CleanTax: A Framework for Reasoning about Taxonomies

David Thau
Dept. of Computer Science
University of California
Davis, CA 95616
dthau@ucdavis.edu

Shawn Bowers
Genome Center
University of California
Davis, CA 95616
sbowers@ucdavis.edu

Bertram Ludäscher
Dept. of Computer Science & Genome Center
University of California
Davis, CA 95616
ludaesch@ucdavis.edu

Abstract

The CleanTax framework relates (aligns) taxonomies (inclusion hierarchies) to one another using relations drawn from the RCC-5 algebra. The taxonomies, represented as partial orders with additional constraints, can frequently (but not always) be represented with RCC-5 relations as well. Given two aligned taxonomies, CleanTax can infer new relations (articulations) between their concepts, detect inconsistent alignments, and merge taxonomies. Inference and inconsistency detection can be performed by a variety of reasoners, and in cases where all relations can be described by the RCC-5 algebra, qualitative spatial reasoners may be applied. When inferring new articulations between taxonomies, CleanTax often poses many highly related queries of the nature "given what we know about the relations between two taxonomies, $\mathbf{T_1}$ and $\mathbf{T_2}$, what do we know about the relationship between concept A in $\mathbf{T_1}$ and concept B in $\mathbf{T_2}$?" This context of posing many (millions) of simple, but highly related queries motivates the need for qualitative reasoning systems that can perform batch jobs and leverage reasoning performed in the past to optimize answering queries about similar situations. This paper describes the CleanTax framework and argues for the development of benchmarks that take throughput into consideration, as well as single-query response time.

Introduction

The CleanTax framework assists metadata curators as they attempt to align taxonomies. Imagine a biologist integrating data sets that contain information about various species. Species are organized into taxonomies, and these taxonomies evolve over time as new information is learned about the taxa. Because of this, the meanings of the species names may differ from data set to data set, depending on the taxonomy used by each data set; complicating data sharing and discovery. To address this problem, biologists are publishing alignments between well-known taxonomies.

CleanTax facilitates the creation and utilization of taxonomic alignments by detecting logically inconsistent alignments, and inferring unstated articulations between concepts. CleanTax also graphically displays taxonomies, taxonomy alignments, and taxonomic merges.

Taxonomies are frequently under-specified, described only by the subsumption relationships of the concepts. However, additional unstated constraints are often assumed, such as that concepts are composed of the disjoint union of their children. These constraints may impact the logical consistency of articulations, as well as the entailment of additional articulations. CleanTax enables the exploration of these constraints, showing their effects on reasoning and merging across multiple taxonomies. (Thau, Bowers, and Ludäscher 2008)

CleanTax supports reasoning about taxonomies and taxonomic alignments in a variety of logics, depending on the articulations used, and the assumed taxonomic constraints. In some situations, the articulations and constraints may be captured by relations from the RCC-5 algebra, in which case qualitative reasoners are applied. In other cases, constraints or relations are outside the RCC-5 algebra, in which case other logics (currently monadic first-order logic) are applied.

The CleanTax Framework

In the following section, we describe the CleanTax framework in more detail.

Taxonomies

A *taxonomy* $T = (N, \preccurlyeq_N, \Phi)$ consists of a set of *names* (or *taxa*) N, a partial order (isa-hierarchy) $\preccurlyeq_N$, and a set of additional constraints Φ over N. Typical constraints that might be in Φ include:[1]

- *non-emptiness*: $c \neq \emptyset$ (for some or all $c \in N$)
- *sibling-disjointness*: if $c_1 \prec\!\!\cdot\; p$ and $c_2 \prec\!\!\cdot\; p$ then $c_1 \cap c_2 = \emptyset$
- *parent coverage*: $p \subseteq c_1 \cup \ldots \cup c_n$ (where all $c_i \prec\!\!\cdot\; p$)

When any of these constraints is applied to every applicable taxon in a given taxonomy, we call the constraint a *globally applied taxonomic constraint* (GTC).

Articulations

CleanTax uses the RCC-5 (Randell, Cui, and Cohn 1992) topological algebra as the basis for representing articulations. The RCC-5 algebra uses the same five *basic relations* ($\mathbb{B}_5$) as several biological taxonomic alignments and taxonomic reasoning systems (Berendsohn 2003; Koperski et al. 2000; Franz, Peet, and Weakley 2006).

[1] We write $x \prec\!\!\cdot\; y$ and say that x is *covered by* y, if $x \prec y$ and there is no other $z \in N$ with $x \prec z \prec y$; so x is a "direct" child of y.

Typically, not all of the articulations between concepts in two taxonomies will be given. The type of reasoner employed to infer these unstated articulations depends largely on the constraints appear in Φ. When the parent coverage constraint is not applied to any node in N, all of the relationships fall under the RCC-5 algebra, and therefore a qualitative spatial reasoner may be used. When the parent coverage constraint is applied, articulations are converted into logic formulas (Thau and Ludascher 2007) and then a first-order logic reasoner is applied.

Sometimes the relationship between two concepts is uncertain, and this uncertainty is represented with disjunctions of the RCC-5 relations. The power set of the basic five relations ($\mathbb{R}_{32}$) describes all of the possible disjunctive relations. Many of these $\mathbb{R}_{32}$ relations have been used in the real-world taxonomic alignments we have considered. In practice, the aligned taxonomies we have seen can be described by tractable subsets of the $\mathbb{R}_{32}$ relations (Renz and Nebel 1997). However, this is not necessarily the case, and particularly when integrating data, it may often be the case that a problem falls into a non-tractable subset of the RCC-5.

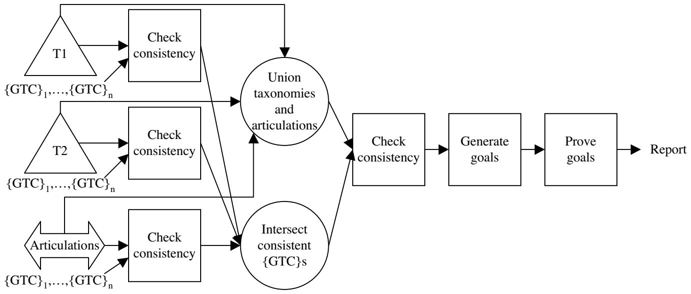
T1
Check consistency
$\{GTC\}_1,\ldots,\{GTC\}_n$
Union taxonomies and articulations
T2
Check consistency
$\{GTC\}_1,\ldots,\{GTC\}_n$
Articulations
Check consistency
$\{GTC\}_1,\ldots,\{GTC\}_n$
Intersect consistent {GTC}s
Check consistency
Generate goals
Prove goals
Report

Figure 1: Overview of the CleanTax framework.

Implementation

The CleanTax framework roughly follows the flowchart shown in Figure 1. Two taxonomies, and the articulations between them, are each checked for consistency under each combination of GTCs. The combinations of GTCs which are consistent for both taxonomies and the articulations are then applied to the combined taxonomies and articulations, and again, consistency is checked for each combination of GTCs. For each surviving combination of GTCs, the relationship between each pair of concepts in the combined taxonomies is determined by iterating through some set of the $\mathbb{R}_{32}$. Once the queries have been asked, the system reports on the relations between each pair of concepts in the taxonomies. The most complex situation tested to date involved one taxonomy $\mathbf{T_1}$ of 218 concepts, one taxonomy $\mathbf{T_2}$ of 142 concepts, and a set of 206 articulations between the taxonomies. There are 142 * 218 = 30956 articulations between the taxonomies, and in a naive brute-force approach, each of the $\mathbb{R}_{32}$ relations (except $\perp$, which never holds as long as the combined taxonomies and articulations are consistent) should be checked, resulting in 928,680 queries for a single combination of GTCs. Various optimizations have been devised to limit the number of questions asked (Thau 2008). However in the described scenario, the best optimization still resulted in tens of thousands of queries.

Reasoner Requirements

The CleanTax framework requires a reasoner which can answer hundreds of thousands of very similar, fairly simple queries. This requirement advocates for reasoners and benchmarks which stress throughput, rather than single-query response times. To maximize throughput, a reasoner should be able to reuse results from previous queries. It should also be able to schedule queries in a way that might maximize the possibility for reuse. Optimally, the scheduler would be able to divide queries into partitions that could be run in parallel, for cluster computing environments.

Conclusion

We have presented here CleanTax, a framework for applying the RCC-5 algebra toward reasoning about aligned taxonomies. This application of the algebra leads to requirements that may not be common among other qualitative spatial reasoning applications. We hope that the application of the RCC-5 algebra in this context can motivate requirements that may generalize to other domains.

References

Berendsohn, W. G. 2003. *MoReTax – Handling Factual Information Linked to Taxonomic Concepts in Biology*. Number 39 in Schriftenreihe für Vegetationskunde. Bundesamt für Naturschutz.

Franz, N. M.; Peet, R. K.; and Weakley, A. S. 2006. On the use of taxonomic concepts in support of biodiversity research and taxonomy. Proceedings of the New Taxonomy Symposium.

Koperski, M.; Sauer, M.; Braun, W.; and Gradstein, S. 2000. *Referenzliste der Moose Deutschlands*, volume 34. Schriftenreihe Vegetationsk.

Randell, D. A.; Cui, Z.; and Cohn, A. 1992. A spatial logic based on regions and connection. In Nebel, B.; Rich, C.; and Swartout, W., eds., *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*. San Mateo, California: Morgan Kaufmann. 165–176.

Renz, J., and Nebel, B. 1997. On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the region connection calculus. In *IJCAI (1)*, 522–527.

Thau, D., and Ludascher, B. 2007. Reasoning about taxonomies in first-order logic. *Ecological Informatics* 2(3):195–209.

Thau, D.; Bowers, S.; and Ludäscher, B. 2008. Merging taxonomies under RCC-5 algebraic articulations. In *Workshop Proceedings of the 17th International Conference on Information and Knowledge Management*. ACM.

Thau, D. 2008. Reasoning about taxonomies and articulations. In *Workshop Proceedings of the 11th International Conference on Extending Database Technology*. ACM.