Assembling Latent Cases from the Web
A Challenge Problem for Cognitive CBR

David Leake

School of Informatics and Computing
Indiana University, Bloomington, IN, USA
leake@cs.indiana.edu

Abstract. Early visions of case-based reasoning stressed its broad applicability. Realizing the dream of near-universal application of CBR will require lowering the boundaries to entry for CBR applications. This position paper proposes, as a step toward that goal, the challenge problem of developing CBR systems with more human-like capabilities to exploit large multi-use memories to assemble "latent cases," e.g., building missing cases from fragments drawn from existing external information sources such as the the Web. This approach is inspired by the Dynamic Memory [1] model of human memory. The paper identifies key research areas for harnessing latent cases, including: (1) structural indexing, (2) case assembly, evaluation, and repair, (3) introspective reasoning, (4) provenance capture and analysis, and (5) storage/recall methods for dynamic cases.

Keywords: Dynamic Memory, Indexing, Retrieval

1 Introduction

Case-based reasoning (CBR) has deep roots in cognitive science. A substantial current of CBR research grew from research on human memory organization and reminding [1], and AI research on case-based models has sometimes given rise to psychological investigations (e.g., [2]). Since those early days, many AI efforts have focused on developing systems which are "cognitive," not in the sense of modeling underlying human cognitive processing, but in the sense of manifesting the robustness associated with human reasoning [3, p. 68]. Part of the desired robustness is the ability to parallel the human ability to solve a stunning range of problems, which people accomplish by drawing on an extensive memory of diverse information, much of which may have been assembled for other tasks. This position paper proposes moving beyond the current pattern of treating each application of CBR to a different task as independent, with each system tuned to a particular task, and attempting to develop more "universal" CBR systems, which apply CBR based on a shared general-purpose memory of diverse experiences and general knowledge developed for many tasks—without the assumption that any of those prior tasks will match the current task of the system.

Schank's Dynamic Memory Theory [1] proposed that human memory is reconstructive, with people retrieving episodes by re-assembling subparts organized by Memory Organization Packages (MOPs). Each subpart is stored independently and may be part of multiple memories, each one reflecting a different perspective. Dynamic Memory theory assumed that such reconstruction was supported by a memory structured and indexed to facilitate this process, and that the stored episodes were initially presented to the system in their entirety. The question we consider is whether CBR systems can be designed to perform assembly of memories from "latent cases"—materials on the Web which contain the information needed, but which may have been stored for other purposes and in multiple parts, without custom indexing to aid their reconstruction. For example, a travel case for an entire trip might be generated from many pieces representing separate parts of the journey, captured from a combination of blog entries, newspaper stories about a celebrity's travel, and general travel guides; a case for a software workflow might be developed by combining a partial provenance trace with inferred information, and a diagnostic case might be pieced together from a combination of a specific anecdote with general medical knowledge retrieved from various Web sources.

2 The Web as a Memory

Many CBR projects are already examining the integration of CBR with the Web (see [4, 5] for a sampling of some of this work), including studying how to draw on various libraries of case experience already stored on the Web. A central component of Plaza's [6] Experience Web proposal is the reuse of other's experiences, as when a reasoner might consider large numbers of hotel experience reports from different people, collected into a single archive. CBR research has generalized from single case bases towards drawing on multiple case bases (see [7] for an overview). However, such models share the fundamental CBR principle that the boundaries and content of cases have already been defined.

In contrast to the CBR assumption of clearly delineated knowledge units, humans dynamically draw from quite heterogeneous knowledge sources. For example, Barsalou's [8] research on ad hoc categories studies the categories that people form to satisfy particular goals, such as the formation of the category "things to take on a camping trip" or "places to look for antique desks." Kolodner's [9] research on retrieval and organizational strategies illustrates how people can reformulate queries to retrieve sought-after information which has not been indexed for their tasks.

Building on observations about the flexibility of human memory during problem solving, we propose developing CBR systems which can make similarly flexible use of knowledge sources such as the Web. viewing the Web as an enormous heterogeneous resource from which not only to collect, but also to assemble, augment, and repair assembled cases on demand. Of necessity, the proposed model makes no assumptions about the required case structure being predefined or the items in memory being pre-indexed for particular purposes.

3 Research Areas for Enabling Dynamic Case Assembly and Reuse

Enabling CBR systems to assemble cases will require general methods for enabling CBR systems to adjust to new tasks and exploit the Web as a heterogeneous memory from which the system, rather than the developer, delineates and builds cases. This goes beyond traditional case mining approaches (e.g., [10]) which aim for a one-shot extraction of a case base from a database. It also goes beyond multi-case-base reasoning [11] and multi-agent case sharing [12] in not assuming that cases will be pre-delineated, or even that tasks will be known in advance. In this vision, instead of the CBR system navigating a pre-delineated knowledge source, the system will navigate the Web in its raw entirety.

Managing the search and assembly process for latent cases will require a strong introspective process to determine the types of information needed in a case and how to search for and assemble them. Developing such capabilities can build on numerous prior projects in integrating introspection into CBR (e.g., [13–19], as well as research on reasoning about information needs from areas such as knowledge planning [20], goal-driven learning [21], and explanation [22, 23]. The following specific focuses are particularly important:

- **Identifying Needed Case Content and Structure**: Initial foundations have been developed for knowledge planning processes to reason about what information is needed and how to obtain it [20, 21]. Broader application of CBR will require methods for facilitating CBR task definitions in a form that can be used by knowledge planning systems.
- **Flexible Structure Matching**: CBR research has made much progress on developing carefully crafted indexing structures which distill relevant problem features into a simplified form to facilitate efficient matching. The latent case approach cannot assume that any of the case parts to be retrieved will have been predefined as problem/solution with respect to the problem at hand. Consequently, to retrieve the parts of latent cases for assembly, capabilities are needed (1) to pre-retrieve a limited pool of potentially relevant items, and (2) to efficiently identify relevance based on the content of those items, which will often require structural comparisons. Some research has studied methods for efficient structural matching [24], but Web context complicates this process by the potential need to compare structures in different formats, in different vocabularies, and possibly requiring inference to be comparable.
- **Using provenance:** Provenance information traces the derivation of data products. Outside of CBR, major initiatives are pursuing provenance capture [25]. Captured provenance is both a source to mine for cases [26] and a potential source for assessing trust [27]. The availability of such information provides a new resource for a certain type of web-based episodic information, traces of Web processing, but with two important caveats, which present challenges as well. First, provenance information captures unfiltered information, only some of which may be relevant in a given context, and which

captures a sequence of events without capturing their rationale. Second, automatic provenance capture tools may capture noisy data or may drop parts of a trace, which may need to be filled in [28]. Thus provenance information is a vital source of raw material—but even provenance-based cases must be evaluated, related to system goals, and possibly repaired.

- **Case assembly**: Retrieved components of latent cases may provide partial or incomplete information. Methods will be needed to guide assembly, and to guide choices when multiple alternatives are available (or to develop case representations which retain conflicting information and record conflicts for future consideration). For example, if building a planning case requires information about how a traveler of interest traveled within a city, and that information is not available, the "case" could contain information on a multitude of options mined from information about other travelers; this notion of case would be generalized from solely representing a unified individual experience.
- **Assessing result quality:** Cases mined on-the-fly may have gaps, alternatives (if multiple pieces of information are retrieved for the same component), or unreliable information. Consequently, applying such cases will require methods for assessing their completeness and the quality of the information they provide—and the ramifications when they fall short. An important resource for this assessment will be provenance information. This will need to include and merge any available provenance information about external sources—how they were generated and from what initial sources—and internal provenance information [29] about how the information was assembled.
- **Storage/Recall of Dynamic Cases**: The cases assembled for new problems become a resource to store, as in the standard CBR cycle. However, true to the spirit of Dynamic Memory Theory, the information assembled for the cases need not be copied. Instead, as appropriate, cases may become packages with pointers to external knowledge sources, enabling cases to change—automatically being updated—as their components change. This removes much of the case-base maintenance burden, replying on the external sources to perform needed updates. For example, consider a case reflecting calculation of the value of a home, based on recent sale prices in its neighborhood. If that case contains pointers to information about neighboring houses in a property database, when the case is next retrieved, it will reflect the updated values for any of those neighbors which have sold in the interim. Achieving this will require new case representations supporting the just-in-time access to and integration of external knowledge.

4 Conclusion

This position paper proposes that to develop cognitive CBR systems which take full advantage of the explosion of information on the Web, CBR researchers should embrace a view of the Web as a general-purpose memory, from which

cases are not only retrieved but assembled. Realizing this vision will require more self-aware CBR systems, with a broader view of retrieval and more flexible retrieval methods, as well as capabilities to consider case variants and repair their problems. The successful development of the desired general methods could, in a deep sense, replace "taking the task to CBR" with bringing CBR to the diversity of the Web. Even if the goal of this proposal is not fully realized, the methods developed towards the goal could increase the ability of CBR systems to augment their internal memories from Web-based knowledge, helping to increase the impact of CBR in a Web-based world.

Acknowledgments

This material is based upon work supported by a grant from the Data to Insight Center of Indiana University. I thank the anonymous reviewers for very helpful comments.

References

1. Schank, R.: Dynamic Memory: A Theory of Learning in Computers and People. Cambridge University Press, Cambridge, England (1982)
2. Read, S., Cesa, I.: This reminds me of the time when ...: Expectation failures in reminding and explanation. Journal of Experimental Social Psychology **27** (1991) 1–25
3. Brachman, R.: Systems that know what they're doing. IEEE Intelligent Systems **17**(6) (2002) 67–71
4. Bridge, D., Plaza, E., Wiratunga, N., eds.: Reasoning from Experiences on the Web, Seattle, WA, ICCBR (July 2009) ICCBR 2009 WORKSHOP.
5. Bridge, D., Delany, S., Plaza, E., Smyth, B., Wiratunga, N., eds.: WebCBR: Reasoning from Experiences on the Web (ICCBR 2010 Workshop), Alessandria, Italy, ICCBR (July 2010)
6. Plaza, E.: On reusing other people's experiences. Kūnstliche Intelligenz **23** (2009) 18–23
7. Plaza, E., McGinty, L.: Distributed case-based reasoning. Knowledge Engineering Review **20**(3) (2005) 315–320
8. Barsalau, L.W.: Ad hoc categories. Memory and Cognition **11** (1983) 211–227
9. Kolodner, J.: Retrieval and Organizational Strategies in Conceptual Memory. Lawrence Erlbaum, Hillsdale, NJ (1984)
10. Yang, Q., Cheng, S.: Case mining from large databases. In: Case-Based Reasoning Research and Development: Proceedings of the Fifth International Conference on Case-Based Reasoning, ICCBR-03, Berlin, Springer-Verlag (2003) 691–702
11. Leake, D., Sooriamurthi, R.: Case dispatching versus case-base merging: When MCBR matters. International Journal of Artificial Intelligence Tools **13**(1) (2004) 237–254
12. Ontañón, S., Plaza, E.: Collaborative case retention strategies for cbr agents. In: Case-Based Reasoning Research and Development: Proceedings of the Fifth International Conference on Case-Based Reasoning, ICCBR-03, Berlin, Springer-Verlag (2003)

13. Arcos, J.L., Mulayim, O., Leake, D.: Using introspective reasoning to improve CBR system performance. In: Proceedings of the AAAI 2008 Workshop on Metareasoning: Thinking About Thinking. (2008) 21–28
14. Bonzano, A., Cunningham, P., Smyth, B.: Using introspective learning to improve retrieval in CBR: A case study in air traffic control. In Leake, D., Plaza, E., eds.: Proceedings of the 2nd International Conference on Case-Based Reasoning (ICCBR-97). Volume 1266 of LNAI., Berlin, Springer (July 25–27 1997) 291–302
15. Fox, S., Leake, D.: Using introspective reasoning to guide index refinement in case-based reasoning. In: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Hillsdale, NJ, Lawrence Erlbaum (1994) 324–329
16. Leake, D.: Learning adaptation strategies by introspective reasoning about memory search. In: Proceedings of the AAAI-93 Workshop on Case-Based Reasoning, Menlo Park, CA, AAAI Press (June 1993) 57–63
17. Leake, D., Powell, J.: A general introspective reasoning approach to web search for case adaptation. In Bichindaritz, I., Montani, S., eds.: Proceedings of the Ninth International Conference on Case-Based Reasoning, Berlin, Springer Verlag (2010) In press.
18. Oehlmann, R., Edwards, P., Sleeman, D.: Changing the viewpoint: Re-indexing by introspective questioning. In: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Atlanta, GA (1994)
19. Ram, A., Cox, M.: Introspective reasoning using meta-explanations for multistrategy learning. In Michalski, R., Tecuci, G., eds.: Machine Learning: A Multistrategy Approach. Morgan Kaufmann (1994) 349–377
20. Ram, A., Hunter, L.: The use of explicit goals for knowledge to guide inference and learning. Applied Intelligence **2**(1) (1992) 47–73
21. Ram, A., Leake, D.: Learning, goals, and learning goals. In Ram, A., Leake, D., eds.: Goal-Driven Learning. MIT Press (1995)
22. Leake, D.: Goal-based explanation evaluation. Cognitive Science **15**(4) (1991) 509–545
23. Sormo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning—perspectives and goals. Artificial Intelligence Review **24**(2) (2005) 109–143
24. Falkenhainer, B., Forbus, K., Gentner, D.: The structure-mapping engine: Algorithm and examples. Artificial Intelligence **41** (1989) 1–63
25. Moreau, L., Ludäscher, B., Altintas, I., Barga, R.S., Bowers, S., Callahan, S., Chin, Jr., G., Clifford, B., Cohen, S., Cohen-Boulakia, S., Davidson, S., Deelman, E., Digiampietri, L., Foster, I., Freire, J., Frew, J., Futrelle, J., Gibson, T., Gil, Y., Goble, C., Golbeck, J., Groth, P., Holland, D.A., Jiang, S., Kim, J., Koop, D., Krenek, A., McPhillips, T., Mehta, G., Miles, S., Metzger, D., Munroe, S., Myers, J., Plale, B., Podhorszki, N., Ratnakar, V., Santos, E., Scheidegger, C., Schuchardt, K., Seltzer, M., Simmhan, Y.L., Silva, C., Slaughter, P., Stephan, E., Stevens, R., Turi, D., Vo, H., Wilde, M., Zhao, J., Zhao, Y.: Special issue: The first provenance challenge. Concurr. Comput. : Pract. Exper. **20**(5) (2008) 409–418
26. Leake, D., Kendall-Morwick, J.: Towards case-based support for e-science workflow generation by mining provenance information. In: Proceedings of the Ninth European Conference on Case-Based Reasoning, Springer (2008) 269–283
27. Briggs, P., Smyth, B.: Provenance, trust, and sharing in peer-to-peer case-based web search. In: Proceedings of the Ninth European Conference on Case-Based Reasoning, Springer (2008) 89–103
28. Cheah, Y., Plale, B., Kendall-Morwick, J., Leake, D., Ramakrishnan, L.: A noisy 10gb provenance database. In: Proceedings of the Second International Workshop

on Traceability and Compliance of Semi-Structured Processes, Berlin, Springer-Verlag (2011) In press.
29. Leake, D., Dial, S.: Using case provenance to propagate feedback to cases and adaptations. In: Proceedings of the Ninth European Conference on Case-Based Reasoning, Springer (2008) 255–268