Germany, 2002.

[19] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, Baltimore, MD, 1998. ACM Press.

[20] S. Muggleton. Stochastic logic programs. In L. de Raedt, editor, *Advances in Inductive Logic Programming*, pages 254–264. IOS Press, Amsterdam, Netherlands, 1996.

[21] C. Nédellec, C. Rouveirol, H. Adé, F. Bergadano, and B. Tausend. Declarative bias in ILP. In L. de Raedt, editor, *Advances in Inductive Logic Programming*, pages 82–103. IOS Press, Amsterdam, Netherlands, 1996.

[22] D. Page and M. Craven. Biological applications of multi-relational data mining. *SIGKDD Explorations*, 5, 2003. This volume.

[23] C. R. Palmer, P. B. Gibbons, and C. Faloutsos. ANF: A fast and scalable tool for data mining in massive graphs. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 81–90, Edmonton, Canada, 2002. ACM Press.

[24] A. Srinivasan, R. D. King, S. H. Muggleton, and M. Sternberg. The predictive toxicology evaluation challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 1–6, Tokyo, Japan, 1997. Morgan Kaufmann.

[25] S. Tsur, J. D. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. Query flocks: A generalization of association-rule mining. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 1–12, Seattle, WA, 1998. ACM Press.

[26] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.