Case-Based Goal Trajectories for Knowledge Investigations

Vahid B. Eyorokon, Uday Panjala, and Michael T. Cox
Wright State University
{eyorokon.3, panjala.6, michael.cox}@wright.edu

Abstract

Humans seek to gain knowledge and structure data by many means including both bottom-up and top-down methods. But often, people have a specific purpose to their activity that drives the process, that is, they have particular questions that need answering in support of some broader investigation. These questions often change as answers point in various directions during an investigation, whether the investigation is formal (e.g., scientific, legal, journalistic, or military) or simply an informal browsing of the internet. Here we take a mixed-initiative approach to knowledge discovery, and we present a system called Kyudo that supports the process using a conversational case-based reasoning process. Cases in Kyudo are sequences of knowledge goals or questions that form arcs through a multidimensional knowledge space and that form the core activity in a dialogue between the user and system. As the system gains more experience and therefore more cases, it is able to detect similarity in knowledge goals and prompt the user with additional relevant goals that can short circuit the human reasoning process to minimize tangents or false starts. In this paper we present a distance-based mechanism that reduces the total length of a goal trajectory through guidance that accelerates the human reasoning process and aids effective knowledge discovery.

Introduction

Modern information retrieval techniques using statistical inference have dramatically changed the way we interact with data. Parse techniques translate natural language questions into relational database queries (Yahya et al. 2012) or lambda calculus representations (Berant et al. 2013) that can be executed against a structured knowledge base. This has resulted in a shift in the balance between the amount of work a human must do to answer questions and the amount of work offloaded by augmentation systems. For example, previous search techniques were designed to deliver the most relevant content to a human user, but left them the task of reading and interpreting that content to inform their goals.

Our proposed approach for complex knowledge goal reasoning is a case-based reasoning (CBR) (Kolodner 1993; de Mantaras et al. 2005) system called *Kyudo*[1] (Eyorokon et. al. 2016) which reuses past experience in an interactive fashion. Kyudo is a part of a larger mixed-initiative information system called Ronin (Bengfort and Cox 2015). Interactive CBR operates similarly to conversational case-based reasoning systems, which incrementally elicit a target problem through an interactive dialog with the user, attempting to minimize the number of questions before a solution is reached (Aha, McSherry and Yang 2005). To provide an adaptable, investigative system, the methodology we are exploring guides the user in a finite length interactive dialogue, removing the requirement to minimize session length to facilitate an ongoing discovery process. Additionally, the system itself is a learning agent with the goal of predicting future knowledge goals, and acquiring the information in advance to provide specific guidance to the user.

We first present the concept of a case-based knowledge-goal trajectory and situate it within interactive question-answer dialogues. Next we examine distance-based representation for measuring similar knowledge goals, and then we examine case retrieval of trajectories. After discussion of evaluation, we close with related research and a conclusion.

Goal Trajectories and Dialogues

In knowledge investigations, a *knowledge goal* represents a user's desire for information, often expressed as a question or query (Ram 1990;1991). Achievement of a knowledge goal is decomposed into a hierarchical plan involving simpler sub-questions to augment the investigative process. During a dialogue with a data source, the user performs queries and continues to chain sub-goals together to work towards a solution. During this process, the search plan can change as the user discovers new information and forms new questions; indeed, the questions themselves can change. Like attainment goals (i.e., goals to achieve world states), that are subject to transformation (Cox and Veloso 1998), we claim that a knowledge goal is also subject to change. Therefore, as interactive reasoning changes a knowledge

[1] Kyudo is "the way of the bow" in the Japanese martial art of archery.

goal, the path that leads to the final information can be represented by a *goal trajectory*, $GT[1..n]$, a sequence of n knowledge goals each represented as a directed graph $G = (V, E)$ with vertices V and edges E.

$GT[1..n] = G[1] \mid GT[2..n] = \langle G[1], G[2], \ldots G[n] \rangle$

Each graph is a tree having root question $q' \in V$ and sub-questions as children along edges $\{(q', q) \in E \mid q', q \in V, q \neq q'\}$. Note that $V \subset Q$, with Q defined in the next section.

Goal trajectories can be influenced either through the direct interactive manipulation of goals (Cox and Zhang 2007); via other users in the system issuing similar queries that provide the basis for recommending new goals; or by monitoring new information that has been added to the knowledge base. An investigative dialogue can be therefore seen as a planning problem where knowledge goals are not static and must be responsive to goal changes. We believe that a system can leverage goal change to provide guidance by proposing medium steps towards a series of predicted goals. This guidance will accelerate the user who is likely to take short steps toward a goal, yet not provide uncanny or mystifying advice by proposing longer, unintuitive steps.

In the context of an interactive system that attempts to assist a user whose goals are changing, a dialogue is a discrete investigative session. In the dialogue, the user provides an initial knowledge goal (represented as a natural language question) to initiate the process. The system provides answers through traditional information retrieval techniques or by answering fact-based sub-questions via a structured knowledge base. It also provides guidance in the form of new goals that may accelerate the goal change process towards a final goal or to prevent the user from pursuing fruitless paths. During the dialogue, the system tracks goal changes by recording when new questions are posed. The dialogue is completed by a final knowledge-goal, presumably the target of the investigation. An investigation dialogue, I, is represented with an initial goal gt, user u, start time t_1[2], end time t_n, length n, and a Boolean (=⊤ if successful).

$$I = \{(gt, u, t_1, t_n, n, succ) \in GT \times User \times \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \{\top, \bot\}\}$$

$$User = \{(name, age, gender, mstatus) \in \Sigma^* \times \mathbb{N} \times \{male, female\} \times \{single, mar, divor, separ, wido\}\}$$

By tracking goal changes and the goals that proceeded them, a system can construct a goal trajectory from a dialogue. In fact, a one to one relationship exists between a dialogue and a goal trajectory. Since dialogues occur in context, cases can be retrieved based on the context of the user, which therefore improve guidance. It is for this reason that it is not correct to inspect or compare dialogues on their own, but rather to embed contextual information into the goal itself, so that similar goals may be proposed to the user. Dialogues are summarized as an ordered set of goals, subgoals (e.g. simpler goals designed to add information to the top level goal) and guidance that cause goal change.

Distance-Based Representations

A distance-based approach requires the encoding of knowledge goals into a vector representation so that related knowledge goals are nearer in the space. By computing the relative distances of these goals in the space, a goal trajectory can be said to have a length, i.e., the sum of the distances between the knowledge goals that compose the trajectory. The trajectory magnitude is the distance between the starting and final knowledge goal.

If the length of a trajectory is much greater than its magnitude, then the user has followed a complex and circular path to their final goal. Part of the reason for this could be tangential or false paths. A short circuit of a complex knowledge goal means that the guidance provided by the system brings the ratio of a goal trajectory's length and magnitude closer to 1, preventing complexity or poor reasoning.

When the length is similar or equal to the magnitude of the goal trajectory, this indicates a straight forward line of investigation that proceeds in an ordered fashion directly towards the final goal. In cases like this, guidance will accelerate the user towards the final goal, such that they may skip intermediate knowledge goals and arrive at their final goal much faster. However, here a system must be careful; it cannot simply point the user to the end of the goal trajectory as that might cause confusion when it is not clear how the path to that knowledge was deduced. Instead, acceleration skips over knowledge goals in the investigative chain, possibly providing feedback about why the acceleration would help. Feedback allows a user to closely associate the reasoning path with the system's guidance and trust it.

The two dialogues in Figure 1 are both within Kyudo's concierge corpus where the nodes are questions in goal space and their connection forms their dialogue. Both dialogues share their initial questions and lie one on top of the other. Their shared questions are about a family with children looking for an activity that is family oriented. Through their investigation, they learn that there is a fair in town. At this point, the tangent was introduced in Dialogue 48. The questions that appear in red are part of Dialogue 48 and are tangential and can be seen to diverge from the family related questions about the local fair as the user begins asking about bars and the best beers. Dialogue 48 then resumes its initial family-related inquiry, and both paths converge on the final knowledge goal about transportation to the fair.

[2] Time is a positive integer representing the number of seconds elapsed since the UNIX epoch (i.e., January 1, 1970).

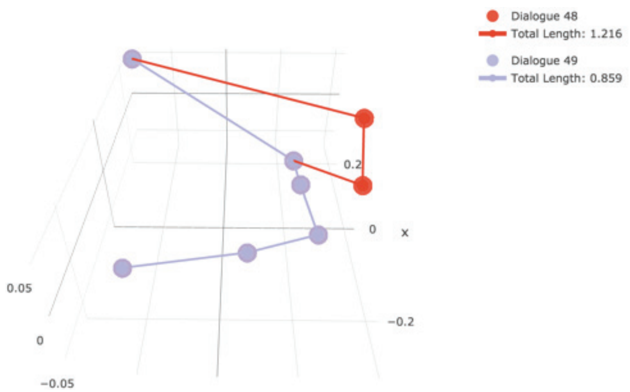

Figure 1. Two trajectories in a 3Dl goal space. Dialogue 48 is the current trajectory and contains tangential questions shown as red nodes. Dialogue 49 is the efficient trajectory where its questions are light purple nodes overlaid on Dialogue 48.

What is important to note is that the two tangential questions completely changed the direction of Dialogue 48's investigation in goal space. As investigations become more complex, tangents can become more frequent and can significantly increase the time taken to satisfy the knowledge goal of the user's investigation when they pursue unrelated and irrelevant paths of investigation.

The Knowledge-Goal Space

A *goal space* is a multidimensional representation of all possible knowledge goals where a point in this space is a goal location. Similar questions should thus be close as measured by Euclidean-like distances. Features of the natural language question group into *concept*, *context*, and *task*. Each is represented as a vector that together determine a goal space location.

$$Q = \{(c, k, \tau) \in Concept \times Context \times Task\}$$

1. **Concept Vector**: The concept vector uses a Term Frequency Inverse Document Frequency (TF-IDF) vector of the words in the question. Because questions are so short, TF-IDF is a good measure of the importance of infrequent words in the corpus. Moreover, this vector is reduced by a truncated singular value decomposition such that only the 50 best components of the vector remain. It can be described as follows, where μ is a string from the alphabet Σ, ρ is a parse tree, and $\xi \subseteq Topics$ is a set of topic labels.

$$Concept = \{(\mu, \rho, \xi) \in \Sigma^* \times G \times 2^{Topics}\}$$
$$Topics = \{topicLabel\}$$

2. **Task Vector**: We have identified 6 potential tasks related to why the knowledge goal is being solved including factual questions like "who" or "what", explanation questions like "why" or "how", as well as existential and permission tasks. This vector is simply a Boolean vector of the tasks of these questions based on a lightweight syntactic analysis as shown below

$$Task = \{(who, what, when, where, how, why) \in \{\top,\bot\} \times \{\top,\bot\} \times \{\top,\bot\} \times \{\top,\bot\} \times \{\top,\bot\} \times \{\top,\bot\}\}$$

3. **Context Vector**: The context vector embeds the user-specific information into the goal including time of day, location of query, as well as relative position of the knowledge goal in a dialogue.

$$Context = \{(u, t) \in User \times \mathbb{N}\}$$

These component vectors are easily computed from a natural language question using a lightweight parsing technique. The final knowledge goal representation is simply the ordered union of the concept, task, and context vectors. Because the TF-IDF vector has to be computed on a medium to large corpus of questions ahead of computing the vector at run time, three corpora were used: Free917, WebQuestions, and our concierge dialogues.

The distribution of questions from this corpus is shown using a principle component analysis (PCA) of the two or three most informative directions of each dimension, then mapped to two or three dimensions respectively. A two dimensional representation is shown in Figure 2. It is important to note that because PCA acts as a coordinate transformation, negative values can occur for some dimensions within the data, but the overall clustering of data points is still preserved.

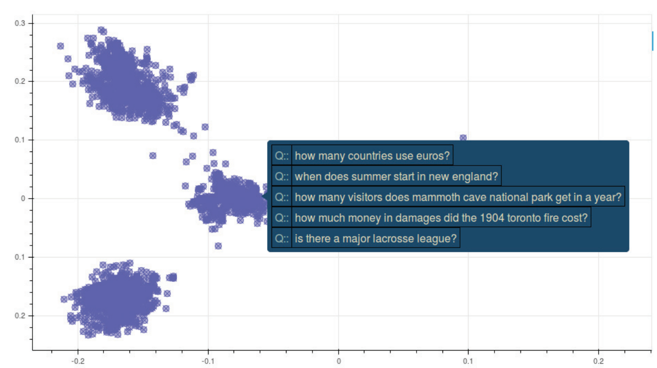

Figure 2. Two dimensional PCA projection of knowledge goals from Free917, WebQuestions, and a concierge corpora of question/answer pairs. Distinct clusters have formed, generally related to task. Questions at the cursor are enumerated with KNN.

Measuring Similarity

For each new question posed to Kyudo, the parser produces a grammatically structured parse tree of constituent phrases. Each token from the phrase is queried to Wikipedia to fetch the related semantic entities from the first section of the page called the extract. This extract is filtered for stop words and we calculate the TF-IDF measure over all the knowledge goals we have in the casebase. This provides an addition to Kyudo's concierge corpus to expand Kyudo's knowledge of semantically related keywords and therefore any conceptually related knowledge goal in Kyudo's casebase.

A weight is assigned to each of the three components, i.e., concept, context and task of the knowledge goal. The responses retrieved for a question with a greater weight on context would fetch more related questions matching the context metadata rather than matching either task or concept related keywords. Figure 3 shows the related questions fetched based on the method discussed above.

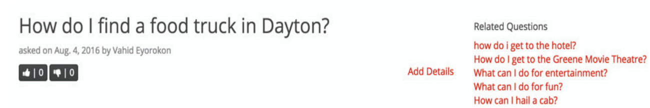

Figure 3. Kyudo application showing the related questions fetched for the natural language question asked.

Goal Trajectory Retrieval

Kyudo's interface engages the user in a conversation where the user asks a series of questions or knowledge goals specifying whether each one is a new goal or a sub-goal (see Figure 4). The series of knowledge goals asked becomes a dialogue or goal trajectory and is used to retrieve similar goal trajectories from Kyudo's casebase. Retrieval is based on: concept, context and task. For each new goal trajectory initiated by a user, Kyudo will first retrieve all stored cases of goal trajectories that pass contextual checks of time and profile similarity based on a *k-nearest neighbor* (KNN) algorithm (Aha, Kibler and Albert 1991; Cover and Hart 1967). In the time check, the start time for each retrieved case is compared to the start time of the new goal trajectory. If the time difference is less than an arbitrary threshold of four hours, the goal trajectory is said to survive and will then be processed for profile similarity. Profile similarity will then check the similarity of the user in the new goal trajectory against users in all surviving goal trajectories for matches in gender, marital status, and an age difference within a threshold of four years. The reasoning behind analyzing time and profile similarity is because the answer to questions for a given domain may depend on contextual attributes of the dialogue and its user. For example, within our concierge domain the answer to the question "What can we do for fun?" would have a different answer for a middle aged married man with his family asking it at noon than it would for a group of single twenty year old males asking the question at midnight. While the words used in both utterances of the question are identical, the underlying context is drastically different thereby requiring a different answer.

If the user profile of the processed goal trajectory matches in all three attributes, the goal trajectory survives to be processed by the final check of initial similarity.

Initial Similarity Filter

The Initial Similarity Check measures task and conceptual relevance of the knowledge goals that appear in the beginning of each surviving goal trajectory against the knowledge goals in the current goal trajectory whose length is n. For each goal in the current trajectory, a document is created consisting of the question text and all extracts of each constituent concept queried to Wikipedia. This effectively transforms our current dialogue into a list of documents. A similar process is performed on each surviving dialogue.

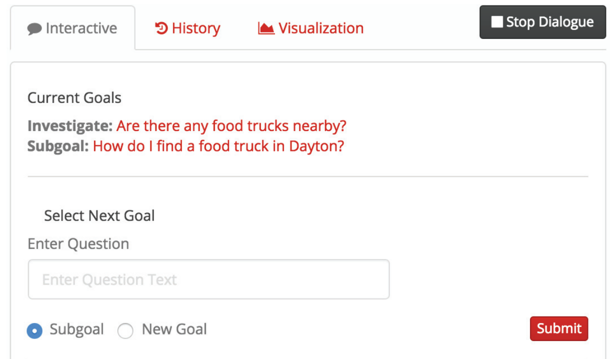

Figure 4. Kyudo's reasoner application interface

TF-IDF similarity scores are calculated for each document in the current dialogue against the document that coincides the same index of a surviving trajectory. This produces a list of similarity scores from 0 to 1 where 1 is a perfect match. If a score is at least 0.85, that document is considered to be a similarity match allowing Kyudo to identify conceptually related questions.

A sequence or list of similarity matches l is maintained, and if the number of similarity matches for a surviving goal trajectory is greater than or equal to a set threshold of $n/2$, where n is the length of the current dialogue, then the goal trajectory is considered a match. These matching goal trajectories are then ranked by their number of initial similarity matches and the *ids* for the top three dialogues with the greatest number of matches are returned. Figure 5 shows Kyudo's goal trajectory retrieval algorithm.

Evaluation

Inefficient goal trajectories are dialogues which contain tangents or false starts unrelated to the user's knowledge investigation thereby prolonging the time it takes the user to satisfy their knowledge goal. Trials were performed to measure the effect that tangents have on the length of a dialogue in goal space. Each trial consisted of one goal trajectory known to contain tangents called the current trajectory and one efficient goal trajectory without any tangents which was retrieved by Kyudo.

The questions within trajectories were transformed into three dimensional vectors where their (x, y, z) coordinates corresponded were their coordinates in the first, second and third principal components. Euclidean distances for each pairwise coordinates were found and the sum of these distances was the total length of the trajectory. The length of the efficient trajectory was compared against the length of the current trajectory and the difference was said to be the savings. The savings ratio is the savings over the length of the current trajectory. Five such trials were performed on dialogues which were created to demonstrate this tangent recognition.

Algorithm 1

```
function CONTEXT-FILTER(case-base, context-pair)
    passing ← []
    for case in case-base do
        if (|case.u.agecontext-pair.u.age| < 4) ∧
(|case.t1context-pair.time| < 4) ∧ ··· then
            passing ← case | passing
            return(passing)
        end if
    end for
end function
function SIMILARITY-MATCH(old-case, cur-case)
    count ← 0
    for i ← 1 to length(cur-case) do
        if similarity(old-case[i], cur-case[i]) >=
0.85 then
            inc(count)
            if count > length(cur-case)/2 then
                return(count, old-case)
            else
                return(−1, old-case)
            end if
        end if
    end for
end function
function SIMILARITY-SCORE(case-base, cur-case)
    sequence ← []
    for case in case-base do
        sequence ← SIMILARITY-MATCH(case,
cur-case)|sequence
        return(sequence)
    end for
end function
SIMILARITY-SCORE(CONTEXT-FILTER(case-base,
cur-case.k), cur-case)
```

Figure 5. Retrieval algorithm to find the best matching cases.

Preliminary results show a significant decrease in trajectory length (see Table 1). An average savings of 36.6% in the total length of trajectories was achieved where tangents were avoided. While this average savings may not be statistically significant, it is suggestive of Kyudo's capacity to identify and remove tangents within a dialogue. This translates into a more efficient knowledge investigation because of a shorter path towards knowledge discovery.

Trials	Retrieved Trajectory (Dialogue)			Current Trajectory (Dialogue)			Savings	Ratio
	ID #	Length	# Nodes	ID #	Length	# Nodes		
1	41	0.967	6	40	1.412	8	0.445	31.5%
2	49	0.859	6	48	1.216	8	0.357	29.4%
3	53	1.174	8	54	1.834	10	0.660	36.0%
4	37	1.755	6	36	2.819	8	1.064	37.7%
5	56	1.006	6	57	1.957	9	0.951	48.6%
Avg.		1.152	6.4		1.848	8.6	0.695	36.6%

Table 1. Percentage savings on the length of the retrieved goal trajectories compared with that of current goal trajectories. Note that the number of nodes is the number of questions in a dialogue.

Related Research

This work is related to (Powell 2011) which demonstrates a case-based reasoning approach to the adaptation of knowledge which can be dynamically mined from web-based resources. Their novel method utilizes large web-based data sets, similar to Wikipedia used by Kyudo, in order to solve the problem of adaptation or revision of a case to make it applicable to the user's current task. Our work with Kyudo adopts a similar web-based data mining approach and incorporates dialogues which track the overall sequence of goal changes and the evolution of the questions (Ram 1991) being asked. Our work differs in that goal changes, along with the addition of conceptual knowledge and contextual relevance, allow Kyudo to retrieve not just related questions, but entire dialogues. While their work improves their systems ability to adapt cases, our work focuses more on tracking the goal changes within the current dialogue thereby allowing Kyudo to guide the user's current investigation based on our casebase and short circuit the user's knowledge discovery process.

Our approach builds upon previous work (Aha et al. 2015), particularly a taxonomy of knowledge goals, to create a multidimensional representation of a knowledge goal. This representation is defined by a knowledge goal space with which we can compare goal similarity using distance metrics. This implementation therefore allows us to use a simple nearest neighbor algorithm to provide guidance to the user; a simplification that improves upon many challenges regarding case-based learning.

The work of (Aha, McSherry and Yang 2005) is highly relevant due to Kyudo's conversation based interface. Their work highlighted the importance of refining the user's question to solve a problem. Such problems faced by the user are usually vaguely or briefly defined and lack adequate detail for the system to provide a meaningful solution. Their method of implementing a conversational style interface to extracting details of a target goal was proven to be effective and close to the natural way humans communicate problems. Our work builds on this style of conversation based interface and also tracks the user's decomposition of goals into sub-goals. By allowing users to map out a 'plan' to solve their target goal, Kyudo can better understand the context of why goals change and identify false tangents to better provide guidance.

The research described in Cox and Zhang (2007) emphasizes the need for systems with the potential to maximize the individual strengths of humans and AI. Their work argues the need to "model planning as a goal-manipulation task" or mixed-initiative planning as opposed to traditional search based planners. Mixed-initiative planning regards plans as being a series of goal adaptations (Cox and Veloso 1998), and this overall evolution as being the byproduct of human and AI collaboration. Goal transformation and goal reasoning as discussed in Cox and Dannenhauer (2016) change the goals using metacognition.

Kyudo's Reasoner application engages the user in a dialogue, and the user provides Kyudo the general trajectory of the dialogue. As the dialogue progresses, Kyudo's understanding of both the target goal and retrieval improves, which enhances (Ram 1990) its ability to guide the direction of the conversation and streamline the user's knowledge discovery process. The dialogue constitutes a plan with the goal being to obtain an investigative solution.

The work done in de Mantaras et al. (2005) has proven to be invaluable as it succinctly describes the overall architecture of any case-based reasoning system. In their paper, they describe retention, reuse, revision and retrieval as being the main components of a case-based reasoning system. Kyudo adheres closely to their methodology which has provided a framework from which Kyudo was modeled. By providing our team with the terminology necessary for effective communication, the process of Kyudo's development has proven to be more practical.

Conclusion

Knowledge investigations made by humans often follow different trajectories based on the target goal of their investigation. By providing guidance towards a target knowledge goal, Kyudo accelerates the user's investigation process and helps to avoid false tangents. Our work represents a knowledge goal as a vector comprising of concept, context and task components. Kyudo incorporates Case-based reasoning principles for goal retrieval by calculating the TF-IDF of the given natural language question to expand its own knowledge by fetching the related semantic entities from Wikipedia. This enables Kyudo to provide the user with more accurate information vital to achieve successful knowledge discovery and retrieve matching goal trajectories.

Current work is being done to further develop Kyudo's guidance capabilities and recognition of false tangents within a dialogue. Further evaluation is needed on examples that use randomly chosen dialogues. Adaptation of related dialogues has yet to be developed and is another area of future research.

Acknowledgements

This research is funded by ONR under grants N00014-15-C-0077 and N00014-15-1-2080. We also thank the anonymous reviewers for their comments and suggestions.

References

Aha, D. W.; Kibler, D.; and Albert, M. K. 1991. Instance-based learning algorithms. *Machine Learning 6*: 37-66.

Aha, D. W.; McSherry, D.; and Yang, Q. 2005. Advances in conversational case-based reasoning. *The knowledge engineering review 20*(03):247–254.

Bengfort, B., & Cox, M. T. 2015. Interactive reasoning to solve knowledge goals. In D. W. Aha (Ed.), *Goal reasoning: Papers from the ACS workshop* (pp. 10-25).

Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, volume 2, 6.

Cover, T. M.; and Hart, P. E. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*(1): 21–27.

Cox, M. T.; and Veloso, M. M. 1998. Goal transformations in continuous planning. In *Proceedings of the 1998 AAAI fall symposium on distributed continual planning*, 23–30.

Cox, M. T.; and Dannenhauer, D. 2016. Goal transformation and goal reasoning. In *Proceedings of the 4th Workshop on Goal Reasoning*. New York, IJCAI-16.

Cox, M. T.; and Zhang, C. 2007. Mixed-initiative goal manipulation. *AI Magazine* 28(2):62-73.

de Mantaras, R. L.; McSherry, D.; Bridge, D.; Leake, D.; Smyth, B.; Craw, S.; Faltings, B.; Maher, M. L.; T Cox, M.; Forbus, K.; et al. 2005. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review* 20(03):215–240.

Eyorokon, V., Bengfort, B., Panjala, U., & Cox, M. T. 2016. Case-Based Goal Trajectories for Knowledge Investigations. A. Coman & S. Kapetanakis (Eds.), *Twenty-Forth International Conference on Case-Based Reasoning Workshop Proceedings: Synergies between CBR and Knowledge Discovery* (pp. 202-211). Atlanta, GA.

Kolodner, J. L. 1993. *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann Publishers.

Powell, J. 2011. A generic web-based knowledge discovery framework for case-based reasoning. Ph.D. Dissertation. Computer Science, Indiana University, Bloomington, IN.

Ram, A. 1990. Knowledge goals: A theory of interestingness. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 206–214.

Ram, A. 1991. A theory of questions and question asking. *Journal of the Learning Sciences 1*(3-4):273–318.

Yahya, M.; Berberich, K.; Elbassuoni, S.; Ramanath, M.; Tresp, V.; and Weikum, G. 2012. Natural language questions for the web of data. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 379–390. Association for Computational Linguistics.