Towards a Model of Pattern Recovery in Relational Data

B. Rode
Cycorp
3721 Executive Center Drive, Suite 100
Austin, Texas 78731, U.S.A.
ben@cyc.com

Keywords: multiple competing hypotheses, information extraction and link analysis

Abstract

This paper describes some fundamental issues associated with using rule-based reasoning to discover evidence for the instantiation of complex threat types in relational data. Based on the considerations raised, I delineate a general model for the pattern recovery process.

1. Introduction

By *pattern recovery* I mean the process whereby a human or an application discovers that the existence of an entity satisfying a generic definition is strongly implicated by some subset of the contents of a relational database. One might wonder why more effort has not been spent to date in constructing a formal theoretical framework to describe this process, given that the challenge of identifying the signature of a 'threat scenario' or a 'threat group' is a signal priority in coping with the asymmetric threat environment, but lack of progress becomes more understandable once some of the associated difficulties are clarified. A fundamental problem is that we lack consensus regarding the meaning of 'pattern'. For this treatment, I'll take a 'pattern' to be a set of quantified rules that are regarded as characterizing instances of a certain type. Although it can be argued that many threat types of interest to the intelligence community are characterized by a recursive hierarchical structure lending itself to rule-based representation, I should emphasize that my choice doesn't amount to advocacy of the rule-based approach *per se*: rather, it is born of the conviction that my points can most readily be illustrated within a rule-based context, and the suspicion that many pattern-recovery systems, rule-base or otherwise, will face challenges that are at least analogous to those I describe.

We may take it as a given that any type which is sufficiently interesting to warrant defining is sufficiently interesting to warrant reifying. By the same token, it is unlikely that a type this interesting can be provided with a *complete* definition: inevitably, there are facts about any really interesting class that we will not know. Consider the definition of some type taken to be a natural kind—say, 'cat'. If the average person were asked to define this, it would not be unusual for them to respond with a series of assertions: *every cat has a tail, every cat has a set of whiskers, every cat has four paws,* and so on. All of these statements share a common formal template, $\forall x(Cx \rightarrow *)$, where * is a formula stating a condition with *x* as an open variable; i.e., these definitional assertions state *necessary* conditions[1] on being a cat. That the common definition should take this form is not surprising: giving *sufficient* conditions for what it takes to be a cat is a harder business, and one best left to trained biologists. However, one cannot use assertions of the form $\forall x(Cx \rightarrow *)$ to deductively *conclude*, of some individual *a*, that *Ca*. What I am doing when I reason to *Ca* is sometimes called *abductive* reasoning: using satisfaction of the consequent of a conditional sentence as a guide for *hypothesizing* the antecedent--and a definition that is as heavy on necessary conditions as most common sense definitions implicates this sort of abductive reasoning if the task at hand is to find instances of the type in question. In fact, the actual case is more complicated in that, in the real world, *necessary* has to be weakened to *defeasibly* or, perhaps, *probabilistically necessary*, and *sufficient* has to be weakened to *defeasibly* or *probabilistically sufficient.* The conditions given, even where they can be stated, can't be said to hold 'no matter what', but only 'under normal circumstances', or, alternatively, with a certain measure of probability.

This represents a complication because the propriety of using a pattern component for defeasible reasoning as opposed to a check on abduction is constrained by the problem space. E.g., suppose we have a 'contract kill' pattern which specifies that the agent who hires the hitman

[1] Some readers will have noted that not all of the conditions I have stated *are* necessary. The conditional employed in every example I have given really can't realistically be taken to be the material conditional of first-order logic, an issue which will be dealt with in due course.

and the agent who initiates the final after-payment must be the same person, and further, that our pattern matching technology has identified a wire transfer that it thinks constitutes a sufficient condition in context for defeasibly inferring that the transfer constitutes part of the after payment, and also that the same technology identifies a meeting in a restaurant as a candidate occurrence in which the hiring was consummated. *If* the source data is such that we are reasonably certain few individuals in the data use aliases, or that all of the cases where an alias is used are known, then a reasonable way of checking the hypothesis[2] is to check whether the initiator of the transfer is the same as the other attendee at the meeting; but if there is reason to believe the 'aliasing' level is high, using pattern equivalence information as the basis for a constraint check is *not* recommended: indeed, the desirable procedure in this case might be to use the same equivalence knowledge to infer that the other participant at the meeting and the initiator of the suspected after-payment are the same person, and at least one of the different names used was an alias. In addition, the line between reasoning abductively and reasoning probabilistically from a set of sufficient conditions is dangerously thin: that is, some problems may call for abduction where others call for defeasible or probabilistic inference, but which method is employed is partly a function of how the pattern itself is represented. In short, the heuristic nature of the definitions in play insures that the question of how the patterns are used and represented cannot, finally, be separated from the strategy of pattern recovery that is being pursued.

2. Sketch of a Pattern Recovery Model

We are now in a position to provide a sketch of how the pattern recovery task might plausibly be carried out. Suppose we have an invariant pattern, **P**, consisting largely in quantified rules giving defeasibly necessary conditions on the instantiation of some concept, **C**. Suppose further that we hypothesize an instance of **C**, *i*, and use the rules of **P** to deductively conclude all we can about it, skolemizing additional terms as necessary. If we now go to the trouble of unifying skolemized terms based on knowledge of what is equivalent to what (a well-specified pattern should tell us how to do this) we will be left with a set of ground sentences that we can conjoin. If we take the additional steps of consistently replacing all of the individual denoting terms in the conjunction with variables and existentially quantifying over these, we will be left with an existential description *E*. If we now imagine **P** to be associated with a contingent pattern **S** and a relational database $\mathcal{D}$[3] (and assuming there to be a well-specified scheme for translating between the content of $\mathcal{D}$ and the language in which *E* and the elements of **P** and **S** are expressed), the task of recovering **P** from $\mathcal{D}$ consists in inferring *E* from $\mathcal{D}$ using **S** together with whatever methods are available to us, including not only standard deductive inference, but also abduction, graph-based pattern matching, constraint reasoning, and probabilistic reasoning. In the real world the dataset will almost certainly not contain sufficient data to support deductive derivation of the existential. Nondeductive methods will have to be involved; the challenge consisting in the fact that we can't give a problem-independent 'recipe' for selecting these techniques. This suggests that the centerpiece of the scheme should consist in an electronic medium where a the capabilities of a number of specialized modules are posted, along with a data model that contains information about the content of the dataset and features such as a quantitative measure of 'aliasing'. An 'inference strategist' with access to this information, together with the invariant pattern, the contingent pattern, and general world knowledge should be implemented to try to apply the modules towards inferring *E* in a way that we are justified in believing has reasonable chances for success.

[2] I shall refer to any conclusion arrived at by other-than-deductive means, as an *hypothesis*.

[3] For present purposes I assume a single dataset as the locus of the source data. A natural extension of the model features integration across multiple knowledge sources.

3. Conclusion

For better or for worse, this paper has served to reiterate how hard the unconstrained pattern recovery problem is. While invoking an inference strategist is not quite as bad as saying 'magic happens', it has to be admitted that at this stage of the game, we have only the sketchiest and most analogical ideas of how this sort of agent might work. To be sure, humans manage equally hard tasks using data that is as obfuscated as anything in use in the asymmetric threat regime, but for a comparatively narrow range of problem domains for which we are evolutionarily optimized. The asymmetric threat problem, by its very nature, expands that vista to problem domains for which humans are not naturally well-suited. Nevertheless, we do not have the luxury of turning aside from the challenge, given what's at stake. And we have at least some hope that persistence in this regard will pay off. The best, and, indeed, the only viable strategy open to us is to continue ahead in the assurance of reason and perseverance eventually yielding results.

Bibliography

McCarthy, John. 'Approximate Objects and 'Approximate Theories' in *Proceedings of KR- 2000: Principles of Knowledge Representation and Reasoning.* Morgan-Kaufmann. 2000.

Senator, T. 'Ongoing Management and Application of Discovery Knowledge in a Large Regulatory Organization: A Study of the Use and Impact of NASD Regulation's Advanced Detection System (RADS)' in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining,* pp 44-53. Boston. 2000.