Identification of Variables in Model Tracing Tutors

Can Model Tracing Tutors Track Unstructured Problem Solving?

http://www.physics.rutgers.edu/~shapiro/tutor/

Chun Wai Liew
Computer Science
Lafayette College
liew@cs.lafayette.edu

Joel A. Shapiro
Physics & Astronomy
Rutgers University
shapiro@physics.rutgers.edu

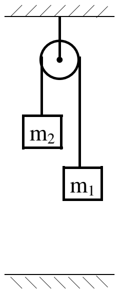
Don Smith
Computer Science
Rutgers University
dsmith@cs.rutgers.edu

Our Goal

To automatically provide concept-based feedback to students in college level introductory physics courses. Problems are presented as English text with an accompanying diagram. Students provide a set of equations as their answer. The following is an exerpt from "Physics for Scientists and Engineers" by Raymond A. Serway.

Example Problem

When two unequal masses are hung vertically over a light, frictionless pulley as shown to the right, the arrangement is called Atwood's machine. Determine the acceleration of the two masses and the tension in the string.



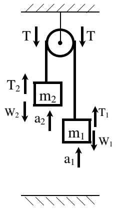
Concept-based Decomposition

The student is expected to recognize the properties of the objects (*e.g.,* the mass of block one) and the relationships among these properties. For this example based on Atwood's machine a student would need to employ Newton's second law on each block, knowledge that the string's length is constant, and how forces are transmitted around an ideal pulley.

A complete concept-based decomposition would result in the set of equations shown below. These could be developed automatically from a problem definition and first principles or provided manually for a problem.

Concept-based Decomposition

$T = T_1$
$W_1 = m_1 g$
$T_1 - W_1 = m_1 a_1$
$T = T_2$
$W_2 = m_2 g$
$T_2 - W_2 = m_2 a_2$
$a_1 = -a_2$

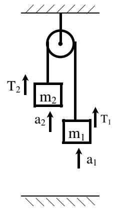


Student Solutions

Solutions provided by students, even correct solutions, tend to look very different than the complete concept-based decomposition shown to the left. Two typical solutions to this example are shown. Solution B is the most common

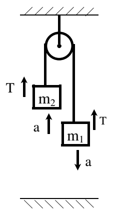
Student Solution A

$T_1 - m_1 g = m_1 a_1$
$T_2 - m_2 g = m_2 a_2$
$T_1 = T_2$
$a_1 = -a_2$



Student Solution B

$T - m_1 g = -m_1 a$
$T - m_2 g = m_2 a$



Focus on the Variables

What information is needed for an automated tutor to *understand* the student's use of variables and equations? Must the student explicitly map each variable to its corresponding physics concept? Must each equation be an instantiation of a single physics concept?

We've focused on the semantics of variables and developed techniques that can infer a variable's dimension from its use in the set of equations provided as a solution. This is a critically important first step to inferring the full semantics of the variable and equations used by the student to solve the problem.

Using scaffolding

It is important for a student to *know* the meaning of each variable used in their solution. Requiring complete explicit definitions is often a pedagogically useful requirement for a beginning student. However, with time and experience the student adopts standard conventions that can be used to reduce the need for explicit definitions of all variables.

As a student matures in her use of variables and equations, forcing her to completely and explicitly define each variable proves to be a frustrating and can cause her to lose interest and motivation.

Inferring Dimensions of Variables

To remove the need for explicit specification of variables we rely on basic physics naming conventions used in conjunction with an analysis of the variables as they are used in the equations. The process is performed in 3 steps.

1 - Use standard physics naming conventions to determine possible names for each variable

2 - Develop constraints on each variable's dimension from the equations

3 - Apply constraints using locality-based heuristics

This process may find one possible assignments of dimensions to variables, no consistent assignments, or multiple consistent assignments. Only when the result is inconsistent or ambiguous would the student be required to provide definitions for the variables.

Standard Naming Conventions

Our naming convention is derived from introductory physics texts, such as "Physics for Scientists and Engineers" by Raymond A. Serway, and augmented by the matching hierarchy shown below. A student's variable may be matched by several entries in our table and each match may provide multiple candidates for the variable's dimensions.

1 - Prefix match: an entry in the table matches a prefix of the student variable. For example, student variables tx, tstop, and t_1 would all be matched by a table entry of t.

2 - Preemptive prefix match: same as a prefix match except that all shorter prefix matches are discarded. For example, student variables thetax and thetay would be matched by a table entry of theta which would prevent the table entry of t from matching .

3 - Exact Match: This case requires the student's variable to be an exact match to the table entry. This is a rare match and used from variables such as G, the Universal Gravitational constant.

An example

$T_1 - m_1g = m_1a_1$

The following example highlights how the dimensions of variables are inferred. We use a single equation from Student Solution B shown above.

The example is worked through three separate steps

1 - apply a standard naming convention

2 - constraints on each variable's dimension are derived from the equations

3 - constraints are applied using locality-based heuristic

Handling Incomplete and Incorrect Information

Constraint propagation handles complete and correct information by finding all possible consistent assignment of dimensions. Experience has shown that there is an unambiguous interpretation 90% of the time.

However, when information is incomplete or incorrect the challenge is not only to recognize that there is a problem but to localize the problem so that effective feedback can be provided to the student. To do this we added heuristic control structures to constraint propagation that enable the method to localize an inconsistency within the smallest containing subexpression.

One of the important key observations was that students often use the same variable to represent different properties. For example, a_1 might represent an acceleration in one equation but a distance in another equation. To handle these cases and provide a basis for effective feedback our approach treats each instance of a variable independently and iteratively adds identity constraints until all constraints have been applied. The process proceeds in four successive stages. Each successive stage builds on the results of the previous stage.

1 - dimension from leaf nodes are propagated up the equation parse tree
2 - within equations information is propagated down the tree
3 - identity of variables within equations is asserted and validated
4 - identity of variables across equations is asserted and validated

Naming Conventions

$T_1 - m_1g = m_1a_1$

From variable naming conventions we get the following matches for each variable. The matches employed are not case sensitive and cover all of introductory physics and not just Newtonian Mechanics. When considered independently T_1 could be any one of five different dimensions, m_1 could be any one of three, g could be anyone of three, and a_1 could be any one of two. Standard naming conventions yield 90 possible combinations consistent with the variables used.

T_1 : time [s], tension [kg·m/s^2], kinetic energy [kg·m^2/s^2], temperature [K], thickness [m]

m_1 : mass [kg], angular magnification [dimensionless], integer [dimensionless], magnification [dimensionless], magnetization [C/m·s]

g: acceleration [m/s^2], density of states [s^2/kg·m^2], Newton's G [m^3/kg·s^2]

a_1: acceleration [m/s^2], distance [m], amplitude [several]

Experimental Results

An Andes corpus containing 6000 logs and 9865 sets of equations was analyzed. The dimensions of 90% of variables used by students to solve introductory physics problems can be inferred from their equations using localized constraint propagation.

This success rate has been confirmed on both complete and incomplete solutions as well as correct and incorrect solutions. Naming conventions were based on all of Newtonian Mechanics and not stylized for the corpus being investigated.

Ambiguous Variables	Ambiguous Constants	Percent of Corpus
0	0	83 %
0	> 0	6 %
1	> 0	3 %
> 1	> 0	6 %
Inconsistent Dimensions		2 %

Equations-based Constraints

$T_1 - m_1g = m_1a_1$

The equation's parse tree can be annotated with constraints that encode dimensional analysis. For example, only terms with the same dimensions can be added, subtracted or equated. When factors are multiplied dimensions are added.

By applying these constraints to the example equation the ninety possible combinations are reduced to the single possibility highlighted in red. In this case , T_1 is a tension, m_1 is a mass, and g and a_1 are accelerations.