International Journal on Artificial Intelligence Tools
Vol. 15, No. 6 (2006) 867–874

World Scientific
www.worldscientific.com

ADVANCES IN KNOWLEDGE ACQUISITION AND REPRESENTATION

LAWRENCE B. HOLDER
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX 76019, USA
holder@cse.uta.edu

ZDRAVKO MARKOV
Department of Computer Science
Central Connecticut State University
New Britain, CT 06050, USA
markovz@ccsu.edu

INGRID RUSSELL
Department of Computer Science
University of Hartford
West Hartford, CT 06117, USA
irussell@hartford.edu

The articles in this special issue represent advances in several areas of knowledge acquisition and knowledge representation. In this article we attempt to place these advances in the context of a fundamental challenge in AI; namely, the automated acquisition of knowledge from data and the representation of this knowledge to support understanding and reasoning. We observe that while this work does indeed advance the field in important areas, the need exists to integrate these components into an end-to-end system and begin to extract general methodologies for this challenge. At the heart of this integration is the need for performance feedback throughout the process to guide the selection of alternative methods, the support for human interaction in the process, and the definition of general metrics and testbeds to evaluate progress.

Keywords: Knowledge acquisition; knowledge representation; automated reasoning.

1. Introduction

One of the fundamental challenges of AI is the automated acquisition of knowledge from data along with the representation of this knowledge to support reasoning and understanding. The articles in this issue span the spectrum of this challenge from feature selection for learning to new formalisms of knowledge representation to support for specific reasoning capabilities. In this article we attempt to place these advancements in the context of the aforementioned fundamental challenge and discuss future directions in the field.

2. Knowledge Acquisition

The field of knowledge acquisition has been heavily studied from initial automated machine learning approaches (e.g., perceptrons[20]), to semi-automated knowledge engineering,[2] to more robust automated approaches (e.g., neural networks,[21] symbolic rule learning,[18]) and most recently to proven, practical methods,[23] successful applications[16] and mature theoretical frameworks.[11] Here we focus on three basic components of successful knowledge acquisition. First, the ability to acquire knowledge requires the ability to focus attention on the relevant features of the problem, i.e., feature selection. Second, a staple of any knowledge acquisition approach is the ability to learn rules relating the selected features to the categories of interest. Third, in today's fast-paced, dynamic environment, the abilities to acquire knowledge incrementally and detect when knowledge is changing are crucial. In addition to these basic components, we also consider the use of learning in performing a collaborative assessment of knowledge, i.e., collaborative filtering, and the novel approach of using emergent behavior of multi-agent systems as a method for acquiring new knowledge.

The first step in any knowledge acquisition task is the collection of data and selection of the features to represent the data. Numerous feature selection techniques have been proposed in the literature,[15] and we will not detail them here. However, their importance continues to be stressed, most recently with the advent of support vector machines[3] (SVMs) and their efficient utilization of higher-dimensional features for simple, regression-like learning methods. In fact, Li *et al.*'s article in this issue describes a least-squares regression approach to feature selection for use with support vector machines to improve learning for polyp detection to help identify colon cancer. SVMs represent one of the most advanced methods for learning, and feature selection is crucial to the success of this approach.

While SVMs are among the best learning methods for accurate classification, they currently suffer from the disadvantage of not producing easily-interpretable knowledge for human consumption and automated reasoning systems. More traditional rule learning systems overcome this disadvantage by using a symbolic representation for learned knowledge that still achieves accurate knowledge acquisition, but supports the comprehension of the acquired knowledge and integration of this knowledge into more traditional rule-based reasoning systems. The ability to interpret learning knowledge is important in domains in which an explanation of actions taken by an automated system is necessary for validation of the actions and improving the understanding of practitioners in the domain. For example, the ability to detect anomalies in order to identify potential security intrusions necessitates a symbolic rule-learning approach both in terms of representing the relevant features of the task and explaining the assessment of level of security risk. Tandon and Chan's article in this issue describes a similar approach in learning symbolic rules representing the arguments to computer commands in order to detect anomalies that may suggest attempts at intrusion. They note that approaches relying on purely quantitative measures or just the symbolic commands can be circumvented by experienced intruders.

As with methods for computer intrusion detection, many knowledge acquisition methods are being motivated by the heightened attention to security tasks like counter-terrorism, emergency response and border monitoring. One of the common characteristics of security-related tasks is the dynamic nature of the data, which typically arrives via multiple data streams in real time, and the dynamic nature of the knowledge (e.g., patterns of behavior), which undergo constant change. The ability to handle this dynamic environment is a challenge to current knowledge acquisition methods. Coble *et al.*'s article in this issue describes one approach when the streaming data describes a changing network of nodes and links and the task is to learn patterns in the network and detect when these patterns change over time. Their approach makes efficient use of the learning capability so as to not have to learn on the entire, cumulative data, but only on the latest increments, yet still react to changes in the central tendency of the patterns over time. They successfully demonstrate their approach on a simulated counter-terrorism domain.

Similar to the challenge of acquiring knowledge from multiple data streams is the task of assimilating such knowledge from multiple sources. For example, in counter-terrorism domains there are typically multiple, possibly conflict sources of knowledge about some phenomena (e.g., level of threat of a person or activity). An abstract approach to combining learned knowledge is represented by the recent advances in ensemble learning,[5] however, these approaches assign weights on the knowledge based on performance, rather than on what humans rely more heavily on, that is, the notion of trustworthiness of the source of the knowledge. Trust in the source is an important component to any intelligence analysis task. However, it is becoming increasingly important in perhaps less critical activities related to assessment of items in non-security-related domains. For example, the collaborative filtering techniques[14] used by many on-line vendors utilize customer profiles and reviews to help assess items of interest to those with similar profiles. However, customers with similar profiles do not always share the level of expertise of other customers with perhaps less-similar profiles. However, we value this expertise and desire to include it in the assessment of items of interest. Donavan and Smyth's article in this issue describes an approach to including trustworthiness, in addition to profile similarity, in the assessment of items of interest to the user. Their results show that the inclusion of trust improves the accuracy of assessments compared to similarity alone. Incorporation of trust information remains an important open issue in other methods for knowledge acquisition.

Further generalizing this notion of knowledge acquisition from multiple sources takes us to a truly cutting-edge approach to acquiring new knowledge through the emergent behavior of networks of individuals (e.g., multi-agent systems,[10] social networks[1]). In such domains it is difficult to represent data as feature vectors, or even as more relational static structures (e.g., graphs). The domain is more-appropriately represented as a dynamic process of collaboration and competition among a community of active agents. Such an approach to knowledge acquisition can identify emerging behaviors such as hubs and authorities[12] and communities.[8] In the case of Kobti *et al.*'s article in this issue, emergent behavior can identify specialized networks, such as kinship

and economic networks, as well has hubs that serve to connect these networks and provide stability to the overall system. Such knowledge provides significant explanatory power for group behaviors and predictive power for determining the group's future. Kobti *et al.* demonstrate this power in their successful analysis of the mysterious behavior of an ancient society of Pueblo Indians.

The diverse sub-areas described in this section lie on the cutting edge of techniques and applications of knowledge acquisition. The entire spectrum of this challenge, from feature selection to emergent knowledge acquisition, represents advances in traditional approaches as well as innovated techniques. While much progress remains to be made in each of these areas, the next challenge is the design of integrated knowledge acquisition systems that draw from best practices and utilize appropriate methods where needed and motivated by the ultimate representation and use of the knowledge to support improved understanding of the domain and automated reasoning to infer new knowledge. Explicit feedback mechanisms should be in place to allow these targeted uses of the knowledge to inform and guide the acquisition of the knowledge and a more fully-automated approach to selecting appropriate knowledge acquisition methods.

3. Knowledge Representation

Knowledge representation and its use for reasoning and understanding complete the grand AI challenge initiated by the knowledge acquisition task. Like knowledge acquisition, knowledge representation and reasoning have a long history in AI research, and even well beyond that into the realm of philosophy and declarative thought. First-order logic has been and continues to be the substrate of most knowledge representations, and the treatment of the topic in this issue is no different. Articles in this issue use first-order logic as the basis of a stochastic logic, of categorical grammars for the coordination problem, and of advanced quantification for the logic of determination of objects. First-order logic is also used as the basis of methods for structural verification of proofs derived from automated theorem proving and the definition of ontologies and their mapping between domains. Yet, alternatives to first-order logic exist, most commonly for specialized applications like natural language processing and information retrieval. In this section we consider these advanced areas of knowledge representation and reasoning and attempt to integrate them into a more coherent, general view of the field.

The coupling of logic and uncertainty has been studied for many years.[22] While several representations have been developed, few combine the representation with practical algorithms for determining the uncertainty values of logical forms in the representation. One promising approach to this practical task is Pless *et al.*'s (this issue) use of the expectation-maximization[4] (EM) technique to determine the certainty values associated with their stochastic logic. Such an approach provides a practical alternative to richer, yet computationally intractable, approaches to stochastic logics, such as Bayesian networks.[17] Another emerging approach that lies between these two and is showing promise in a number of domains is Markov logic networks.[19]

Even without the introduction of uncertainty into logic representations, several challenges still exist in order to make practical use of first-order logic. Using logic to represent the semantics of natural language is one such challenging practical use, yet progress continues to be made. Biskri *et al.*'s article in this issue describes the use of categorical grammars to solve the coordination problem in natural language understanding, i.e., the relationship of a conjunctive set of phrases (e.g., "John runs fast and walks slow."). Descles and Pascu's article in this issue defines several new logic variable quantifiers (e.g., typical value or atypical value) beyond the tradition for all and there exists and which are commonly used in natural language (e.g., "Typically, mammals have hair."). These augmentations to the traditional use of logic for natural language allow the more-accurate representation of the meaning of the language, as well as support the automated reasoning with this knowledge to infer other knowledge about the topic of the discourse.

Another form of logical knowledge supporting the understanding of both natural and structured languages (e.g., XML) is the ontology[9]. Ontologies represent knowledge of a domain typically in taxonomic form, e.g., an "isa" hierarchy of concepts in a domain. Ontology mappings, a form of meta-knowledge about ontologies, allow the translation of concepts in one domain into the concepts of another domain. Several approaches for learning ontology mappings have been developed,[6] but Wong *et al.*'s article in this issue describes a knowledge engineering approach with specific application to mapping between ontologies defining concepts in security management. Their approach demonstrates the power of human augmentation of automatically-acquired, partial knowledge, which helps to complete our integrated vision of an end-to-end system for data to knowledge for reasoning, where full automation is not feasible. The user and domain expert are crucial participants in such a system despite our goal of lesser dependence on their intervention.

In addition to possibly incorrect acquired knowledge, another source of inaccuracy in a fully-automated system is occasional mistakes in the automated reasoning phase of the system due to errors in the implementation of the reasoning system. Most reasoning systems are built around automated theorem provers to answer questions targeted toward information and knowledge retrieval or the deductive inference of new knowledge. While we would hope that the theorem provers are correct, it behooves us to incorporate checks to ensure correctness. Sutcliffe's article in this issue describes a powerful form of checking based on a structural analysis of the proofs or derivations generated by an automated theorem prover. Such tools help to identify mistakes in the reasoning process and prevent the inclusion of incorrect, and potentially harmful, knowledge into the system.

Our discussion until now has focused on a traditional logic-based approach to knowledge representation and reasoning. However, when knowledge representation and reasoning is sought for a specific application or domain, non-logical approaches can perform better. Louwerse *et al.*'s article in this issue describes one such scenario in which knowledge representation is used to support natural language understanding. Here, they

use latent semantic analysis[13] (LSA) as the basis of a knowledge representation for natural language. Specifically, knowledge is represented as a correlation matrix over the corpus of the language, rather than a logical representation of the concepts and relations in the domain of discourse. They argue that this LSA-based representation is better suited to the types of reasoning typically performed in natural language-based tasks. Another scenario in which logic performs poorly as a representation of knowledge is in time series domains. Nilsson's article in this issue shows that a wavelet-based representation of time series data drawn from heart-rate sequences better supports the common information retrieval task of finding similar sequences in such domains.

The above approaches to knowledge representation to support both general-purpose and domain-specific reasoning tasks represent a small sample of the open issues and ongoing progress in knowledge representation and reasoning. Yet, this area represents the ultimate target of our initial challenge of transforming data into a form that allows us to better understand and reason about the domain. Therefore, as mentioned earlier, this ultimate use of learned knowledge should influence the knowledge acquisition process. How this influence is accomplished is an open issue, but clearly performance goals for reasoning (e.g., speed and correctness of inference, precision and recall of retrieved information, accuracy of prediction) represent quantitative feedback measures. While these measures have been used to influence and refine the knowledge acquisition process on an individual basis, their collective influence and iterative feedback into the process remains an open issue.

4. Conclusions

In this article we set out to describe the articles in this issue in the context of one of AI's grand challenges: the automated acquisition of knowledge from data along with the representation of this knowledge to support reasoning and understanding. While the articles lie on the cutting edge of the pursuit of this challenge, they represent only a small subset of the work on this challenge, which encompasses the fields of machine learning, knowledge representation and reasoning, automated theorem proving, search, uncertainty, natural language processing, and domains ranging from medical informatics to security and sociology. Still, we can draw some observations and generalizations from this work. First, most of the components of this challenge exist, and in many cases, multiple approaches exist for each phase of the challenge. While several domain-specific, end-to-end approaches have been implemented, general methodologies for solving this challenge are still under investigation. Second, as with the knowledge discovery process[7], this challenge requires an iterative approach, adapting and refining based on performance feedback and supporting human interaction and guidance. While such feedback has been incorporated into approaches to individual phases, feedback over the process as a whole is an open issue. Third, evaluation of approaches to this challenge requires new benchmark datasets or testbeds. Datasets would need to specify not only the initial data from which knowledge is acquired, but also the types of knowledge desired, alternatives for its representation, and goals for reasoning over this knowledge. Likewise, testbeds

akin to games and simulators are necessary to exercise and evaluate the various components of the challenge. One possible evaluation testbed is the automated acquisition of semantic knowledge from the world-wide web to support natural language, semantic-based (rather than keyword-based) querying of and reasoning about the web content knowledge. Such evaluations will drive the field toward integration of methods rather than just parallel pursuit of complementary sub-problems, with the ultimate goal of addressing the grand challenge of knowledge acquisition, representation and reasoning.

References

1. N. Boccara, *Modeling Complex Systems*. Springer, 2003.
2. B. Buchanan and M. Shotcliffe (Eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, Massachusetts, 1984.
3. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
4. A. Dempster, N. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):1–38, 1977.
5. T. Dietterich, Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science*, Volume 1857, 2000.
6. A. Doan, J. Madhavan, P. Domingos and A. Halevy, Learning to map between ontologies on the semantic web. *Proceedings of the Eleventh International WWW Conference*, 2002.
7. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* Volume 39, Number 11, 1996.
8. G. Flake, S. Lawrence and C. Giles, Efficient Identification of Web Communities. *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.
9. T. Gruber, A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993.
10. Z. Guessoum, Adaptive Agents and Multiagent Systems. *IEEE Distributed Computing*, Volume 5, Number 7, 2004.
11. M. Kearns and U. Vazirani, *An Introduction to Computational Learning Theory*. MIT Press, 1994.
12. J. Kleinberg, Authoritative sources in a hyperlinked environment. *Journal of ACM*, Volume 46, 1999.
13. T. Landauer, P. Foltz and D. Laham, Introduction to Latent Semantic Analysis. *Discourse Processes*, Volume 25, pp. 259-284, 1998.
14. W. Lee, Collaborative learning for recommender systems. *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 314–321, 2001.
15. H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.
16. G. Paliouras and V. Karkaletsis, *Machine Learning and Its Applications*. Springer, 2001.
17. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
18. R. Quinlan, Induction of decision trees. *Machine Learning*, 1(1): 81-106, 1986.
19. M. Richardson and P. Domingos, Markov Logic Networks. *Machine Learning*, 62, pp. 107-136, 2006.
20. F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan, Chicago, Illinois, 1962.

21. D. Rumelhart, G. Hinton and R. Williams, Learning Internal Representations by Error Propagation. In D. Rumelhart and J. McClelland (Eds.), *Parallel and Distributed Processing*, Volume 1, Chapter 8. MIT Press, Cambridge, Massachusetts, 1986.
22. F. Taroni, C. Aitken, P. Garbolino and A. Biedermann, *Bayesian Networks and Probabilistic Inference in Forensic Science*, John Wiley & Sons, 2006.
23. I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufman, 2005.