Provenance and Case-Based Reasoning*

David Leake
Computer Science Department
Lindley Hall, Indiana University
Bloomington, IN 47405, U.S.A.
leake@cs.indiana.edu

Introduction

Computational science takes a multidisciplinary approach to scientific investigation, tightly linking scientific research with computational studies and processes such as numerical simulation, data management, and visualization to study complex phenomena such as weather systems. The scientific importance of such processes has led to significant interest in recording the *provenance* of the data products, to support tasks such as assessing data quality. How to capture and use provenance information has become an important e-Science research area (Simmhan, Plale, & Gannon 2005). An interesting question is the possible relationships between this burgeoning provenance research and the capture and reuse of process information in artificial intelligence.

This invited talk presents ongoing research exploring ramifications of provenance tracing and reuse for two different facets of case-based reasoning (Mantaras *et al.* 2005). First, it considers the opportunities presented by provenance capture as a source of process information from which to mine cases. Second, it considers the use of provenance tracking *within* case-based reasoning systems, to enable them to make better use of their cases. By mining the provenance traces of external processes, a case-based reasoning system can harness increasingly extensive libraries of process information to extend its case library. By tracing the internal provenance of its own cases, a case-based reasoning system can better manage and refine its own knowledge.

Exploiting Provenance for Case Capture

Captured provenance data can be mined for cases to be used to predict workflow properties and for case-based support of human workflow generation. The talk sketches a range of challenges for mining and exploiting this information, and also highlights issues for how to integrate support into existing interfaces, how to provide useful advice, and how to balance automated and interactive processes. It illustrates these challenges and initial solutions in the context of ongoing research to integrate case-based reasoning with large-scale provenance capture to support workflow generation for e-Science.

Exploiting Provenance Considerations within CBR Systems

Case-based reasoning is inextricably bound to memory of the system's problem-solving process: The storage and reuse of cases is at the heart of CBR. However, with few exceptions (e.g., (Veloso 1994)), CBR systems seldom remember how cases were derived by the system. We have begun to explore the use of internal provenance information for tasks such as recovering from delayed feedback by propagating feedback information to related cases, predicting case quality based on quality loss through repeated adaptations, and using provenance to target maintenance (Leake & Whitehead 2007). The results illustrate how internal provenance information can be exploited to guide the case-base maintenance process. This provenance-based maintenance approach contrasts to prior maintenance work which only detects and fills gaps, or responds to problems revealed by feedback or inconsistencies: It exploits provenance information to enable active maintenance, by making *a priori* suggestions of candidates for case replacements or confirmations.

References

Leake, D., and Whitehead, M. 2007. Case provenance: The value of remembering case sources. In *Case-Based Reasoning Research and Development: Proceedings of the Seventh International Conference on Case-Based Reasoning, ICCBR-07*. Berlin: Springer-Verlag.

Mantaras, R.; McSherry, D.; Bridge, D.; Leake, D.; Smyth, B.; Craw, S.; Faltings, B.; Maher, M.; Cox, M.; Forbus, K.; Keane, M.; Aamodt, A.; and Watson, I. 2005. Retrieval, reuse, revision, and retention in CBR. *Knowledge Engineering Review* 20(3).

Simmhan, Y.; Plale, B.; and Gannon, D. 2005. A survey of data provenance in e-Science. *SIGMOD Record* 34(3):31–36.

Veloso, M. 1994. *Planning and Learning by Analogical Reasoning*. Berlin: Springer Verlag.

* This material is based upon work supported by the National Science Foundation under Grant No. OCI-0721674. This talk presents joint work with Scott Dial, Joseph Morwick, and Matthew Whitehead.