Challenges in Presenting Argumentation Results

Laura Rassbach, Elizabeth Bradley

University of Colorado
Department of Computer Science
Boulder, CO 80309-0430
laura.rassbach@colorado.edu, lizb@cs.colorado.edu

Abstract

We present the initial user interface for the Calvin system. Despite designing for ease-of-use and simplicity, users had significant problems with this initial interface. Possible solutions to these problems are also presented.

1. Introduction

Displaying the results of a qualitative reasoning system to users in a useful and understandable way is almost as important as creating a system that generates the right results in the first place. Even when a system is capable of significantly aiding users with some task, they will tend to choose not to use it when the interface is baffling or impossible to use. However, designing a clear and comprehensible user interface is hardly a trivial task. We encountered significant challenges while designing a user interface for Calvin, an argumentation system for problems in cosmogenic isotope dating (a geological field). Although we believed that Calvin's initial interface would be clear and easy to use, we found that users struggled to understand the interface and did not use it as intended. Significant changes are now required to our initial interface. This paper discusses these challenges and our ongoing plans for improving the user experience with Calvin.

2. Calvin

Experts in cosmogenic isotope dating frequently need to identify what geologic processes are most likely to have affected a set of data they are examining. This process is difficult and time consuming, requiring significant amounts of expertise. Calvin is a qualitative reasoning system aimed at automating portions of this process. The nature of isotope dating (very little data and few unassailable interpretations of any data) makes it difficult to arrive at a definitive answer for any dataset, so Calvin generates a complete argument for each possible process. An argument is similar to a proof in first-order logic, but typically contains both evidence for and against the conclusion in question. Calvin is able to make judgments about which argument is 'best,' but it is not our goal to remove the expert from decision-making. Instead, we intend for the expert to be able to carefully examine Calvin's reasoning and make a final judgement about the process. Calvin's judgments about the best arguments are intended to guide experts' attention to the most likely possibilities.

Internally, the argument for a particular process is a collection of trees. The nodes in each tree are created from rules in Calvin's knowledge base, with every rule used in the argument represented as a node in one or more trees. The root of every tree in the collection for a particular process is a rule directly referring to that process. Because these argument trees are different from proofs in that they are defeasible, Calvin needs some way to judge the relative strengths of different support for a conclusion. Calvin uses a system of 2-element confidence vectors. The first element of the vector is called a 'match' and indicates how closely the current set of data matches Calvin's rule base. Match values are further composed of a truth and a degree. The 'truth' can be either true or false (matches or doesn't match the rule base) and the 'degree' refers to how far the actual data is from any threshold in the rule. For instance, if Calvin's rule base states that elevation greater than 10,000 feet is evidence for snow cover (one possible geologic process), elevation values of 1,000, 8,000, 11,000, and 15,000 feet would all generate different match values. The other dimension of confidence is a quality measure, and can be 'poor,' 'okay,' 'good,' or 'definite.' High elevation is only poor quality evidence for snow cover, whereas ages correlated with boulder size is good evidence (because smaller boulders will be covered with snow for more of the year).

3. User Interface 1.0

Figure 1 shows a screenshot of our initial user interface. We designed the initial user interface to display the complete set of arguments for each process in a simple and intuitive way. We expected this task to be relatively simple because experts quickly grasp Calvin's underlying logic when it is explained to them in an informal setting. When Calvin completes its analysis of a dataset, it brings up a window from which the user can access all the arguments

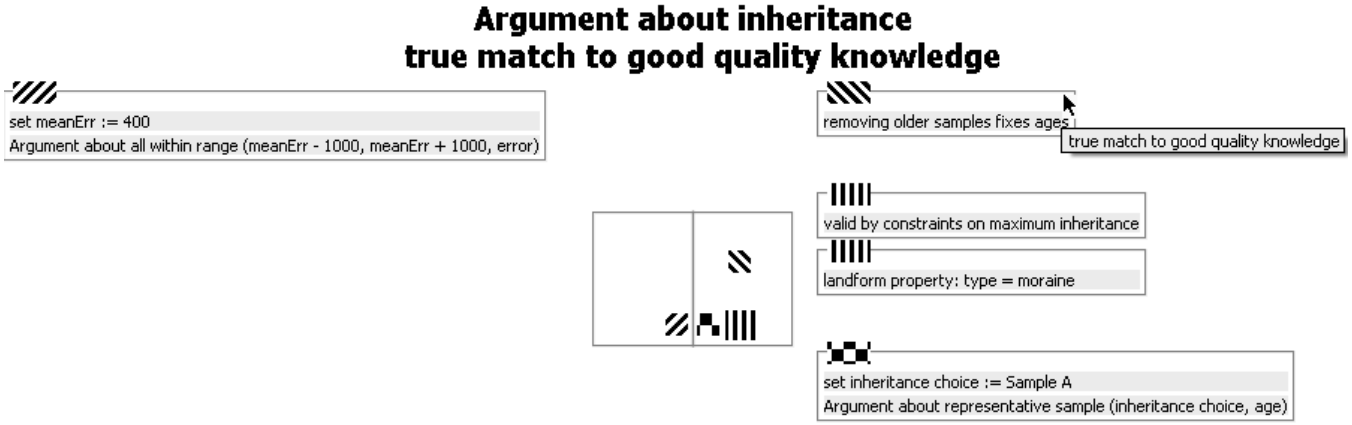

Figure 1: Screenshot of Calvin's initial user interface. Calvin's user interface has several components. A single screen is intended to describe the entirety of Calvin's knowledge about a particular conclusion. At the top of the screen is the conclusion under consideration and a gestalt of Calvin's confidence in that conclusion. Evidence related to the conclusion is divided into evidence against, on the left, and evidence for, on the right. This split is somewhat more apparent in the original because items on the left have red color markings and items on the right are green, however, in this screenshot colors have been replaced with textures for printing. Each distinct piece of evidence is enclosed in a box with text describing Calvin's reasoning. Every box and every text item has a tooltip allowing the user to view Calvin's confidence in these individual items. Users can also click on some text items to get more detail about Calvin's reasoning (either a subargument or the results of a simulation). Finally, the grid in the center provides a visual display of how many items of evidence Calvin has found at each 'level' of confidence: larger boxes represent more items.

it has generated. The first argument displayed is the one with the highest overall confidence.

The display for an individual argument includes the top-level conclusion being argued about (for instance, whether snow cover is a factor for this dataset) and the overall confidence in that conclusion. This confidence is converted directly into text from its internal representation (a typical confidence might read 'very true match to okay quality knowledge'). Evidence for and against the conclusion is sorted into two separate columns, sorted from strongest to weakest. Every piece of evidence contributes an individual confidence to the overall confidence in the conclusion. This contribution is displayed in two ways: as a color-coded block with each piece of evidence and as text available as a tooltip. Users can examine sub-arguments by clicking on their summary; they are displayed in the same format as the top-level arguments. Users can also get a more detailed description of any complex calculations performed by Calvin. For example, Calvin's rule base sometimes needs to determine whether a set of data points falls on a straight line. Users can click on this piece of evidence to get a description of how closely the points fit a line, the statistical significance of the fit, and a graph of the points and the best-fit line.

In the center of the argument display is a final representation of how the individual evidence items contribute to the Calvin's confidence in the overall argument conclusion. This representation is a grid with a location for each possible confidence value. Matches are listed from left to right, most extreme false through most extreme true, and quality is listed from top to bottom highest to lowest quality. Calvin sizes each of the grid cells according to the number of pieces of evidence it has at the corresponding confidence level for the argument. The cells are the same color as the block displayed with the evidence items.

4. Interface Weaknesses and Solutions

Observing experts using this initial interface has revealed a number of problems. The first and most obvious of these problems is that experts refer to Calvin's highest-confidence argument as the 'answer' and do not perform any further analysis when this answer is wrong. In particular, even when the confidence in the initially displayed argument is 'somewhat true match to poor quality knowledge' (the lowest possible 'true' confidence), users do not appear to make the inference that Calvin was unable to generate any good arguments for any process with the current data. Instead, they report a wrong answer and appear to make the inference that Calvin is quite certain of this answer.

An extension to this issue occurs when Calvin is able to generate quite good arguments for several processes. This often happens in datasets with very few samples. In this case, experts have expressed concern with the fact that Calvin came up with any answer at all, since there is quite good evidence for many processes. Instead of looking at the other arguments generated by Calvin, they look only at the best one, and appear to assume that all other arguments generated by Calvin must have been sufficiently inferior to reject them completely. Experts using Calvin appear generally uninterested in looking at the full spectrum of arguments it has generated.

We intend to address both of these problems by presenting a summary of all the arguments generated by Calvin. This will serve as the initial user view, rather than the single best argument generated. This will permit users to get a sense of perspective before viewing individual arguments in more detail.

Even while viewing a single argument in isolation, experts do not use Calvin's interface as intended. In particular, they seem almost uninterested in the evidence that has led Calvin to draw a specific conclusion. Although they appear to glance briefly at it, they seem not to understand it or to think critically about it without prompting. When using Calvin independently, experts did not choose to examine sub-arguments by clicking on them.

It is not yet clear precisely why users behave in this way. We believe that the size and color of the labels may make them difficult to read, and their positioning on the screen makes them less prominent. In addition, it has been observed by non-expert users that the text actually displayed in these labels is somewhat obscure and difficult to follow. We plan to make these descriptions of evidence physically easier to read and less obscure, but expect to need more iterations before experts are examining them closely and critically.

Finally, experts observed that the confidence system was somewhat confusing. This is especially true because understanding the report of a confidence level requires understanding Calvin's internal representation of confidence. However, even after discussing the meaning of various confidence values, experts seemed to have difficulty understanding what a specific confidence meant. Users did appear able to understand the centered, colored grid displaying the confidence contributed by all the evidence after an explanation of its purpose.

Clearly we need to alter Calvin's interface to make confidence more approachable to the user. We are considering a number of changes. The first and most simple is obviously to 'tweak' the language used to describe a confidence value to the user so it is more natural. If this proves to be impossible or insufficient, we may attempt to map Calvin's 2-dimensional confidence values onto a single dimension (e.g. 'no evidence,' 'little evidence,' 'some evidence,', etc.). In addition, we plan to experiment with the idea of never combining the confidence in evidence against some conclusion with evidence for it. Instead, confidence in the evidence for and against a conclusion could be presented separately. Doing this might have the additional benefit of encouraging users to examine the actual evidence that has produced these confidences.

5. Conclusion

Creating a usable and intuitive user interface for displaying the results of a qualitative reasoning system is an interesting challenge. Even when the underlying system is comparatively intuitive, conveying its results in a way that users understand is more complicated than simply placing them in a graphical interface. Despite our efforts, Calvin's user interface will require significant changes to meet our usability goals.