Journal of Visual Languages and Computing 21 (2010) 249–262

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Reasoning about spreadsheets with labels and dimensions ☆

Chris Chambers *, Martin Erwig

Oregon State University, USA

ARTICLE INFO

Keywords:
Spreadsheet
Dimension
Unit of measurement
Static analysis
Inference rule
Error detection

ABSTRACT

Labels in spreadsheets can be exploited for finding formula errors in two principally different ways. First, the spatial relationships between labels and other cells express simple constraints on the cells usage in formulas. Second, labels can be interpreted as units of measurements to provide semantic information about the data being combined in formulas, which results in different kinds of constraints.

In this paper we demonstrate how both approaches can be combined into an integrated analysis, which is able to find significantly more errors in spreadsheets than each of the individual approaches. In particular, the integrated system is able to detect errors that cannot be found by either of the individual approaches alone, which shows that the integrated system provides an added value beyond the mere combination of its parts. We also compare the effectiveness of this combined approach with several other conceivable combinations of the involved components and identify a system that seems most effective to find spreadsheet formula errors based on label and unit-of-measurement information.

1. Introduction

Spreadsheets are widely used [1] end-user programs that contain many errors [2]. To improve the quality of spreadsheets a variety of approaches to prevent, detect, and remove errors from spreadsheets have been investigated. Since preventive approaches, in principle, have to interfere with the creation process that makes spreadsheets so attractive to end users, much research has instead focused on the detection and removal of errors.

One type of error that can be detected in spreadsheets is dimension errors, which occur when units of measurement are used incorrectly in formulas. Units of measurements can be employed as a concrete notion of types that is well known among end users [3], and are used to characterize different kinds of values, much like traditional, more abstract, type systems used in general-purpose programming languages. For example, a floating point number, which has just one type, can nevertheless represent different kinds of quantities, such as length or time values.

Several systems [4,5,3] have been developed in order to deal with dimension errors. Among these dimension inference [4] is a method that can be used to automatically find dimension errors in spreadsheets. This approach has been shown to work reliably and effectively in many cases, however, it does not take full advantage of the information provided in the spreadsheet as it does not utilize the structure of the spreadsheet and focuses on ensuring that formulas are dimension correct.

In contrast, there are several systems that are designed to directly take advantage of the labels and the structure of spreadsheets. These purely label-based approaches, such as UCheck [6] or the system described in [7], are designed to find formula errors caused by inconsistent label usage. This technique operates in two distinct analysis phases. The first phase defines header or label

☆ This work is partially supported by the National Science Foundation under the Grant ITR-0325273 and by the EUSES Consortium (http://EUSESconsortium.org).

* Corresponding author.
E-mail addresses: chambech@eecs.oregonstate.edu (C. Chambers), erwig@eecs.oregonstate.edu (M. Erwig).

1045-926X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jvlc.2010.08.004

information for the entire spreadsheet. UCheck is able to infer this while most others require users to annotate the labels for every cell. Once the headers are determined for a sheet, labels are assigned to cells based on headers and formulas. In the second phase this information is analyzed to find errors in formulas by identifying inconsistent labels.

One thing to note when looking at UCheck and dimension inference is that both systems rely on header and label information. However, this information is used quite differently by the two systems. By combining dimension analysis with the purely label-based approaches the structure of the spreadsheet could be used to help strengthen the reasoning of the system. To some degree this was already tried in the SLATE approach [5]. However, SLATE only transforms labels and dimensions and does not identify errors. Moreover, the fact that SLATE is a stand-alone spreadsheet system that cannot be integrated into Excel together with the additional time required of a user to annotate a spreadsheet renders the approach currently impractical.

In recent work [8] we have designed an integrated system that combines label-based reasoning with dimension inference. This approach was achieved by gathering both label and dimension information about a cell. In many cases a spreadsheet contains dimensions on only one axis, with the labels on the dimension free axis going unused in dimension inference. These unused labels can help to provide structural information that can be exploited by the reasoning system behind UCheck, and create a system that can check a spreadsheet for unit of measurement errors as well as detect label errors.

When run on the EUSES repository this combined system was able to detect several errors that neither UCheck nor dimension inference were able to discover. This brief evaluation shows the validity of a combined label and dimension checking system, but was this the best possible combination?

In previous work [9], a system architecture was developed to study how two specific systems could be integrated, in this case, WYSIWYT and UCheck. In this work, errors are introduced into spreadsheets and several combinations of the systems are tested to determine the best possible combination of reasoning systems. The process showed that when systems are combined the results are positive and more errors can be detected.

In this paper we will describe several different options, along with the original integrated label and dimension analysis [8], that can be pursued to check a spreadsheet for label and dimension errors. We will evaluate these different systems by running them on the 487 dimension and formula contained sheets in the EUSES corpus, as determined in, [4] and determine the efficiency and correctness for each combination.

The rest of this paper is structured as follows. In Section 2 we illustrate the issues involved in adding label reasoning to dimension inference with a small example. In Section 3 we formalize spreadsheets and present models of dimensions and labels. The original combined analysis method is described in Section 4. In Section 5 we describe the different modifications made to the system. Section 6 reports on the evaluation of these prototypical implementations. We discuss related work in Section 7 and give conclusions in Section 8.

2. A motivating example

To explain how the integration of spatial and semantic label analysis works, we will show how both dimension inference and the integrated system work on the spreadsheet in Fig. 1. This spreadsheet is calculating how far specific cars can travel on a full tank of gas based on the result of a drive using only five gallons of gas.

When the spreadsheet is checked with dimension inference, it would first identify the headers for all cells. When the headers are analyzed for dimension information, B1, C1, D1, E1, and F1 all map to a valid dimension. This would allow the system to check all formulas in this spreadsheet for dimension correctness. In this case, the system would detect that there is an error in cell D4 where the formula is trying to add miles and gallons.

Upon further inspection it could be noted that the spreadsheet contains another error. In this particular example, the cell F2 has the formula E2*D3. The dimension for E2 is Gallons, and the dimension for D3 is Miles per Gallon, which, when multiplied together result in the dimension Miles. This result contains no dimension errors, but it does not seem right. E3 is actually total Gallons for the Camry, while D2 is the MPG for the VW Bug. Logically, the result does not make sense, however, plain dimension inference would have no way to catch this.

By integrating label reasoning and checking that formulas are both dimension *and* label correct, the system presented in this paper is able to identify a previously unnoticed error. The first step is to determine which header axis (row or column) will be used as the dimension axis. In this case there are several dimensions on the horizontal axis (row 1), but dimensionless labels on the vertical axis (column A). Therefore, the system then identifies labels and dimensions for each cell. For example, the cells in row 2 would have the label "VW Bug".

With this information assigned, the system can then check to ensure that formulas are dimension and label correct. When the system checks the formula in F2 it can identify that it is multiplying a cell, E2, with the unit Gallons and the label "VW Bug" with the cell D2, which has the unit "Miles/Gallon" and the label "Camry". While the dimensions work out in this formula, the system will identify an inconsistency with the labels and be able to report this to the user, as shown in Fig. 2. The cells F4 and F5 are shaded yellow to indicate problems caused by the propagation of errors.

3. Representations for spreadsheet analysis

In this section, we will formalize the notions of spreadsheets, dimensions, and labels in preparation for the formal rule system that is discussed in Section 4.4.