Reasoning as Search: Supporting Reasoning with Distributed Memory*

Kristian J. Hammond

Northwestern University,
Department of Electrical Engineering and Computer Science,
2133 Sheridan Road,
Evanston, IL 60208

Abstract. The central idea behind Case-Based Reasoning has always been the notion that reasoning can be supported using memories of past problem solving. One bottleneck in this work has often been the development of case libraries needed to support this reasoning rather than the transformation of the cases themselves. In much of our work, we have taken an approach in which we treat web-recourses as the distributed knowledge engineering that can be integrated into memory- or case-based reasoning systems. We have been working on how we can take the core view of "Reasoning as Remembering" and transform it into "Reasoning as Search". The primary issues in this work are how to map problem-solving or task needs onto the queries required to find initial candidates, filter those candidates for relevance and then manage the exploitation of the results. I will outline how we have done this in two systems we have built recently, News at Seven and Baleen, systems that track the world of social media, news and the Web to support narrative generation.

Keywords: Case-based reasoning, CBR, narrative generation, search, intelligent information systems, dynamic memory, memory-based reasoning.

1 The Problem

One of the most persistent issues in Case-Based Reasoning (CBR) has been the need for robust case libraries that can used to support reasoning. In much the same way that rule-based systems were hampered by the need to build up domain level rule sets, CBR has been throttled by the need for large scale case bases.

There are many ways to address this problem that range from the use of existing libraries that were created for other reasons to the incremental learning of these libraries from completed problem solving exercises or the manual production of the cases needed to support a given system.

In most instances, the libraries and the languages that are used to represent cases within them are usually highly structured. Likewise, the queries that are used to

* Some of this material is based upon work supported by the National Science Foundation under Grants No. III-0917261 and 0535231. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

A. Ram and N. Wiratunga (Eds.): ICCBR 2011, LNAI 6880, pp. 1–5, 2011.
© Springer-Verlag Berlin Heidelberg 2011

access them are also usually highly structured and aimed at getting those and only those cases that are genuinely relevant to the problem at hand.

There are exceptions, but in general, there remains an ongoing need for the production of extensive, high quality and robust case-libraries that are both well structured and welled indexed.

Much of the work we have been doing over the past few years has been aimed at trying to find ways in which reasoning can be supported in the absence of these requirements.

2 The Opportunity and Challenge

One approach to this problem is to step back from the dual assumptions that case libraries have to be associated of a single point of view, problem solving approach or even system and that they have to be well organized or structured. That is, shift perspective towards repositories that are the product of the distributed submission of information where there is little, if any, control over the structure of the items submitted. Which is to say, use the Web.

The opportunity here is that there is an ever-growing mass of information online that can be minded if you know exactly what you are looking for and know how to go about doing so. This information is accessible using search tools that have been tuned for the retrieval of the most relevant content given a specific query.

The challenge, of course, is that this “information” is not information at all, but instead unstructured free text that is only informative when processed either by humans who know something about what they are reading or by systems that have similar knowledge. And while there are many both commercial and openly available tools for doing web search, many are tuned for goals and needs that are very different than those of a case-based reasoning system.

The ultimate challenge is to find a way to look at different tasks from the point of view of information retrieval, the trade-offs between precision and recall and support of the kinds of modifications that have to be performed if we are to actually used the ”cases” that are found through search.

In order to deal with this challenge, we decided to focus on content generation as a task in general. This focus provided us with mechanisms for characterizing the goals and needs of our systems in terms of lexically based search along with a framework for doing syntactic and lexical manipulation of cases to fit the generation needs of a system.

3 Filling in the Narrative: News at Seven

News at Seven is a fully automated approach to creating broadcast news. Starting with a set of user preferences, the system finds relevant stories, edits these stories for broadcast presentation, uncovers relevant background visual materials (e.g., videos and still images), and augments the primary text using sources such as blogs and other commentary found on the Web. These resources are then used to drive the construction of new material that is, in turn, used to control a set of animated

characters who reside in a virtual "performance world" we have created for them. The final output of the system is an online Flash presentation that uses animated avatars with generated speech and is modeled after traditional nightly news broadcasts. The system is robust, scalable, and highly flexible.

News at Seven can be used to generate a variety of program styles or dynamics, with specific anchor activities and interactions including traditional news but also entertainment stories (in an "Access Hollywood" style) and film reviews (paralleling Ebert and Roper).

It is this last dynamic that I want to focus on for the moment. The interaction that the system generates for this particular dynamic is driven by two elements. The first, and obvious one, is the name of a movie. The second is what people are saying about it in both the aggregate and the specific.

Fig. 1. News at Seven reviews Star Trek

Once the system is handed a movie title, it performs a series of searches aimed at figuring out, in general, what people think of this film. Although it is looking for an overarching opinion, it is not looking for a consensus so much as a characterization. Occasionally, everyone either hates or loves a film but more often, there are divided reactions to it. The goal of the system at this phase is to capture this characterization. To do this, the system builds a set of queries aimed at social media sites and platforms and gathers statements of opinion about the film that it categorizes as being positive, negative or neutral.

While doing this, there are some important elements of the processing that the system must attend to.

First, it has to make sure that the statements it is looking at are actually related to the movie whose title it used to craft its query. To do this, it needs to filter the statements using terms that tend to be used by moviegoers to describe their own actions. This allows the system to ignore a significant portion of the material that the search has pulled back.

Second, it has to do a credible job of determining sentiment with regard to the statements it has pulled down. While noisy, this can be done using standard sentiment terms, identification of negation, and proximity. Fortunately, at this stage, the system is looking for a fairly course grained characterization that it can use to drive the next stage of actual content generation.

The initial wave of retrieved content is used to determine what narrative *arcs* or *angles* are going to drive the generation process. In effect, the system is trying to figure out what kind of story it wants to tell. Is it a story about a blockbuster, about a film that has mixed reviews, a nice film that people like but not a lot of people see, etc? Once this is determined, the arc that dominates determines the dynamic (in terms of both what information is gathered and how it is presented) that will hold.

If the arc that holds is a mixed opinion arc, for example, the system then crafts a series of searches aimed at finding specific statements associated with the different film elements (actors, story, production value) that are strongly positive or negative. These statement are then used, with some minor syntactic modification, to create a back and forth dialog between two anchors. In cases of ambiguity, lack of syntactic clarity, grammatical problems, the statements are simply filtered out in favor of clearer and more useful ones. The result is a strikingly natural dialog that moves easily from point to point (driven by the narrative arc) focused on positive and negatives aspects of the film (provide by the secondary search) pulled together with interstitial material that the system uses to join the individual statements together.

Two relevant points here.

First, the tension between precision and recall that is the standard driver in search is muted here by the two roles that search plays. In the first phase of determining overall sentiment, recall is more important and any lack of precision results in noise the system is already dealing with. In the second phase, recall is completely irrelevant in that the system is actually looking for specific statements that fit its needs rather than a collection of statements.

Second, the two central processes of CBR, goal based retrieval and modification are still in play. The difference is simply that the retrieval is supported by a substrate of lexically based information retrieval rather a more structured approach and the modification is more syntactic in nature, changing the tense and structure of a statement to fit a new set of rhetorical goals.

One final note about News at Seven. One aspect of the system is that it searches for and modifies both video and stills for its use in crafting a user experience. In doing this, it is using the same core technology that it uses to find and modify text selections. In both these instances, it is defining an information goal, crafting a query aimed at finding a solution to that goal and then modifying the results it retrieves to best fit the problem at hand.

4 Serving the Narrative: Baleen

Baleen is part of the next wave of systems doing machine-generated content that are being developed by the company Narrative Science. Like News at Seven, the system is designed to use search to satisfy specific information goals. It however, is aimed at satisfying those goals using news rather than opinion.

Baleen's goals come from a parent system that is generating narratives as a product of data analysis rather than through search. At any given point in time, this parent system may be in the midst of generating a story in which the facts of the matter are clear in the data, but exogenous information that might explain those facts might not be directly available. In those situations, Baleen is called with explanatory goals that direct it to find the specific item that can then be incorporated into the ongoing story.

To make this clear, consider the example of the stock market. Occasionally companies (and their associated stock) undergo interesting changes. 52 week highs or lows, increase in price, increase in volume, etc. The stories associated the these changes can be written through a combination of tracking the current state of the world and then comparing the data pulled with historical information as well as data about other similar companies and how they are doing.

In crafting these stories, however, there is often a gap in the narrative in that the data alone does not explain *why* these changes have taken place. It is in these moments when Baleen is invoked.

Baleen is called with an information goal. In this instance the goal is to explain, for example, a sudden stock drop. Baleen uses existing data to determine the various ways a company can be described (its names and nick names) and the kind of industry the company is within. The latter is used to provide precise information about what actions (and thus terms) are appropriate to use in describing both positive and negative events. These three elements are combined to create an extremely focused query that is then used to access news sources.

The results of this search are then filtered using the same sort mechanism applied by News at Seven to make sure that the individual results are actually on point. The system is very aggressive about removing individual results in that, as with News at Seven, it is really only looking for a single result that can be modified to fit its needs. In order to verify results, however, it does need to look to see if entries cluster together, ignoring those individuals for which no cluster exists.

In the end, a single result is selected and then modified to fit the system's needs.

5 Conclusion

Both these systems are part of an ongoing push to make more and more use of the ambiguous, fluid and growing wealth of both structured and unstructured data that is the Web. Both however, are also cast in the mold of CBR with the notion that the massive knowledge engineering effort that is the Web will eventually create a case library for us that will be able to support most, if not all, of our reasoning needs.